Contextual and Ethical Issues with Predictive Process Monitoring

Ogunbiyi, Oluniyi

UNIVERSITY of WESTMINSTER

Contextual and Ethical Issues with Predictive Process Monitoring

by

Oluniyi (Niyi) Ogunbiyi

A thesis submitted in partial fulfilment for
the degree of Doctor of Philosophy

in the

School of Computer Science and Engineering
Health and Social Care Modelling Group

January 2022

Abstract

This thesis addresses contextual and ethical issues in the predictive process monitoring framework and several related issues. Regarding contextual issues, even though the importance of case, process, social and external contextual factors in the predictive business process monitoring framework has been acknowledged, few studies have incorporated these into the framework or measured their impact. Regarding ethical issues, we examine how human agents make decisions with the assistance of process monitoring tools and provide recommendation to facilitate the design of tools which enables a user to recognise the presence of algorithmic discrimination in the predictions provided.

First, a systematic literature review is undertaken to identify existing studies which adopt a clustering-based remaining-time predictive process monitoring approach, and a comparative analysis is performed to compare and benchmark the output of the identified studies using 5 real-life event logs. This curates the studies which have adopted this important family of predictive process monitoring approaches but also facilitates comparison as the various studies utilised different datasets, parameters, and evaluation measures.

Subsequently, the next two chapter investigate the impact of social and spatial contextual factors in the predictive process monitoring framework. Social factors encompass the way humans and automated agents interact within a particular organisation to execute process-related activities. The impact of social contextual features in the predictive process monitoring framework is investigated utilising a survival analysis approach. The proposed approach is benchmarked against existing approaches using five real-life event logs and outperforms these approaches. Spatial context (a type of external context) is also shown to improve the predictive power of business process monitoring models.

The penultimate chapter examines the nature of the relationship between workload (a process contextual factor) and stress (a social contextual factor) by utilising a simulation-based approach to investigate the diffusion of workload-induced stress in the workplace.

In conclusion, the thesis examines how users utilise predictive process monitoring (and AI) tools to make decisions. Whilst these tools have delivered real benefits in terms of improved service quality and reduction in processing time, among others, they have also raised issues which have real-world ethical implications such as recommending different credit outcomes for individuals who have an identical financial profile but different characteristics (e.g., gender, race). This chapter amalgamates the literature in the fields of ethical decision making and explainable AI and proposes, but does not attempt to validate empirically, propositions and belief statements based on the synthesis of the existing literature, observation, logic, and empirical analogy.

Table of Contents

List of Figures

List of Tables

List of Algorithms

Acknowledgement

I would like to express my immense gratitude to all those who have supported me throughout my studies.

First, I would like to thank my supervisory team, Dr Artie Basukoski (Director of Studies) and Prof Thierry Chaussalet for their guidance, support, and constructive feedback during this journey.

I would also like to acknowledge Dr Andrzej Tarczynski, Prof Sophie Triantaphillidou and the staff of the Graduate School. Finally, I would like to thank my colleagues at the Health and Social Care Modelling Group (HSCMG) at the University of Westminster for their friendship and camaraderie.

Declaration of Authorship

I, Oluniyi Ogunbiyi, declare that all the material contained in this thesis entitled "Contextual and Ethical Issues with Predictive Process Monitoring" is my own work. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- Part of this work has been published as:

1. Ogunbiyi, O., Basukoski, A. and Chaussalet, T., 2020. Comparative analysis of clustering-based remaining-time predictive process monitoring approaches. *International Journal of Business Process Integration and Management*.

2. Ogunbiyi, N., Basukoski, A. and Chaussalet, T., 2020. Investigating Social Contextual Factors in Remaining-Time Predictive Process Monitoring—A Survival Analysis Approach. *Algorithms*, *13*(11), p.267.

3. Ogunbiyi, N., Basukoski, A. and Chaussalet, T, 2021. Incorporating Spatial Context into Remaining-Time Predictive Process Monitoring. In *Proceedings of ACM SAC Conference, Virtual Event, South Korea, March 22- March 26, 2021 (SAC'21)*, 8 pages.

4. Ogunbiyi, N.; Basukoski, A.; Chaussalet, T. Investigating the Diffusion of Workload-Induced Stress—A Simulation Approach. *Information* **2021**, *12*, 11.

5. Ogunbiyi, N.; Basukoski, A.; Chaussalet, T. An Exploration of Ethical Decision Making with Intelligence Augmentation. *Soc. Sci.* **2021**, *10*, 57

I use the pronoun "we" throughout the thesis to acknowledge the contribution of my supervisory team comprising of Dr Artie Basukoski (Director of Studies) and Prof Thierry Chaussalet. Below is a breakdown of the contribution to the research project:

Niyi Ogunbiyi:

Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing—original draft preparation and editing, Visualisation and Project administration

Supervisory Team:

Validation, Resources, Review and Supervision

Dedication

Lovingly dedicated to:

My Nan, Mrs Maria Ogunbiyi…for your unconditional love

My parents, Professor Isaac and Mrs Olusanu Ogunbiyi…for instilling in me a love for learning

My wife, Sophie…for completing me

My children, Mikey and Imogen…for being my “Why”

CHAPTER ONE

1 INTRODUCTION

1.1 Motivation

Predictive process monitoring (PPM) has gained traction as a research field over the last decade, as evidenced by the steady increase in the number of related papers. (See Figure 1.1).

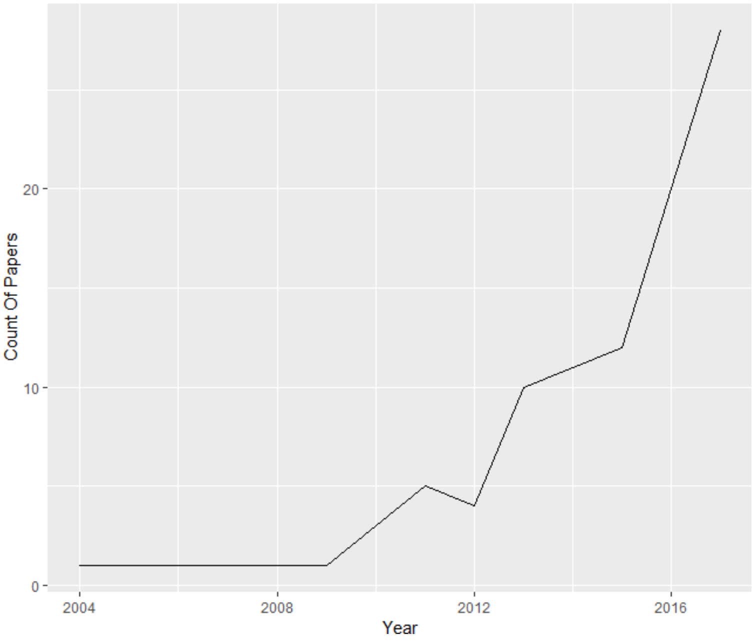


Figure 1.1 –Predictive Process Monitoring Papers by Publication Year

It is also an important topic from a practitioner perspective. For example, Johnston (2004) 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 (van Dongen, Crooy and van der Aalst, 2008). However, the path taken by the case may lead to it deviating from the average (e.g., because 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 remaining-time 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.

Effectively predicting process outcomes in operational business management is important for Customer Relationship Management (e.g., ‘will this customer’s order be completed on time?’), Enterprise Resource Planning (e.g., ‘what level of resourcing will be required to manage running cases/process instances?’) and Operational Process Improvement (e.g., ‘what are the common attributes of cases that consistently complete late?’), among others. Grogori et al (2004) propose a link between customer attraction and retention and “highly consistent and *predictable* quality” of process execution. Various approaches have been proposed to tackle this problem (see Panagos and Rabinovich, 1996; Eder et al, 1999; van Dongen, Crooy and van der Aalst, 2008). However, all these approaches have limitations particularly with regards to prediction accuracy.

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” (Metzger et al, 2015). 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 Evermann, Rehse & Fettke, 2017; Mehdiyev, Evermann & Fettke, 2017).

This thesis addresses contextual and ethical issues with predictive process monitoring. To illustrate how the issue of context is related to quality of prediction: in attempting to predict how long it would take to walk to a specified destination, we could collate data on the time it has taken other persons who have travelled that route to make the same journey by foot. However, that approach fails to consider other factors that might influence the journey time (e.g., difference in crowd density, road conditions, weather, etc at different points in time). We incorporate spatial and social context into the predictive process monitoring workflow with a view to increasing its predictive power. In addition, we also utilise workload (process context) to predict stress (social context) to discover the relationship between these contextual factors

With regards to ethical issues, one of the obstacles to the adoption of predictive process monitoring tools for operational support, is a lack of understanding of the factors that influenced the prediction. This is particularly the case when users are making decisions which may potentially cause harm or distress (e.g., rejecting a loan application) based on the prediction provided by the PPM tool. We propose a model describing how human agents utilise predictive process monitoring (and more generally, AI) tools to make decisions and make recommendations to assist them more easily identify where algorithmic discrimination may be present.

1.2 Background

We begin this section by addressing the *positioning*, *purpose,* and *requirements* of prediction in BPM.

Regarding *positioning*, Van der Aalst (2016:31) proposes a BPM lifecycle with four continuous phases (see Figure 1.2). Any process starts in the design phase, followed by implementation and configuration of the designed process. The implemented process is monitored and adjusted incrementally as required. However, if the process significantly fails to meet its critical requirements, it is often necessary to diagnose the root cause of problems and redesign the process. The literature positions prediction in the **design** phase (see Panagos and Rabinovich, 1996; Eder et al, 1999) and the **enactment/monitoring** phase (see Reijers, 2007; van Dongen, Crooy and van der Aalst, 2008) of the lifecycle.

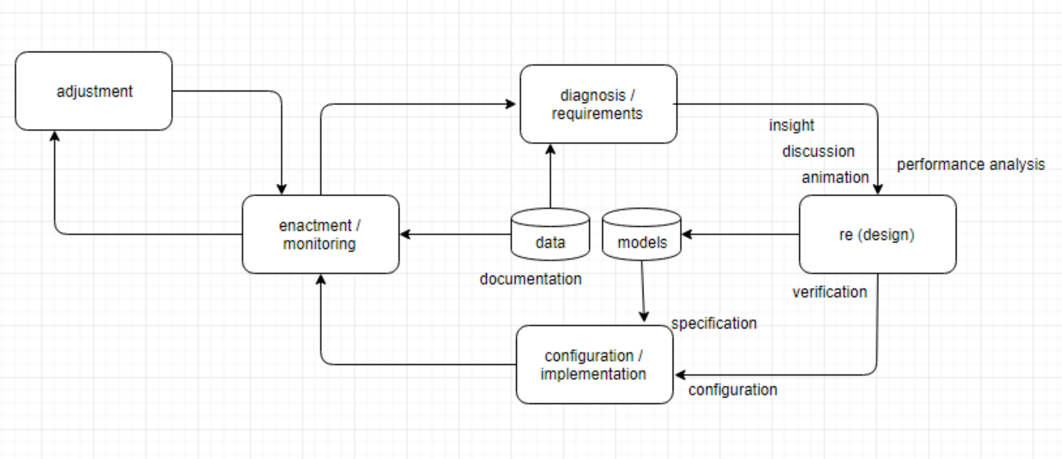


Figure 1.2 – BPM Lifecycle (Source: Van der Aalst, 2016:31)

The literature base also appears to indicate that the *purpose* of prediction differs depending on the phase in the lifecycle where it is made. For example, Van der Aalst (2013) posits that prediction at the enactment/monitoring phase is useful for operational decision making ("solving the concrete problem at hand") as opposed to an "abstract future problem" which is often the focus of design time prediction. A similar distinction is made between design and run time prediction by Reijers (2007). This explains why we adopt the phrase "predictive process *monitoring*", as it indicates the purpose and phase of the prediction we will be investigating. This paper outlines four *requirements* that an effective operational process predictive model must satisfy – accuracy, nearly instantaneous results, ease to use, non- interference with the efficient operation of the BPMS. Accuracy is suggested as the most important requirement based on earlier research undertaken by Yokum and Armstrong (see Yokum and Armstrong, 1995).

Di Francescomarino et al (2018) presents a comprehensive survey of the predictive process monitoring research field. They categorise approaches by the prediction target and the type of algorithm used for prediction.

In terms of prediction target, the authors identified three main categories of prediction target namely: numeric or continuous targets (e.g., remaining time, cost), categorical targets (e.g., risk class, outcome of a case) and activity sequence (e.g., next activity in a trace). In terms of types of algorithms utilised, two main groups of approaches were identified, namely: those that rely on an explicit model (e.g., annotated transition systems) and those that leverage machine learning and statistical techniques (e.g., regression and classification models, etc).

In the first half of the thesis (chapters 2-4), we focus on using a variety of machine learning techniques to predict the remaining time for an inflight trace. As mentioned in section 1.1, accurately predicting the remaining time for process instance is extremely valuable for customer relationship management (e.g., notifying a customer when their case is likely to complete), Just-In-Time production (e.g., to facilitate timely ordering and delivery of required components and services), etc. Though most of the approaches we propose can be easily adapted to predict other numeric or categorical targets, we chose to focus on a single target (. i.e., remaining time) for the sake on parsimony in evaluating the results of our experiments. For example, including outcome-based prediction in our evaluation would have required adopting a different set of metrics for evaluating the results (e.g., AUC, ROC).

We aim to improve current approaches by addressing the issue of lack of context in the predictive process monitoring workflow. Van der Aalst (2016: 318) identifies four pertinent contextual types:

Case context – These are the properties or attributes of case. In this research study, example of these would include request type (i.e., internal vs external), request category (e.g., parking/recreation services, etc). It is common for enhanced due diligence to be applied for certain cases e.g., additional approval required for high-risk cases or amounts above a specified threshold. Using case context to stratify cases on this basis will improve the predictive power of the model. This is the easiest service type to incorporate into the process model as the case attributes are often contained in the event log. Using tools such as decision trees, it is relatively straightforward to find out if there is a relationship between a response variable (e.g., completion time) and a particular case attribute. van der Aalst et al (2007) uses the case perspective to explore whether there is a relationship between the length of time it takes to settle an invoice and the invoice amount. However, it could be argued that the same result could also be answered using traditional data mining approaches i.e., process mining is superfluous.

Process context – This considers similar cases that may be competing for same resources. Schellekens (2009) showed that case interaction and the availability of resources are significant factors to consider when predicting completion time for cases. Another key process contextual factor that should be considered for prediction purposes is the current backlog (or workload).

Social context - This encompasses the way human resources collaborate in an organisation to work on the process of interest. As event logs often capture the details of the resource completing an activity, process mining is suited to organisational/social analysis and there has been some research focus on this area. For example, Nakatumba & van der Aalst (2009) investigated Yerkes-Dodson law of arousal to discover the point where the performance of a worker under time pressure (stress) degrades. van der Aalst et al (2007) derived the relationship between workers from the frequency by which they pass work to each other. Some process mining tools (e.g., PRoM) include the capability to perform Social Network

Analysis with a view to discovering interaction patterns amongst workers, evaluating the role of an individual in social network (e.g., centrality scores), etc. However, we were unable to locate evidence of research that directly explored the impact of social context (e.g., subcontracting) on prediction and where it has been touched on (e.g., see van der Aalst et al, 2007), the analysis has been tangential.

External context – This refers to factors in the wider ecosystem that impacts the process. In the case of this study, that will likely include factors such as the weather, legislation, location, etc. This is likely to be the most difficult context type to incorporate into the predictive model as the data is likely to be located outside the system. In addition, the cause-effect chain between external context and process outcome is likely to be difficult to establish as a certain factor can mediate, moderate, or mitigate another. For example, departmental budget reduction (external context) may impact staffing levels (process context) irrespective of service demand. Finally, a certain external context factor may not be constant over the lifecycle of a process instance (e.g., heavy snowfall might lead to a spike in demand for certain services; however, by the time these cases are completed, this external contextual factor may no longer apply)

In chapter 2, we focus on an important subset of remaining-time approaches i.e., clustering-based approaches. These are a highly interpretable class of approaches that group or bucket identical traces in the training set and build predictive models for each cluster. The clustering and training of these traces are often done during an offline phase. Subsequently in the online phase, an inflight trace is assigned to an appropriate cluster and the predictive model for that cluster is used to predict the remaining time for that trace. As we will discuss in chapter 6, interpretability is essential for engendering trust in predictive process models, hence we chose to focus on this subset of approaches as opposed to recent approaches (e.g., neural networks) which have gained in popularity recently but are not as interpretable.

In chapter 5, we introduce a new prediction target for the predictive workflow: workload-induced stress. As we show in that chapter, the literature indicates that increasing workload is a causal factor for stress in the workplace. The processing speed for activities provides an indicator when workload-induced stress is at its peak. We also examine the concept of emotional contagion – the phenomena of having one person's emotions trigger emotions and related behaviours in others – based on interactions between them. As all the required data (. i.e., workload, processing speed, duration of interaction, etc) can be derived from an event log, we are able to model the effect of workload (process context) on stress (social context). This enables us to better understand the relationship between these contextual factors but also adds a new prediction target to the predictive process mining knowledge base.

Finally in chapter 6, we examine ethical issues regarding the use by human agents of predictive process monitoring tools to make decisions. This primarily applies to outcome-based approaches where a user is influenced by the prediction provided by the tool e.g., decides to reject a loan application due to a prediction that the applicant is highly likely to default. However, where the predictive model is trained on biased data (e.g., data that captures historic discrimination based on features such as race, gender, etc), it's predictions may reflect those biases. Even when these features are removed, the predictive model can still infer these features. We develop a model that describes how human agents make ethical decisions when using predictive process monitoring (and AI) tools and make recommendations to design tools that assist the user determine whether algorithmic discrimination is present in the prediction.

1.3 Aims and Objectives

This research project aims to produce several context-aware predictive process monitoring models. As previously mentioned, the context in which an operational business process is executed has been acknowledged as having a significant effect on the predictive power of a predictive process model. Numerous studies have attempted to incorporate contextual factors into the process monitoring workflow (see Senderovich et al,2017;2019; Denisov, Fahland and Van der Aalst, 2019). However, we are not aware of any studies that assess the relative importance of these factors. This study will aim to address that gap by proposing novel techniques to incorporate relevant contextual factors into the predictive process monitoring workflow. In addition, it will examine the effect that certain contextual factors, have on the predictive power of the model.

Closely related to the aims above is the nature of the relationship between contextual types. This research will aim to shed some light on the nature of the relationship between process and social context.

Finally, we aim to uncover the manner human agents use the predictions generated by predictive process monitoring and other Artificial Intelligence (AI) workflows. We focus on the ethical issues that may arise from these and make recommendations for designing tools that aid human agents to detect algorithmic bias and thus make ethical decisions.

Specifically, the following Research Questions will be addressed:

Chapter 2:

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

How have these approaches been evaluated in the existing literature?

What is the relative performance of these approaches?

Answering these questions is important as it will provide a benchmark of existing clustering-based remaining time approaches which will assist researchers and practitioners to select which are likely to perform best for their predictive process monitoring workflows.

Chapter 3:

What is the relationship between social contextual factors and process completion time?

How does the survival analysis predictive process monitoring approach compare with existing approaches?

The answer to the first question above is vital for determining how to manipulate social contextual factors, specifically group centrality measures for the performers who execute a process instance, to ensure desired process outcomes are delivered. With regards to the second question, if it is proved that survival analysis predictive process monitoring approach compares favourably with other approaches, it advances the state of the art as it provides an approach that can include incomplete (censored) traces in training predictive models. This is essential as some event logs contain a significant proportion of incomplete traces e.g., in cloud systems where many virtual machines have very long lifetimes. Without an approach such as the survival analysis approach, these traces cannot be used to train the model and will have to be filtered out.

Chapter 4:

Do spatial features contribute to the predictive power of remaining-time predictive approaches vis-à-vis other features?

How does spatial-based remaining-time predictive process monitoring approaches compare with existing approaches?

To enable both questions above to be answered, a technique will need to be developed to incorporate spatial context into an event log to create a spatial event log. This opens opportunities beyond predictive process monitoring as it enables a new field of spatial process mining e.g., spatial conformance checking which includes spatial features in checking conformance to the specified process, spatial prescriptive process mining which provides recommendations to process performers based on location, amongst others. Subsequently, the findings from exploring both questions will provide an understanding of the extent to which spatial features improve the power of the predictive process monitoring model.

Chapter 5:

Does the relationship between workload and processing speed exhibit a quadratic relationship as proposed by the Yerkes-Dodson law?

If so, when does this relationship hold and when not?

Do network simulation approaches facilitate the discovery of successful interventions to mitigate the diffusion of workload-induced stress?

The answers to these questions advance the state of the art in several significant ways. Firstly, it provides new empirical evidence for the validity (or otherwise) of the Yerkes Dodson law, a law which was first proposed decades ago (Hebb, 1955). In addition, it sheds light on scenarios where the law holds and those where it doesn't. Finally, the findings enable the prediction of a workload-induced stress – a new and important target for predictive process monitoring workflows.

Chapter 6:

How do human agents make ethical decisions when using PPM (and more generally, AI) tools?

How can PPM tools be designed which assist human agents make ethical decisions?

Finally, the answers to the questions addressed in this chapter advance the understanding of the cognitive and affective processes users navigate as they make decisions with predictive process monitoring tools. The understanding acquired can subsequently serve as input into the design of predictive tools which make it easier for users to detect the presence of algorithmic discrimination and hence, make more ethical decisions.

1.4 Contributions

The systematic literature review contributed to the knowledge base by identifying existing clustering-based remaining-time predictive monitoring approaches (for the first time in the PPM literature base) and proposing a taxonomy for classifying these approaches. It also described a novel approach describing the implementation and execution of a systematic pre-review map (SPRM) step designed to ensure that a systematic literature review is not duplicative. Whilst this step has been proposed (see Brereton et al, 2007), we were unable to locate any studies that had implemented it.

Following on from the literature review, a further contribution to knowledge was made by our pioneering evaluation of the effect of the clustering approach on the performance of the predictive model.

Next, the novel aggregation encoding technique described in Chapter 3 enables us to incorporate social contextual factors into the predictive process monitoring framework and uncover relationships between these factors and remaining time. Besides, the survival analysis approach proposed in this chapter accommodates censoring in event data. Whilst this approach has been used many other fields (e.g., healthcare, marketing, credit risk analysis, etc), it is the first time it has been utilised in the predictive process monitoring workflow.

The innovative approach proposed in Chapter 4 facilitates the addition of spatial context into event logs to create spatial event logs. We also proposed a novel approach for encoding both the training and testing data which is a key enabler for utilising spatial context in the predictive process monitoring workflow. Finally, we demonstrate (for the first time in the PPM literature base) that spatial features improve the predictive power of the model and that spatial ensemble approaches yielded the best result for processes that are likely toexhibit spatial point processes. Based on this we make a strong case for the inclusion of spatial context in event logs, which is not currently typically collected.

In Chapter 5 we apply existing techniques (e.g., GAMs) in a novel way to uncover the nature of the relationship between workload(a key workplace stressor) and productivity from a couple of real-world event log. We identified the factors which drive this relationship (aka the Yerkes–Dodson law).

In the second part of that chapter, we utilised a simulation-based approach in an innovative manner to investigate the diffusion of workload-induced stress in the workplace. We found that in terms of stress management intervention, increasing the recovery rate yields better results vis-à-vis increasing the resilience of the workforce to stress.

In the penultimate chapter, we develop a new model of ethical decision making with AI Augmentation (. i.e., AI workflows with a human agent in the loop) that synthesises the fields of ethical decision making and Explainable AI (XAI). We propose nine belief statements based on the synthesis of the existing literature, observation, logic and empirical analogy which facilitate the design of AI tools which support ethical decision making.

1.5 Thesis Overview

The remainder of this thesis is organised as follows:

Chapter 2 provides the findings from a systematic literature review of remaining-time predictive process monitoring and a comparative analysis of an important subset of these: clustering-based approaches. Chapters 3 and 4 subsequently presents the results of an investigation of the impact of social and external contextual factors on the predictive process monitoring workflow. Subsequently, in Chapter 5 we explore the relationship between workload (a process contextual factor) and stress (a social contextual factor). The penultimate chapter contains an exploration of how human agents utilise the output from the various predictive workflows and ethical issues that might arise. Finally, Chapter 7 details the conclusions and proposed future research.

Figure 1.3 shows the relationship between the various chapters of the thesis in a diagrammatic format.

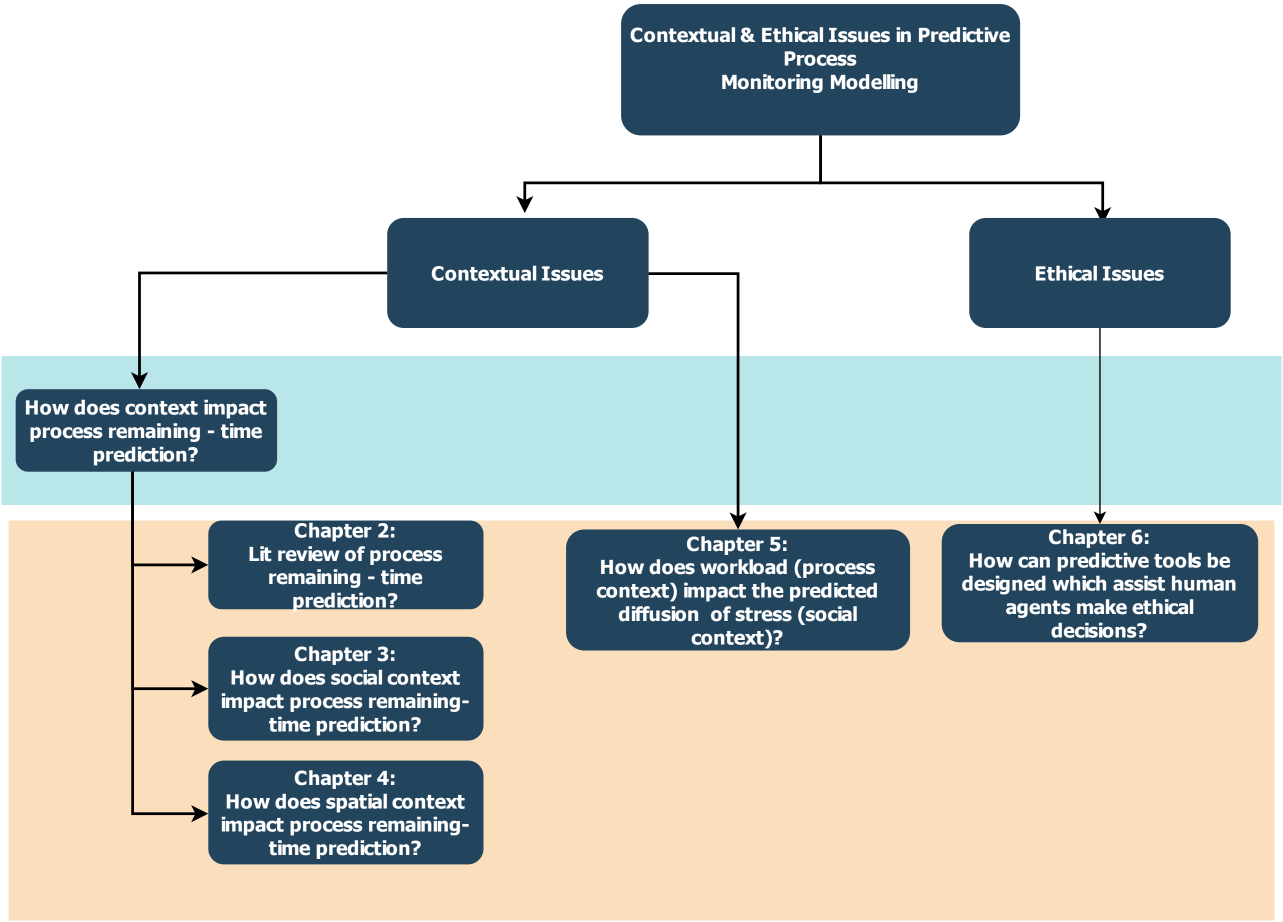

Contextual & Ethical Issues in Predictive Process Monitoring Modelling
Contextual Issues
Ethical Issues
How does context impact process remaining - time prediction?
Chapter 2: Lit review of process remaining - time prediction?
Chapter 3: How does social context impact process remaining-time prediction?
Chapter 4: How does spatial context impact process remaining-time prediction?
Chapter 5: How does workload (process context) impact the predicted diffusion of stress (social context)?
Chapter 6: How can predictive tools be designed which assist human agents make ethical decisions?

Figure 1.3 - Relationship between thesis chapters

CHAPTER TWO

2 COMPARATIVE ANALYSIS OF CLUSTERING-BASED REMAINING-TIME PREDICTIVE PROCESS MONITORING APPROACHES

2.1 Synopsis

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 chapter 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 a comparative analysis is performed to compare and benchmark the output of the identified studies using 5 real-life event logs

This chapter formed the basis of a journal paper accepted for publication in the *International Journal of Business Process Integration and Management*

2.2 Introduction

A critical step in the predictive process monitoring workflow is 'bucketing' (see Figure 2.1) 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 Teinemaa et al., 2017; Verenich et al., 2018). 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 monitoring 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 mapping (SPRM) step designed to ensure that a systematic literature review is not duplicative.

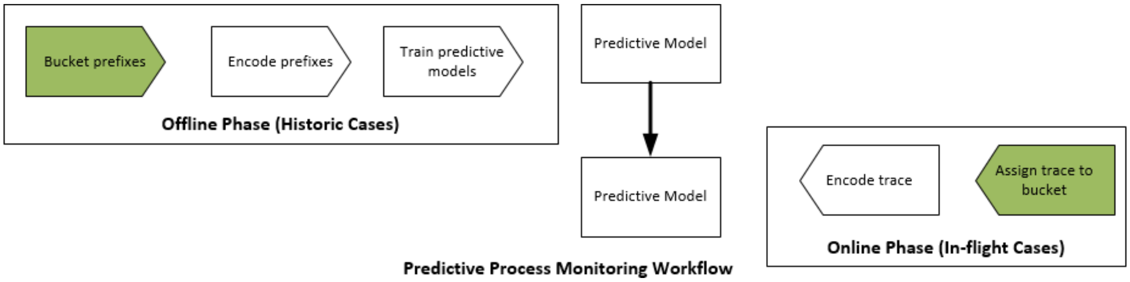


Figure 2.1 – Predictive Process Mining Monitoring Workflow

The remainder of the paper is structured as follows: Section 2.3 details preceding papers which have provided the motivation and methodological basis for this study. Section 2.4 defines key terms which will be built on throughout the paper. Section 2.5 describes the search methodology, including the inclusion/exclusion criteria. Section 2.6 details the clustering-based remaining-time predictive process mining approaches identified. Section 2.7 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.

2.3 Related Works

In terms of predictive process monitoring, Teinemaa et al. (2017) 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 Kitchenham (2004) 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 Verenich et al., 2018) 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.

Marquez-Chamorro, Resinas & Ruiz-Corts (2017) 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 proposed a taxonomy for these approaches. However, their paper does not perform a comparative analysis of these approaches. Metzger et al. (2015) detailed 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. Taleb (2017) 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 notappear 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

2.4 Background

2.4.1 Definitions

Several key terms to be built on throughout this review are formally defined.

Definition 2.1 *Event*. Let ε represent the event universe and T the time domain, A represent the set of activities and P represent the set of performers (i.e., individuals and teams).

An event e is a tuple (#case_identifier(e), #activity(e), #start_time(e), #completion_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), #completion_time(e)), it is common and desirable to have additional attributes such as #performer(e) indicating the performer associated with the event and #trans(e) indicating the transaction type associated with the event, amongst others. For each of these attributes, there is a function which assigns the attribute to the event .e.g. attr$_{\text{start\_time}} \in \varepsilon \rightarrow T$ assigning a start time to the event, attr$_{\text{completion\_time}} \in \varepsilon \rightarrow T$ assigning a completion time to the event, attr$_{\text{activity}} \in \varepsilon \rightarrow T$ assigning an activity label to the event and attr$_{\text{performer}} \in \varepsilon \nrightarrow P$, a partial function assigning a performer (or resource) to events. Note that attr$_{\text{performer}}$ is a partial function as some events may not be associated with any performers.

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).

Definition 2.2 *Terminal activities.* Let Z ⊆ A represent the set of valid terminal activity labels.

e_n is a valid terminal event if #activity_label(e_n) ∈ Z . This event indicates a 'clean' completion of the process instance. Otherwise, the process instance is still in-flight or abandoned.

Definition 2.3 *Trace*. Let ε^* represent the set of all finite sequences over ε. A trace is a (time-increasing) sequence of events, $\sigma \in \varepsilon^*$ such that each event appears only once. i.e. for $1 \leq i < j \leq |\sigma|$: $\sigma_i \neq \sigma_j$ and $\bar{\sigma}_i \leq \bar{\sigma}_j$

A partial trace (σ^p) has a non-valid terminal event as the final event (e_n). It indicates an in-flight (pre-mortem) process instance.

A full trace (σ^f) ends with a terminal event (e_n). It details the journey through the value chain that the process instance followed and indicates a completed (post-mortem) process instance.

Definition 2.4 *Event log.* An event log is a set of traces (full and partial) $L \subseteq C$ for a particular process such that each event appears at least once in the log.

Definition 2.5 *Remaining time.* Let σ^f represent a full trace, $\tau.e_n$ represent the completion time associated with the terminal event, #completion_time(e_n), and t represents the prediction point. For $t < \tau.e_n$,the remaining time $\tau_{rem} = \tau.e_n - t$. It indicates the remaining time to completion of case/process instance. Note that predicting at or after the completion time (i.e. $t \geq \tau.e_n$) is pointless.

Definition 2.6 *Elapsed time.* Let σ^f represent a full trace, $\tau.e_1$ represent the start time associated with the start event, #start_time(e_1), and t represents the prediction point. For $t > \tau.e_1$, the elapsed time $\tau_{ela} = t - \tau.e_1$. It indicates the elapsed time from the start of case/process instance to the prediction time.

Definition 2.7 *Cycle time.* Let σ^f represent a full trace, $\tau.e_1$ represent the start time associated with the start event, #start_time(e_1) and $\tau.e_n$ represent the completion time associated with the terminal event, #completion_time(e_n), The trace cycle time $\tau_{cyc} = \tau.\ e_n - \tau.\ e_1$. It indicates the time taken to complete the process instance from start to finish.

To illustrate the terms above, consider a process for reporting and remediating defects to public goods, e.g., potholes, streetlight 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.

2.5 Search Methodology

This systematic review adopts a combination of the procedure proposed by Kitchenham (2004) and the enhanced procedure (see Brereton et al, 2007).If a recommended step in the procedure is omitted, justification will be provided for the omission.

2.5.1 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?

Brereton et al (2007) 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.

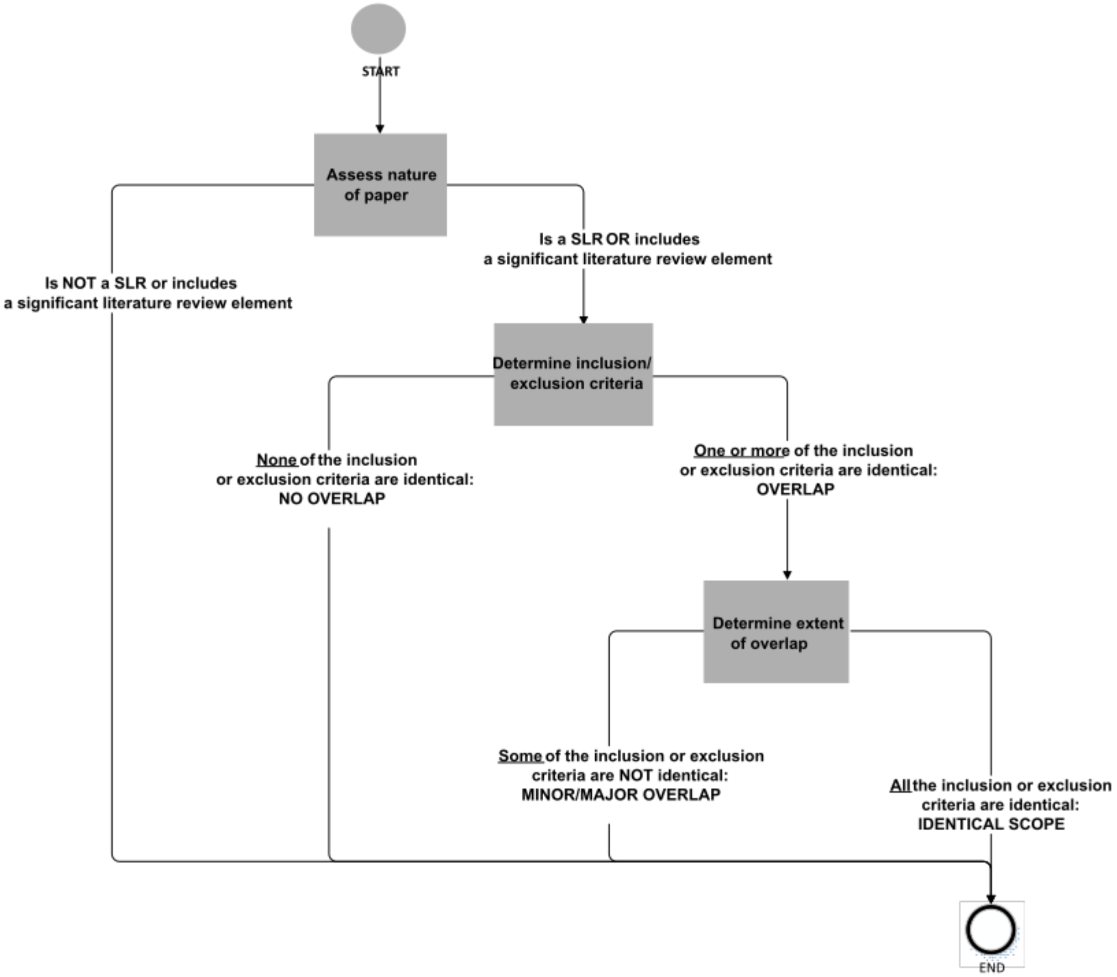

START
Assess nature of paper
Is NOT a SLR or includes a significant literature review element
Is a SLR OR includes a significant literature review element
Determine inclusion/ exclusion criteria
None of the inclusion or exclusion criteria are identical: NO OVERLAP
One or more of the inclusion or exclusion criteria are identical: OVERLAP
Determine extent of overlap
Some of the inclusion or exclusion criteria are NOT identical: MINOR/MAJOR OVERLAP
All the inclusion or exclusion criteria are identical: IDENTICAL SCOPE
END

Figure 2.2 - Systematic Pre-Review Mapping (SPRM) Process

Figure 2.2 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 2.5.2) 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 Verenich et al., 2018)

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.

2.5.2 Identify Relevant Research

Though Brereton et al (2007) 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 Teinemaa et al.,2017; Google Scholar Help, n.d.). 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 Gusenbauer (2019) 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 Anders & Evans, 2010; Freeman et al., 2009; Gehanno, Rollin & Darmoni, 2013).

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

2.5.3 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 2.5.1 , 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.

2.5.4 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 papers (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. Teinemaa et al. (2017) and

Marquez-Chamorro, Resinas & Ruiz-Corts (2017) 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 because of this threshold. Marquez-Chamorro, Resinas & Ruiz-Corts (2017) attempted to address this risk by relaxing this constraint for papers published between 2015 and 2017; however, we took a decision not to include a citation threshold to eliminate this risk

2.5.5 Select Primary Studies

Kitchenham (2004) 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. Teinemaa et al. (2017) 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"

We decided to adopt the same approach as Kitchenham (2004) as there were several challenges with implementing the approach adopted in Teinemaa et al. (2017). 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 2.3 shows a PRISMA Flow Diagram which depicts the flow of information through the different phases of the systematic literature review.

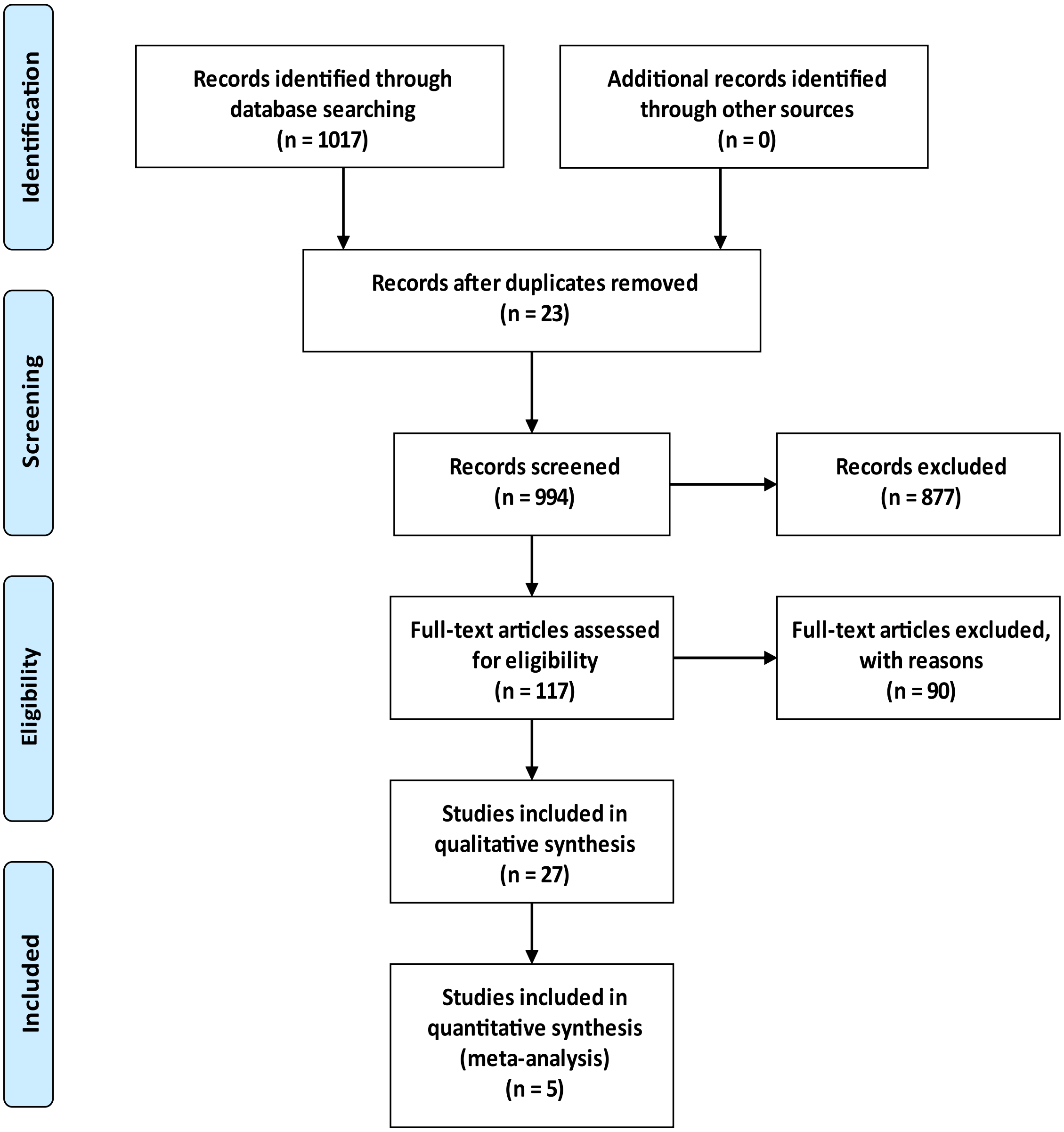

Identification
Screening
Eligibility
Included
Records identified through database searching (n = 1017)
Additional records identified through other sources (n = 0)
Records after duplicates removed (n = 23)
Records screened (n = 994)
Records excluded (n = 877)
Full-text articles assessed for eligibility (n = 117)
Full-text articles excluded, with reasons (n = 90)
Studies included in qualitative synthesis (n = 27)
Studies included in quantitative synthesis (meta-analysis) (n = 5)

Figure 2.3 – PRISMA Flow Diagram

2.5.6 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

2.5.7 Synthesis data

Kitchenham (2004) recommends meta-analysis on the extracted data utilising, amongst others, statistical methods. One of the critical problems with conducting this analysis as highlighted by Marquez-Chamorro, Resinas & Ruiz-Corts (2017), 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. Teinemaa et al. (2017) also calls out this problem and addresses it by implementing an evaluation tool against which it benchmarks eleven outcome-based prediction approaches. A similar tool for evaluating remaining-time clustering-based approaches was implemented in R. The results of the evaluation are detailed in Section 2.7.2.

2.5.8 Assess Study Quality

Kitchenham (2004) 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 2.9.

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 Wallace, Jankowicz. & O'Farrell, 2016), by evaluating criteria such as the number of data sets utilised, 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 2.9 discusses threats to the predictive process modelling validity in additional detail.

2.6 Discussion

As earlier mentioned in Section 2.5.4, the systematic review revealed five clustering-based remaining time predictive process monitoring papers in scope. Table 2.1 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 2.1)

Table 2.1 - List of the clustering-based remaining time predictive process monitoring approaches

Clustering Approach	**Short Title**	**Reference**
Centroid-based	context-aware	Folino, Guarascio & Pontieri (2012)
	low-level logs	Folino, Guarascio & Pontieri (2014a)
Hierarchical	fix-time	Folino, Guarascio & Pontieri (2014b)
Distribution-based	cloud-based	Cesario et al. (2016)
Association Rules	data-driven	Bevacqua et al. (2014)

Two papers, Folino, Guarascio & Pontieri (2012) and Folino, Guarascio & Pontieri (2014a) adopt the ***centroid-based*** approach.

Folino, Guarascio & Pontieri (2012) 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 could adapt to context changes. However, the approach omitted certain contextual factors (e.g., environmental factors) nor did it deal with concurrent behaviour effectively.

Folino, Guarascio & Pontieri (2014a) 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.

Folino, Guarascio & Pontieri (2014b) 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 Cesario et al. (2016) which adopts a ***distribution-based*** clustering method, traces are clustered utilising a probabilistic clustering algorithm. A non-parametric 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, Bevacqua et al. (2014) 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 utilising 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".

2.7 Benchmark

2.7.1 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. 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 2.2 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.

Table 2.2 – Event Log Overview

	BPIC 2012	**BPIC 2014**	**BPIC 2017**	**BPIC 2018**	**BPIC 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	Manufactu ring
Process	Loan Application	IT Service Management	Loan Application	Payments	P2P

2.7.2 Experimental Setup

Four of the five approaches were implemented in R. Folino, Guarascio & Pontieri (2014b) 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 Alelyani, Tang & Liu, 2013). 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 Boulesteix et al., 2018)

As with the methodology used in Verenich et al., 2018, the training & test set were not temporally disjoint

2.7.3 Accuracy

A survey of remaining-time predictive process monitoring approaches (i.e., including non-clustering-based approaches) revealed a variety of measures that assesses how accurate or effective the approach performed compared to specific benchmarks. Table 2.3 shows the distribution of the assessment measures utilised by the papers.

Table 2.3 – Accuracy Assessment Measures

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

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

$$RSME = \frac{1}{N} \sum_{i=1}^{N} (y_{\mathrm{i}} - \hat{y}_i)^2 \qquad (2.1)$$

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 Senderovich et al, 2017).

Another popular measure in the literature MAPE would be skewed towards the end of a case where remaining time tends towards zero (see Teinemaa et al, 2018). 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 Senderovich et al, 2017; Teinemaa et al, 2018). We adopt this evaluation approach through this research project.

The Mean Absolute Error (MAE) is defined as:

$$MAE = \frac{1}{N} \sum_{i=0}^{n} |y_i - \hat{y}_i| \qquad (2.2)$$

2.7.4 Earliness

Unlike the approach used by Teinemaa et al (2018), 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. However, as done by Teinemaa et al (2018), we measured both dimensions of accuracy & earliness

2.7.5 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 Xiong, & Li, 2013). 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 Dua & Chowriappa, 2012).

2.8 Results

Table 2.4 details the global MAE and Standard Deviation (SD) for each dataset/algorithm pair. Figure 2.4 displays the average ranking of each algorithm over the datasets with associated error bars calculated as the standard deviation of the rankings. 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). It is also worth noting that *data-driven* also has the second lowest error.

Table 2.4 - Global MAE ± SD

	BPIC 2012	BPIC 2014	BPIC 2017	BPIC 2018	BPIC 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

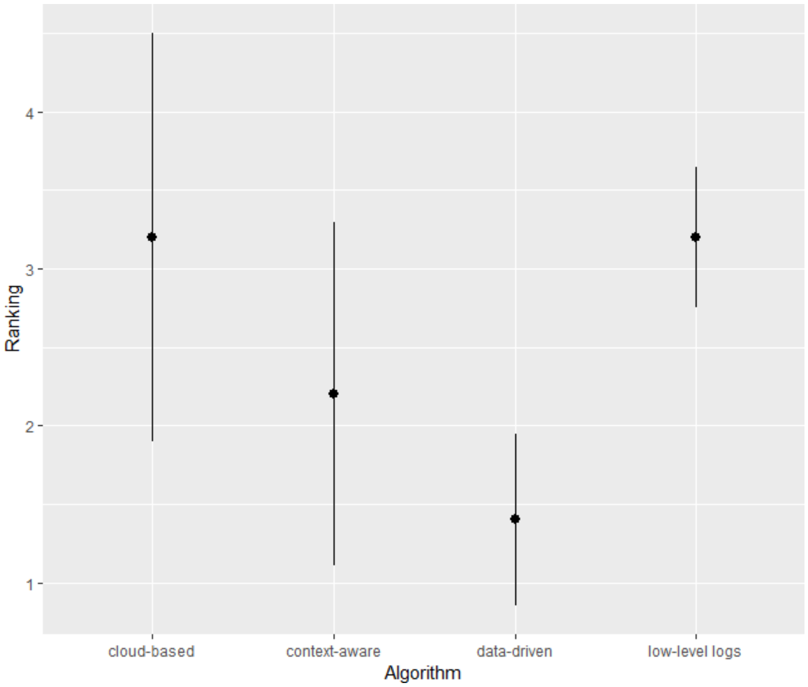


Figure 2.4 - Average Algorithm Ranking with associated error bars.

Figures 2.5 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 Verenich et al.,2018). *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.

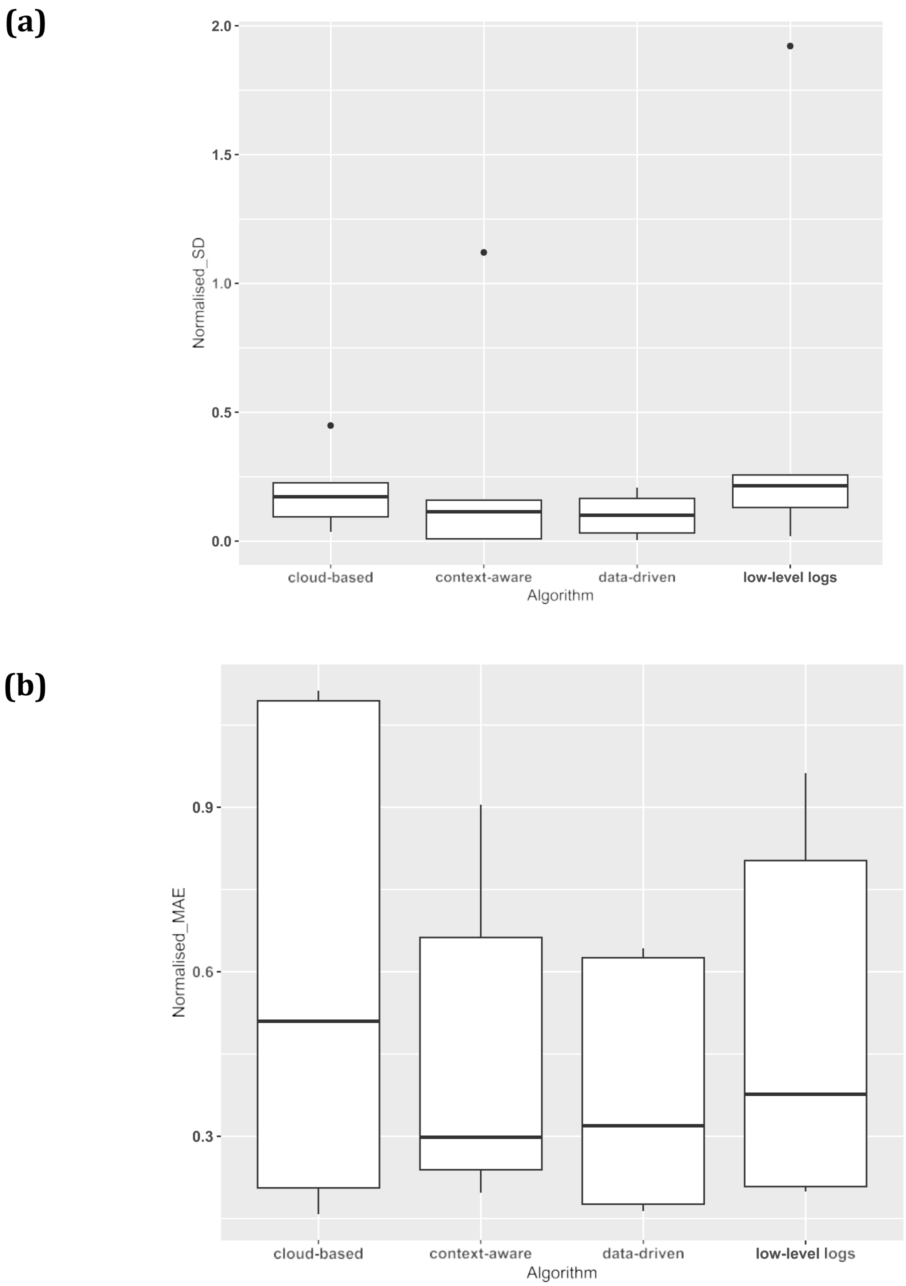


Figure 2.5– Average Normalised MAE (a) and standard deviation(b)

With regards to earliness, Figure 2.6 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 positive correlation appears to be consistent across algorithms though *context-aware* does display a weak negative correlation (see Table 2.5)

Table 2.5 - Pearson product-moment correlation coefficient between trace length and MAE

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

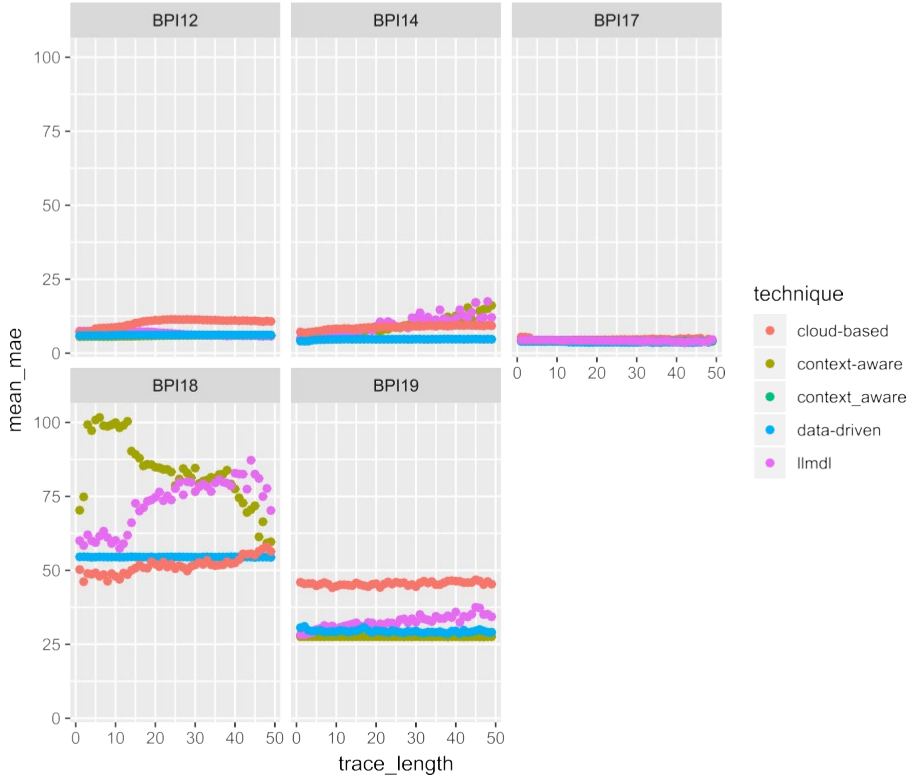


Figure 2.6 – Average MAE per Trace length

2.9 Threats to Validity

As mentioned earlier in Section 2.5.8, *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., Teinemaa et al, 2018) 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 Brereton et al. (2007) and Kitchenham (2004) (see Section 2.5) and fully documenting the approach. Besides, the initial literature base is made available for review and assessment

2.10 Summary

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.

We conclude this chapter by reflecting on the research choices we made on the selection of accuracy measures and exclusion criteria, among others. For example, we excluded outcome-based clustering approaches to keep the scope of the study manageable. In our opinion, this was the right choice as including outcome-based approaches in the assessment would have resulted in a less focused study and required the adoption of a mixture of remaining time and outcome-based accuracy measures (e.g., AUC, ROC, etc). However, we acknowledge that outcome-based predictive process monitoring studies have proposed several interesting clustering approaches. For example, the clustering approach proposed by Di Francescomarino et al (2015) clusters traces based on the string-edit distance between traces and subsequently predicts the outcome based on the payload of the last event in the trace. Such an approach could have been easily adapted to remaining-time predictions and in our opinion would likely have performed comparably with the best performing algorithm in our study (i.e., *data driven*) as they both cluster utilising the control flow perspective.

In terms of selecting remaining time accuracy measures, we chose MAE and rejected both RMSE and MAPE for the reasons outlined in Section 2.7.3. However, it may be argued that reliance on a single measure can miss important features and a single measure may not be appropriate in every context. For example, the RMSE may be more affected by outliers, but where data is very skewed a method with a lower RMSE will do a better job of minimising large errors in the tail. Thus, it may be argued that additional measures should be utilised to address these scenarios.

However, we believe that given the nature of the datasets in our study and the weaknesses inherent in both RMSE and MAPE, utilising only the MAE as we have done, was the correct choice.

In the following chapter, we will examine the impact of social contextual factors on the accuracy of remaining-time predictive process monitoring.

.

CHAPTER THREE

3 INVESTIGATING SOCIAL CONTEXTUAL FACTORS IN REMAINING-TIME PREDICTIVE PROCESS MONITORING – A SURVIVAL ANALYSIS APPROACH

3.1 Synopsis

Though social contextual factors are widely acknowledged to impact the way cases are handled, as yet there have been no studies which have investigated the impact of social contextual features in the predictive process monitoring framework. These factors encompass the way humans and automated agents interact within a particular organisation to execute process-related activities. This chapter seeks to address this problem by investigating the impact of social contextual features in the predictive process monitoring framework utilising a survival analysis approach.

We propose an approach to censor an event log and build a survival function utilising the Weibull model, which enables us to explore the impact of social contextual factors as covariates. Besides, we propose an approach to predict the remaining time of an in-flight process instance by using the survival function to estimate the throughput time for each trace, which is then used with the elapsed time to predict the remaining time for the trace. The proposed approach is benchmarked against existing approaches using five real-life event logs and outperforms these approaches

This chapter formed the basis of a journal paper published in *Algorithms*

3.2 Introduction

Earlier in Section 1.2, we highlighted the importance of contextual factors in predictive process monitoring and identified four pertinent contextual types, namely: case, process, social and external context

Figure 3.1 shows the relationship between the various contextual types.

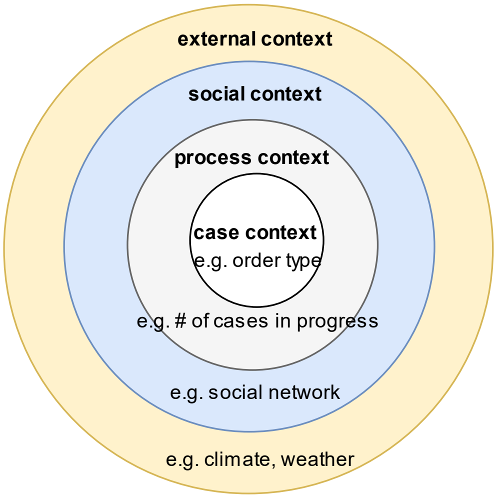


Figure 3.1 - Contextual Factors and Relationship (Adapted from van der Aalst, 2016:319)

To highlight the importance of social contextual factors, van der Aalst (2016:320) argues that "activities are executed by people that operate in a social network. Friction between individuals may delay process instance, and the speed at which people work may vary." It further adds that "process mining techniques tend to neglect the social context even though it is clear that this context directly impacts the way cases are handled". The same argument could be made about process monitoring. This study aims to address that gap by empirically investigating the impact of social contextual factors in the predictive process monitoring workflow.

To date, numerous approaches have been used to predict process remaining time, including Deep Learning (see Tax et al, 2017), Annotated Transition Systems (see Rogge-Solti & Weske, 2013) and Queuing Theory (see Bevacqua et al., 2014), among others. This study uses the survival analysis approach, also referred to as "time-to-event" analysis. This approach dates back to work by John Graunt published in his 1662 book 'Natural and Political Observations upon the Bill of Mortality' which suggested that the time of death should be considered an event that deserved systematic study (Akritas, 2004). While the majority of applications of the approach have been in the healthcare research field, by replacing the event of "death" with other events, the approach has been successfully applied in others fields such as Human Resources Management to determine the time-to-employee-attrition (see Somers, 1996), Marketing to model time-to-customer-churn (see Larivière & Van den Poel, 2004) and Credit Risk Management to determine the time-to-default (see Dirick, Claeskens & Baesens, 2017) However, as yet, time-to-event methods have not been utilised in predictive process monitoring. This study intends to address that gap by proposing an approach that uses survival analysis to predict the remaining time for process instances from several event logs. The primary advantage survival analysis offers over other approaches is that it can deal well with censored observations, i.e., observations where time to the event of interest is unavailable. This contrasts with standard regression approaches which tend to produce results that are neither accurate nor reliable if a high percentage of cases are incomplete. Besides, survival analysis approaches can handle covariates well. We propose a method for censoring event log for use in survival analysis by treating the completion of a process instance as the event of interest.

The remainder of the chapter is structured as follows: Section 3.3 details preceding studies which have provided the motivation and methodological basis for this study. Section 3.4 defines vital terms built on throughout the paper. Section 3.5 describes the proposed approach, while Section 3.6 details the evaluation results. The penultimate section describes the threats to the validity of the study while the final section summarises the findings.

3.3 Related Work

A review of the literature reveals three primary predictive process monitoring approaches: Model-based approaches (see Rogge-Solti & Weske, 2013; Verenich et al., 2017), sequence-to-feature encoding (STEP) approaches (see Cesario et al., 2016; Folino, Guarascio & Pontieri, 2012; Senderovich et al., 2017) and simulation-based approaches (see Rozinat et al.,2009; Veldhoen, 2011).

STEP approaches encode event log into feature-outcome pairs using a variety of techniques such as last state, aggregation, index-based or tensor encoding (see Senderovich et al., 2017; Verenich et al., 2017). However, it is worth mentioning a subset of STEP approaches that have become popular in recent years. i.e., neural-network-based approaches (see Tax et al, 2017; Breuker et al, 2016; Evermann, Rehse, and Fettke, 2017; Pasquadibisceglie et al, 2019). These state-of-the-art models make it relatively easy to include additional features into the prediction model. While most of these approaches focus on the next activity as the prediction target, the approach proposed by Tax et al, 2017 utilises an LSTM (a particular type of a Recurrent Neural Network) to iteratively predict the remaining activities till case completion and associated timestamps. This enables estimation of the remaining-time of the process instance.

With regards to social contextual factors, van der Aalst, Reijers & Song (2005) proposed an approach for discovering social networks from an event log and several metrics based on potential causality, joint cases/activities and special event types. They also apply these concepts to a real-life event log. In Song & van der Aalst (2008), the authors build on these and extend the approach to discover organisational models from event logs. In Nakatumba & van der Aalst (2009), the authors explored the relationship between the effect of workload and service time utilising regression analysison historical event log data. In Everett & Borgatti (1999), the authors extend the standard network centrality measures (degree, closeness and betweenness centrality) which had hitherto been applied to individuals to groups and classes as well.

With regards to survival analysis, Zhang & Thomas (2012) compares approaches that utilise linear regression and survival analysis to model loan recovery rate and amounts. The authors propose an approach to determine the optimal quantile for taking a point estimate from the survival curve. In Dirick, Claeskens & Baesens (2017), the authors extend that approach to predict the time-to-default for credit data sets from Belgian and UK financial institutions.

In this chapter, we utilise the STEP approach, together with a survival analysis technique to build a predictive process monitoring framework utilising the Weibull model.

3.4 Background

3.4.1 Definitions

3.4.1.1 Event, Traces and Event Logs

Several key terms to be built on throughout this chapter are formally defined. We build on the definitions from Chapter 2 (see 2.4.1) and adopt the standard notation defined in van der Aalst (2016:131)

Definition 3.1 *Censored traces*. Let C represent the set of all possible traces, $\mathrm{Z} \subseteq A$ represent the set of valid terminal activity labels and #censored(σ) represent a binary variable indicating whether a trace is censored or not respectively. A function $\mathrm{attr}_{\mathrm{censored}} \in \mathrm{C} \rightarrow \{1,0\}$ which assigns the appropriate value to a trace is defined as follows:

$$\#\mathrm{censored}(\sigma_{\mathrm{i}}) = \begin{cases} 1, & \#\mathrm{activity\_label}(\mathrm{e}_n) \in Z \\ 0, & \#\mathrm{activity\_label}(\mathrm{e}_n) \notin Z \end{cases} \qquad (3.1)$$

3.4.1.2 Survival Functions and Social Networks

Definition 3.2 *Survival Function.* Let *L* represent an event log with a set of trace cycle times {$\tau_{cyc.1}$…. $\tau_{cyc.n}$}, a trace $\sigma_i \in L$ with cycle time $\tau_{i\text{•}cyc}$ and a random time, t_r, the survival function S(t) = P ($\tau_{i\text{•}cyc} > t_r$). It gives the probability that the random time, t_r exceeds the trace cycle time

Weibull Model: Let T=t denote the time-to-completion of a trace σ_i, *f(t)* the probability density function of *T*, the probability density function of the Weibull model is given by

$$f(t) = \alpha\lambda t^{\alpha-1}\ e^{-\lambda t^{\alpha}} \tag{3.2}$$

where $\lambda > 0$ represents the trace completion rate parameter, and $\alpha > 0$ represents the scale or shape parameter

Definition 3.3 *Handover of work.* Let *P* represent the set of performers, *E* represents a set of directed edges and ϕ: $E \rightarrow V^2$ represent an incidence function mapping edges to vertices.

A handover-of-work graph is a directed multigraph permitting loops G =(*P*, *E*,ϕ). For our study, the incidence function maps the handover of work from one performer to another. A handover of work from performer *a* to performer *b* occurs if there are subsequent events (e_i and e_{i+1}) and *a* completes #activity(e_i), while *b* completes #activity(e_{i+1}). Note that the incidence function permits a performer to hand over work to themself. i.e. complete #activity (e_i) and (e_{i+1})

Definition 3.4 *Group Centrality Measures.* Let *L* represent an event log, *P* represents the set of performers and G represent a handover of work graph derived from the log. For a trace $\sigma_i \in L$, X= {# performer (e_1)…..# performer (e_n)}. This denotes the subset of performers who completed the activities ina trace.

Group Degree Centrality: The Group Degree Centrality of X is defined as follows:

$$\text{GD(X)} = |v: (u, v) \in E \wedge u \in X \wedge v \notin P|,\ X \subseteq P \tag{3.3}$$

If the group is defined as the subset of performers who worked on the trace, the group degree centrality specifies the number of non-group members that are connected to group members.

Group Between Centrality: Let $g_{u,v}$ represent the number of geodesics connecting vertices *u* to *v* and $g_{u,v(}X)$ represent the number of geodesics between *u* and *v* passing through some vertex of X. The group betweenness centrality of X is defined as follows:

$$\text{GB(X)} = \sum_{\{u<v\}} \frac{g_{u,v}(X)}{g_{u,v}},\ \text{u,v} \notin X \tag{3.4}$$

This measure shows "the proportion of geodesics connecting pairs of non-group members that pass through the group" (see Everett & Borgatti, 1999).

Group Closeness Centrality: The Group Closeness Centrality is defined as follows:

$$\text{GC(X)} = \frac{|P\backslash X|}{\sum_{v \in P\backslash X} dx_{,v}}, X \subseteq P \tag{3.5}$$

where dx_v denotes the distance between X and a vertex v defined as $dx_v = min_{u\ \in X}\ dist_{u,v}$ where $dist_{u,v}$ is the shortest path between u and v. This measures how close group members are to other non-group members

Group Eigenvector Centrality: Let X* represent a super vertex such N(X*)=N(x_1)UN(x_2)…UN(x_n) where N() denotes the neighbourhood of the vertex and x_i denotes the members of the set X. The group eigenvector centrality is defined as follows:

$$\mathrm{GE}(\mathrm{X}^*) = \frac{1}{\lambda} \sum_{j \in m_{(i)}} x_j \tag{3.6}$$

where M (i) is a set of neighbours of X* and λ is a constant

This is a measure of how connected the members of the group are to influential vertices outside the group.

To illustrate the terms above, we extend the example from Section 2.4.1 considering a process for reporting and remediating defects to public goods, e.g., potholes, streetlight outages. The set of valid activity labels is as follows:

{'Create Service Request', 'Initial Review', 'Assign Service Request', 'Assign Crew', 'Contact Citizen', 'Put Service Request On Hold', 'Close Service Request'}. The set of valid terminal activity labels for the process consists of the sole activity: {'Close Service Request'}. 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'}. This case would be considered 'non-censored' as the terminal event for the case was recorded in the event log. An example of a partial trace for a process instance would be {'Create Service Request', 'Initial View', 'Assign Service Request'}. This case would be considered 'censored' as the terminal event is missing from the trace. Note that, while this case is censored because it is in-flight, the same is true of cases that are abandoned, cancelled, withdrawn or fail to complete for any reason.

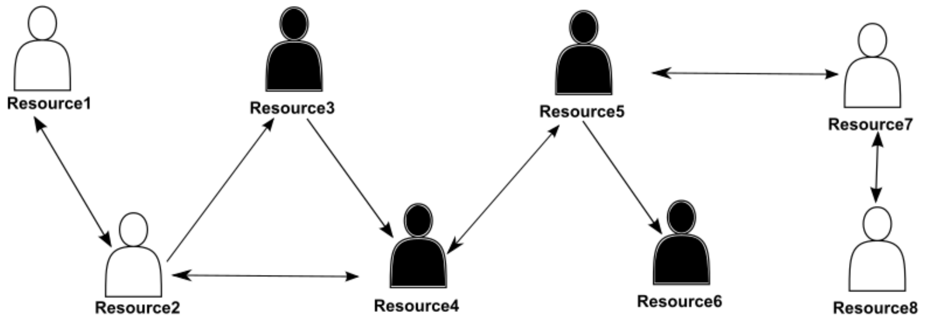


Figure 3.2 – Group Centrality Measures

To illustrate the group centrality measures, consider the handover-of-work graph depicted in Figure 3.2. The shaded performers (Resource3,4,5 and 6) represent the group of performers (hereafter referred to as the group) that executed activities on a selected trace σ, whilst the non-shaded performers did not work on the case.

The group degree centrality for the group is two as there are two non-group performers (Resource2 & 7) connected to the group. Note that, though there are two edges between Resource2 and the group, multiple edges are counted once.

As mentioned in Definition 3.4, the group betweenness centrality measures the proportion of geodesics connecting pairs of non-group members that pass through the group. The binary matrix in Table 3.1 displays *1* where the geodesic between the pair of non-group members passes through the group and *0* where it doesn't.

Table 3.1 – Geodesics passing through group of trace performers

	Resource1	**Resource2**	**Resource7**	**Resource8**
Resource1	-	0	1	1
Resource2	0	-	1	1
Resource7	1	1	-	0
Resource8	1	1	0	-

The group betweenness centrality measure for the trace is 0.66 (8/12).

The group closeness centrality for the trace is 6, i.e., the sum of distances from the group to all non-group performers. Table 3.1 displays these distances. Note that, whilst there are different options for calculating these distances (e.g., minimum, maximum or mean distance), we utilise the minimum distance for our calculation.

Table 3.2 – Distance to Group of Non-Performers

Non-Group Performer	**Distance to Group**
Resource1	2
Resource2	1
Resource7	1
Resource8	2
Σ Distance	**6**

To calculate the group eigenvector centrality, we treat the group as a large pseudonode, produce an adjacency matrix (see Table 3.3) and find the eigenvalues from which we calculate the eigenvector centrality.

Table 3.3 – Adjacency Matrix for Performers

	Resource1	**Resource2**	**Resource7**	**Resource8**	**Group**
Resource1	0	1	0	0	0
Resource2	1	0	0	0	1
Resource7	0	0	0	1	1
Resource8	0	1	1	0	0
Group	0	0	1	0	0

The set of calculated eigenvectors are {$\frac{\sqrt{5}-1}{2}$,-1, $\frac{-\sqrt{5}-1}{2}$, $\frac{\sqrt{5}\mp 1}{2}$,1}, giving an eigenvector value of 1 for the group.

3.5 Approach

3.5.1 Overview

Figure 3.3 provides an overview of the proposed approach used in building and evaluating the predictive model (see Section 3.5.3). The initial step is the determination of the set of terminal activity labels which indicate the ‘successful’ completion of a trace. This set serves as input into the censoring function, which outputs a log where each trace is deemed censored or otherwise. We subsequently encode the log traces and build a survival model using the censored log. We recommend that these steps are performed offline to improve runtime performance.

Subsequently, in the offline phase, the remaining time for in-flight cases are predicted using the survival model.

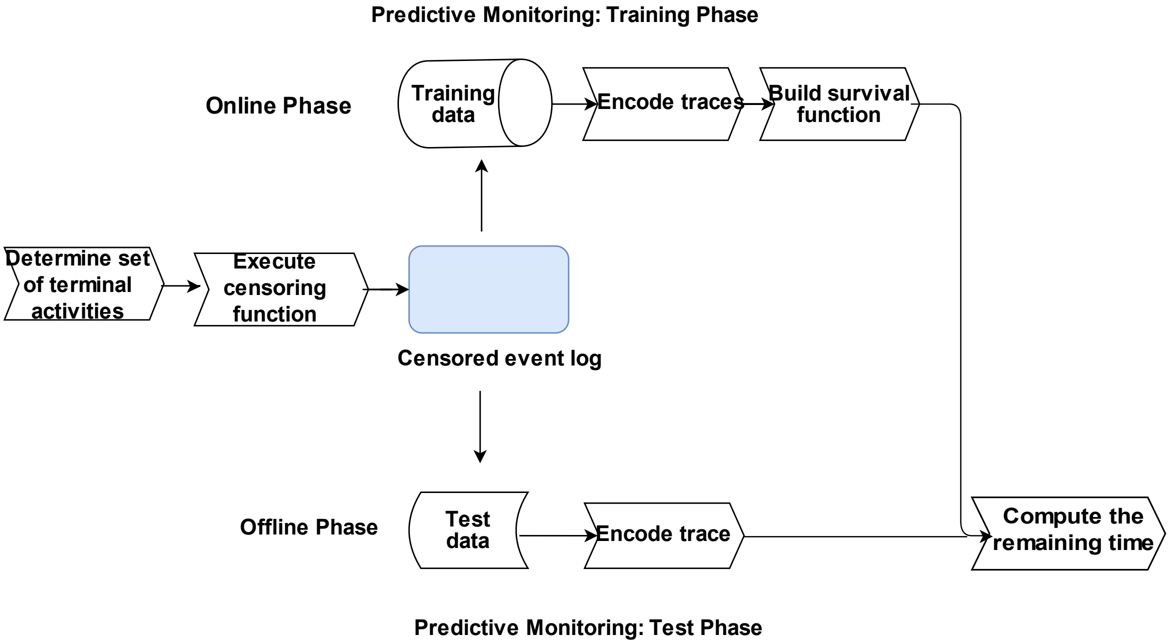

Predictive Monitoring: Training Phase
Online Phase
Training data
Encode traces
Build survival function
Determine set of terminal activities
Execute censoring function
Censored event log
Offline Phase
Test data
Encode trace
Compute the remaining time
Predictive Monitoring: Test Phase

Figure 3.3 - Overview of the proposed approach

3.5.2 Pre-Processing

To determine the set of terminal activity labels, we examined the process description and trace attributes of the respective datasets. However, in practice, this should be determined in conjunction with process subject matter experts. We utilised a function which examines whether the activity label associated with the terminal event for each trace, e_n (see Definition 2.2) is present within the set of terminal activity labels. If it is not, the trace is considered censored; otherwise, it is.

3.5.3 Predictive Monitoring

The approach consists of two phases: offline and online (see Figure 3.3). In the offline phase, the traces in the event log are encoded. We utilise a novel encoding approach, where we compute the grouped centrality measures for each trace (see Definition 3.4) using all the performers (#performer(e_1)…. # performer(e_n)) associated with that trace as well as the start and end event activity labels. While we adopt that approach to explore the impact of social contextual factors on process cycle time, we acknowledge that other encoding approaches, such as index-based encoding which is "lossless" (Verenich et al.,2017), could also be equally adopted. Our approach is in effect a combination of aggregation and last state encoding (Verenich et al.,2018) where the aggregation function computes the group degree (g_deg), betweenness(g_bet), closeness (g_clo) and eigenvalue (g_eig) centrality for each trace based on the set of performers who executed the events in that trace. This approach enables us to treat the performers who execute the activities in a trace as a team and builds on the approach in "the team effectiveness literature where researchers have used several internal team composition variables to predict performance" (Everett & Borgatti, 1999). Formally, given a trace σ_i = {$e_1 … e_n$} $\sigma \in \varepsilon*$ executed by a set of performers P, σ_i is transformed into a feature set {GC(σ_i), #activity(e_1), #activity(e_n), #censored(σ_i)} → τ_{rem} (σ_i) where GC(σ_i) = { $e'_1 … e'_n$} $\in \varepsilon*$ where, for $1 \leq i \leq n$, $e'_1 = e_i \odot (GC_P, u)$ with u = {$g\_deg_{e' \in \text{prefix}(e) \oplus \{e\}}$ e(P), $g\_bet_{e' \in \text{prefix}(e) \oplus \{e\}}$ e(P), $g\_clo_{e' \in \text{prefix}(e) \oplus \{e\}}$ e(P), $g\_eig_{e' \in \text{prefix}(e) \oplus \{e\}}$ e(P)} (De Leoni,xxx)

Table 3.4 displays the encoding for a couple of illustrative traces as described in section 3.4

Table 3.4 – Illustrative Trace Encoding

	group degree	group betweenness	group closeness	group eigenvector	event_1	event_n	censored	remaining time (days)
σ_m	2	0.66	6	1	Create Service Request’	Contact Citizen’	0	6.35
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	W.	.	.	.	.	.	.	.
σ_n	e1	0.1	3	0.4	Create Service Request’	Initial View’	1	5.73

We utilise the parametric Weibull model to build the survival model. Even though it requires that certain assumptions regarding the distribution of the process cycle time are satisfied, this method offers several unique benefits in that it is “simultaneously both proportional and accelerated so that both relative event rates and relative extension in” process cycle “time can be estimated”(see Carroll, 2003)

In the online phase, the in-flight traces are encoded utilising the same approach as in the online phase. The survival model built in the offline phase is used to estimate the total cycle time for the trace, and the remaining time for the trace is computed by subtracting the elapsed time from the estimated cycle time.

Algorithm 3.1 details the survival analysis predictive modelling algorithm

Algorithm 3.1 - Survival Algorithm

```
Input: An event log $L$ over some trace universe $\sigma$ with the associated
feature elapsed time $\tau_{ela}$, cycle time $\tau_{cyc}$, a target measure remaining
time $\tau_{rem}$, a set of terminal activity labels ($T$), an estimation quantile $q$
and a survival analysis (SURV) method

Output: A Survival Analysis predictive model ($SA$-$PM$) for $L$

Method: Perform the following steps:
  i.   Associate a binary variable #censored($\sigma$) with each trace $\sigma \in$ L using
       #activity($e_n$), T (see definition 3)
  ii.  Encode each trace using a suitable encoding function
  iii. Induce a survival function $sa$-$pm$ out of $L$ using method SURV
       {#censored($\sigma_i$), # cycle time($\sigma_i$) .....# attribute$_n$($\sigma_i$)} as input value
  iv.  Let $\sigma_1 \ldots \sigma_n$ denote each trace
  v.   For each $\sigma_i$ do
         a. Estimate the cycle time $\tau i._{cyc\_pred}$ for each trace from $sa$-$pm$
            utilising $q$
         b. Estimate the remaining time for each trace $\tau i._{rem\_pred}$ :
            $\tau i._{cyc\_pred}$ - $\tau_{ela}$
  vi.  End
  vii. Return c {$\tau_{rem\_pred1} \ldots\ldots \tau_{rem\_predn}$}
```

3.6 Evaluation

In this section, we perform two sets of experiments to address the research questions of interest in this study. In the first set of experiments, we evaluate the impact of social contextual features on the cycle time of process traces. In the second set of experiments, we evaluate the proposed survival analysis predictive monitoring approach against similar predictive monitoring approaches. Specifically, we seek to address the following research questions:

RQ4: What is the relationship between social contextual factors and process completion time?

RQ5: How does the survival analysis predictive process monitoring approach compare with existing approaches?

3.6.1 Datasets

Five real-life event logs from the Business Process Intelligence Challenge (BPIC) were used for the experiments as follows: BPIC 12 (van Dongen, 2012), BPIC 14 (van Dongen, 2014), BPIC 15(3) (van Dongen, 2015), BPIC 17 (van Dongen, 2017) and BPIC 18 (van Dongen & Borchert, 2018). The logs were from a variety of domains covering diverse processes. To manage memory requirements, a subset of each event log (except for BPIC 12 and 15(3) where the entire log was used) was selected for the analysis. The event logs were selected on the basis that information about the performer (i.e. the individual resource or team) that executed each event (#performer (e)) is present in the log. This enabled us to create the social networks required to address the research questions. For example, BPIC 20 (van Dongen, 2020) is not considered suitable as the performer attribute is abstracted to a high-level role (e.g., Staff member). Besides, basic feature engineering was performed to add required features such as elapsed time and remaining time to each log.

See Table 3.5 for a summary of the logs used for the experiments.

It is worth highlighting that unlike the other logs, BPIC 14 had resources information at the team levels, so the handover of work was computed at a team rather than at an individual level.

Table 3.5 - Event Log Overview

	BPIC 18	**BPIC 17**	**BPIC 15(3)**	**BPIC 14**	**BPIC 12**
# of events	267830	233928	59681	277577	262200
# of cases	3285	9453	1409	13985	13087
# of traces	3277	5211	1350	13942	4366
# of distinct activities	141	26	277	39	24
Mean trace length	81.53	24.75	42.36	19.85	20.04
Mean throughput time (days)	580.63	24.11	62.23	12.93	8.62
Throughput time SD (days)	580.62	14.893	97.64	27.94	12.13
Domain	PublicAdmin	Financial services	PublicAdmin	Financial services	Financial services

3.6.2 Experimental Setup

As input into both sets of experiments, we created a social network from the handover of work from one performer to the next in each trace. We subsequently created an adjacency matrix and computed four grouped centrality measures – degree, betweenness, eigenvalue and closeness based on the approach recommended in Everett & Borgatti (1999). In the initial sets of experiments, we performed some exploratory analysis by assigning each trace to a cluster using its four grouped centrality score. Before clustering the grouped centrality scores using a centroid-based clustering method (k-means), we empirically estimated the optimal numbers of clusters, k, from each dataset using the elbow method. We subsequently computed a survival curve for each cluster to visually determine how the grouped centrality scores impact the process cycle time. We subsequently created a case network to determine the relationship between these factors and case cycle times. We concluded by exploring the relationship between each group centrality measure and the cycle time.

For the second set of experiments, we implemented a survival analysis predictive monitoring approach named survival in R (as described in section 3.4) which enables evaluation of this approach vis-à-vis similar existing approaches. The code and data for the experiments are located in the GitHub repository: https://github.com/etioro/SocialNetworks.git. With regards to selecting the set of approaches to evaluate against, we considered only testing against the set of clustering-based approaches identified in chapter 2. However, we realised that this would limit the generalisability of the results as clustering-based approaches represent a subset of predictive process monitoring approaches. As such, we decided to widen the set of approaches included in the evaluation set. We evaluated the survival analysis approach against two clustering-based remaining-time approaches identified in the literature (see Cesario et al., 2016 and Folino, Guarascio & Pontieri, 2012) and a couple of methods which used a zero prefix-bucketing combined with a gradient boosting machine (*gbm*) and multilayer perceptron (*mlp*) neural network regressors respectively to predict the remaining time for each trace (Verenich et al., 2018). The same set of features were used to build all and evaluate all the models in the experiment.

We encoded the traces as described in section 3.4. Though this encoding is 'lossy', we adopt this approach as it enables us to capture the social contextual factors associated with each trace and adequately address the research questions.

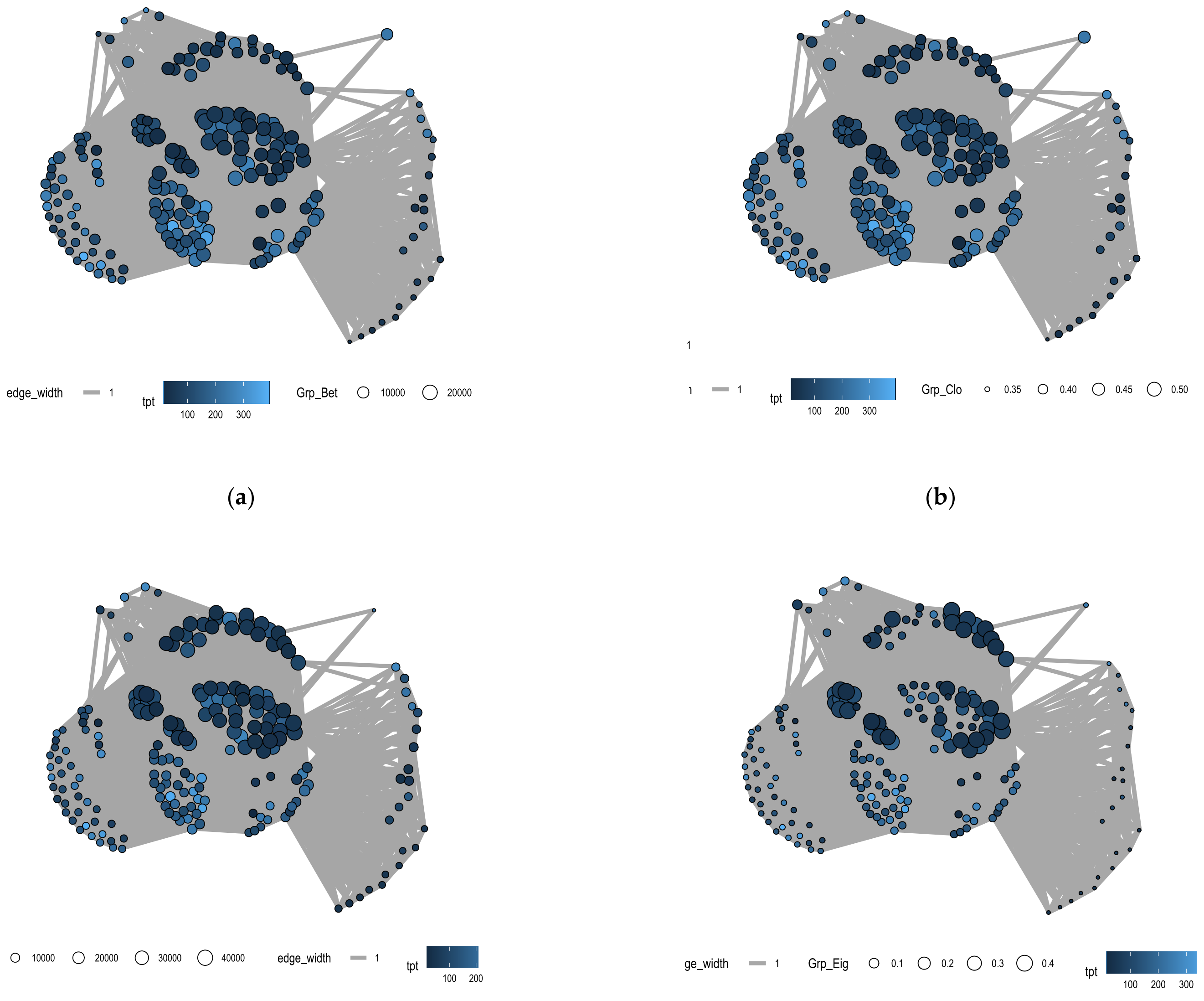
We split each event log into test and training sets (75:25 split, respectively), used the training set to build the survival function and the test set for making remaining-time predictions which are subsequently evaluated. As the survival curve gives a distribution of cycle time estimates, it was necessary to determine an optimal quantile for estimating the cycle time. We explored the approach suggested by Dirick, Claeskens & Baesens (2017) and Zhang & Thomas (2012) for selecting this quantile. This entails fitting a survival curve to the training set and determining which quantile minimised the MAE and RSME. This quantile is used to estimate the cycle time in the test set. However, we found that compared to the median, this method performed poorly; hence we used the median as the optimal quantile for estimating the cycle time.

As with the methodology used in Verenich et al. (2018), the training & test set were not temporally disjoint.

As discussed in Section 2.7.3, we chose to utilise the Mean Absolute Error (MAE) to evaluate the accuracy as other measures such as the Root Mean Square Error (RSME) are susceptible to outliers and Mean Percentage Error (MAPE) would be skewed towards the end of a case where remaining time tends towards zero. We filter the test set to use only non-censored traces to evaluate the MAE, as these are completed traces, whereas censored traces were abandoned or in-flight as at the time of log extraction.

3.7 Results

In the first set of experiments, to explore which of the group centrality factors have the most impact on the cycle, we created a network that visually illustrates the effect high and low values of the grouped centrality measures on case cycle times (see Figures 3.4 a,b,c and d below). Each node represents a case, with cases which share common performers connected and the edges weighted by the number of shared performers. The size of the nodes is proportional to the relevant grouped centrality measure. The network appears to show an inverse relationship between cycle time and group degree and eigenvector centrality while the opposite is the case with the group closeness and betweenness centrality measures



(**a**)

(**b**)

(**c**)

(**d**)

Figure 3.4 (a) BPIC 14 Group Between Centrality Graph; (b) BPIC 14 Group Closeness Centrality Graph; (c) BPIC 14 Group Degree Centrality Graph; (d) BPIC 14 Group Eigenvalue Centrality Graph.

We explore these apparent relationships between the group centrality measures and trace cycle times further. Table 3.6 shows the Spearman Rank Correlation between each group centrality measure and the trace cycle time (with all the values statistically significant at the 95% confidence level in bold font). This test was selected to determine the strength and direction of the monotonic relationship between these measures. The group closeness centrality is the most strongly correlated measure to the trace cycle time, followed by the group eigenvector centrality. The group betweenness and closeness centrality were generally positively correlated while the group eigenvector centrality was generally negatively correlated.

We delve into the team effectiveness literature to sheds some light on these results. As Everett & Borgatti (1999) posits, “maintaining strong ties with people outside the team is an important determinant of team success”. We argue that these results may have implications for team setup as they shed light on the nature of these “ties”. For example, many organisations create specialised cells or SWAT teams to handle certain types of cases e.g., complex cases. As a result, these teams could become isolated from other process performers which increases their probability of “failing” (see Ancona, 1990). The results would seem to imply that connecting the teams to other influential performers in the organisation (high eigenvector centrality) will result in shorter cycle times, perhaps because of the abilityof these performers to resolve issues relatively quickly. This would suggest that where such cells exist, it would be desirable to work cases with performers outside their cell periodically. Intuitively this will increase the sharing of knowledge and experience across the organisation.

On the other hand, lower group betweenness across groups appears linked to lower cycle times. This measure is a proxy for how much the group is becoming a bottleneck across the organisation; perhaps because the performers are perceived as possessing certain desirable traits, e.g., viewed as experts or dependable. Lower group closeness centrality (a measure of the distance of the group to other performers) is also correlated with lowerprocessing time as it indicates greater connectedness between performers.

Table 3.6 - Spearman Rank Correlation between group centrality measure and trace cycle time

	GB	**GC**	**GE**	**GD**
BPIC 18	**-0.058**	-0.014	**0.248**	**0.107**
BPIC 17	**0.063**	**0.176**	**0.063**	**0.093**
BPIC 15(3)	**0.289**	**0.414**	**-0.208**	-0.045
BPIC 14	**0.078**	**0.119**	**-0.183**	**-0.171**
BPIC 12	**0.877**	**0.848**	**-0.475**	-0.003

Progressing to the second set of experiments, Table 3.7 details the Global MAE and Standard Deviation (SD) for each dataset/algorithm pair. The performance of the algorithms is visualised in Figure 3.5, which displays the average ranking of each algorithm over the datasets with associated error bars, calculated as the standard deviation of the rankings. Over the five datasets, the survival analysis approach outperforms all the other approaches and has the lowest error.

Table 3.7 - Global Mean Average Error ± Standard Deviation

	Survival	**MLP**	**GBM**	**Cloud-based**	**Context-aware**
BPIC 18	166.27 ± 46.6	187.37±190.62	76.09 ± 84.18	217.27±162.30	212.34±124.37
BPIC 17	11.158 ± 2.03	12.11 ± 12.67	10.83 ± 9.54	12.18 ± 11.53	12.77 ± 11.23
BPIC 15(3)	23.91 ± 6.12	27.88 ± 40.91	29.07±33.63	42.26 ± 52.27	57.12 ± 59.31
BPIC 14	19.19 ± 11.6	20.78 ± 41.39	23.79 ± 36.37	26.03 ± 27.99	25.53 ± 31.84
BPIC 12	5.83 ± 1.95	8.24 ± 8.72	5.59 ± 5.27	9.55 ± 8.49	9.86 ± 9.12

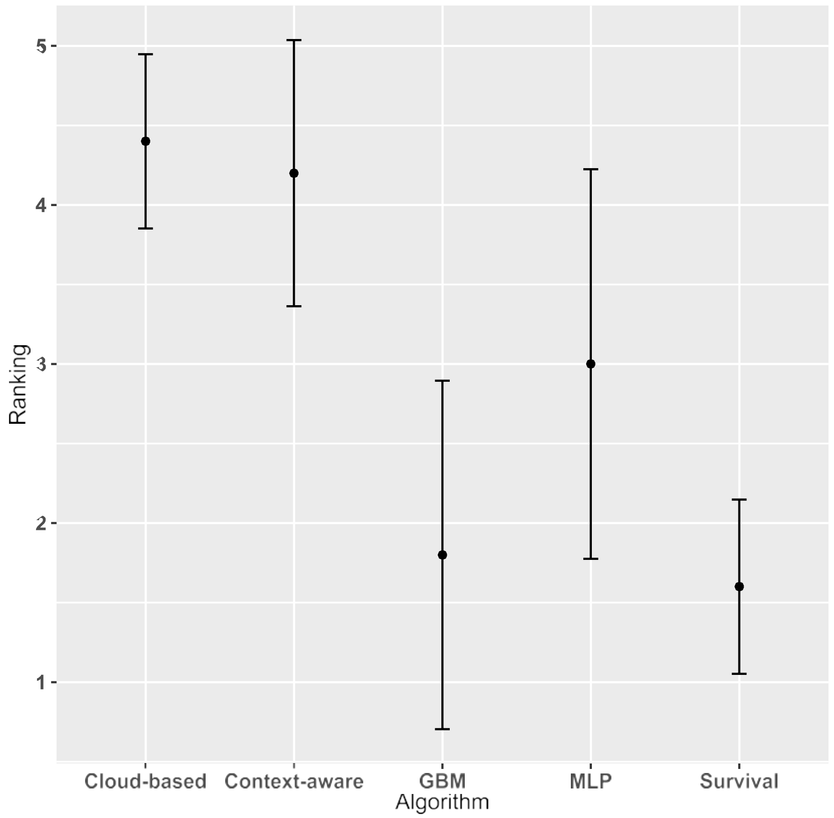


Figure 3.5 - Average Algorithm Ranking with associated error bars.

Figures 3.6 and 3.7 show 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 Verenich et al., 2018). The survival approach has the lowest normalised mean and median MAE (0.659 and 0.463, respectively) further confirmation of its superior performance.

As recommended by Demsar (2006), the non-parametric Friedman test was performed on the ranked data to determine whether there was a significant difference between the algorithms. The test results indicate a statistically significant difference between the various algorithms at the 95% confidence level (p=0.008687). To determine which algorithms differ from the other, we utilise the Quade post-hoc test to perform a pairwise comparison between the various algorithms. Table 3.8 shows the results of the pairwise comparisons (with all the values statistically significant at the 95% confidence level in bold font). The results indicate that the survival methods significantly outperformed all the existing methods except for *gbm* (see results in bold). To determine the explanation for this, we observe that event logs typically contain a portion of incomplete traces which are filtered out by existing approaches as they do not contribute any information towards accurately predicting the remaining time of the trace. Intuition supports this approach as we cannot determine whether an incomplete trace will finish in the next hour, day, or year.

Verenich et al (2018) provides a detailed discussion of generative and discriminative approaches for process monitoring. Discriminative approaches infer a conditional probability P(Y|X) from the training data set where X = (σ_1, σ_2....σ_n} denotes the set of feature variables and Y={ τ_{rem_pred1}, τ_{rem_pred2}..... τ_{rem_predn}} represents the prediction target. The resulting probability distribution is used to make predictions for the test set. However, when there is a significant proportion of incomplete traces in the training data, this approach is not useful as the target (Y). i.e., the remaining time for the trace, is unknown. This is the reason why these traces are typically removed from the training set. However, generative approaches, such as the survival analysis approach proposed, calculate a joint distribution P(X,Y) which is then utilised to derive the conditional probability P(Y|X). This approach can generate synthetic values of X by sampling from the joint distribution. As a result, this approach performs better when an event log has a significant proportion of incomplete trace.

In our experimental data, the percentage of incomplete traces ranged from 39% (BPIC 14) to 69% (BPIC 12). However, the survival analysis approach enables us to “account for [incomplete traces (i.e., censored data)] in the analysis” as this approach can extract information from them (Linden & Yarnold, 2017). This is the main advantage of the approach we propose as it delivers better accuracy for event logs with a significant proportion of incomplete traces

Table 3.8 - Pairwise comparisons using posthoc-Quade test

	Survival	**MLP**	**GBM**	**Cloud-based**
MLP	**0.04173**			
GBM	0.75592	0.07598		
Cloud-based	**0.00042**	**0.04173**	**0.00082**	
Context-aware	**0.00082**	0.07598	**0.00159**	0.75592

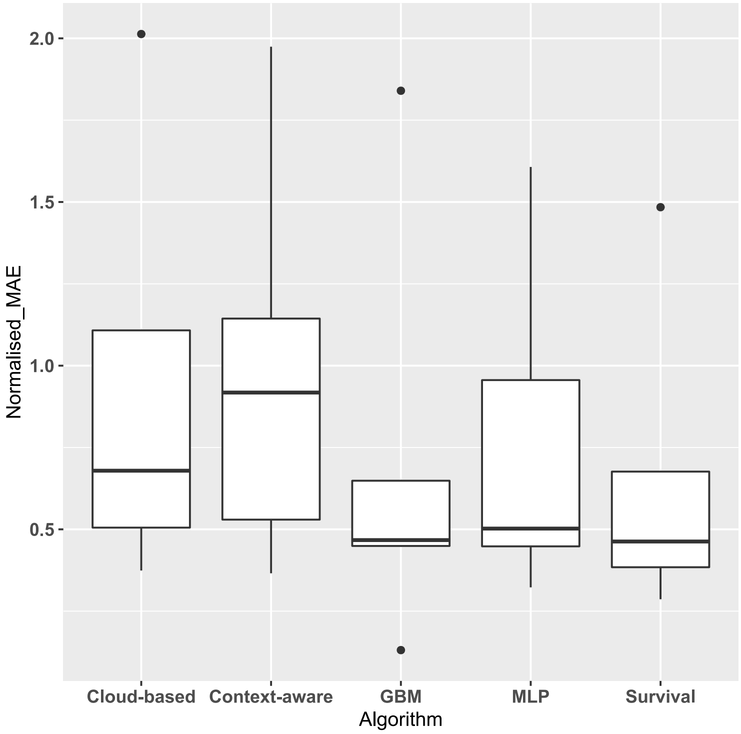


Figure 3.6 - Average Normalised MAE

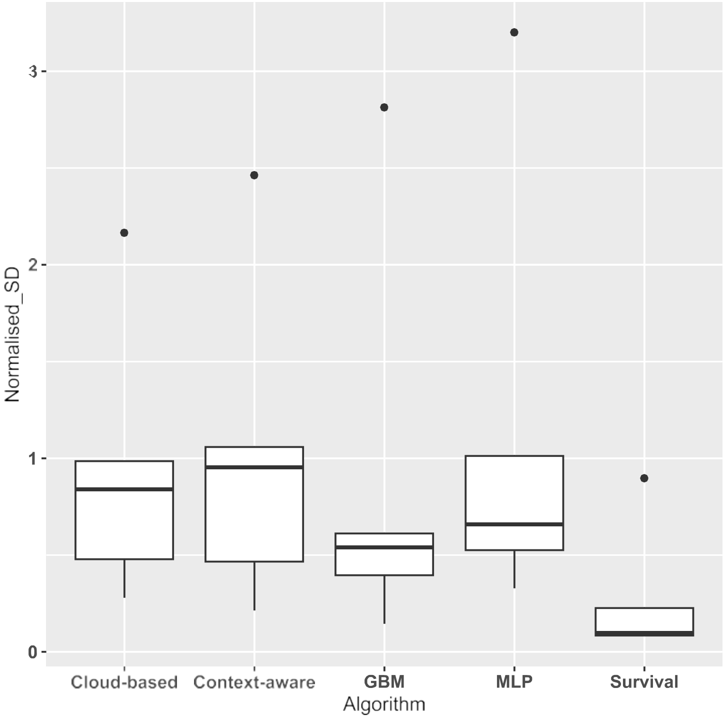


Figure 3.7 - Average Normalised Standard Deviation

To explore the effect of the proportion of incomplete traces on performance, we perform an additional set of experiment utilising a subset of data from a couple of event logs (BPIC12 and BPIC18), selected as they are on opposite spectrums of event log complexity (see van der Aalst, 2016:366). Keeping the size of the event log constant, we incrementally increase the percentage of incomplete traces in the log in steps of 20%, starting from 0% through to 100% (the baseline). We subsequently calculate the normalised MAE for each log using the proposed survival approach. Figures 3.8 and 3.9 display the plots of the normalised MAE by the proportion of incomplete traces in the event log. As expected, both plots indicate a dramatic improvement in performance as the proportion of complete traces in the log increases. However, we observe that this improvement begins to level off once the proportion of complete traces exceeds c.60 %, after which the gain is less significant.

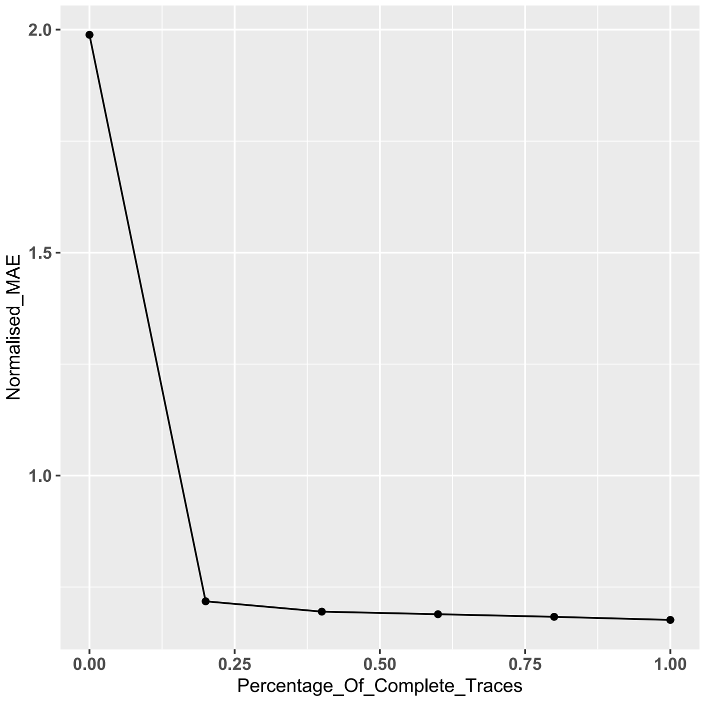


Figure 3.8 - BPIC 12 - Normalised MAE by proportion of incomplete traces

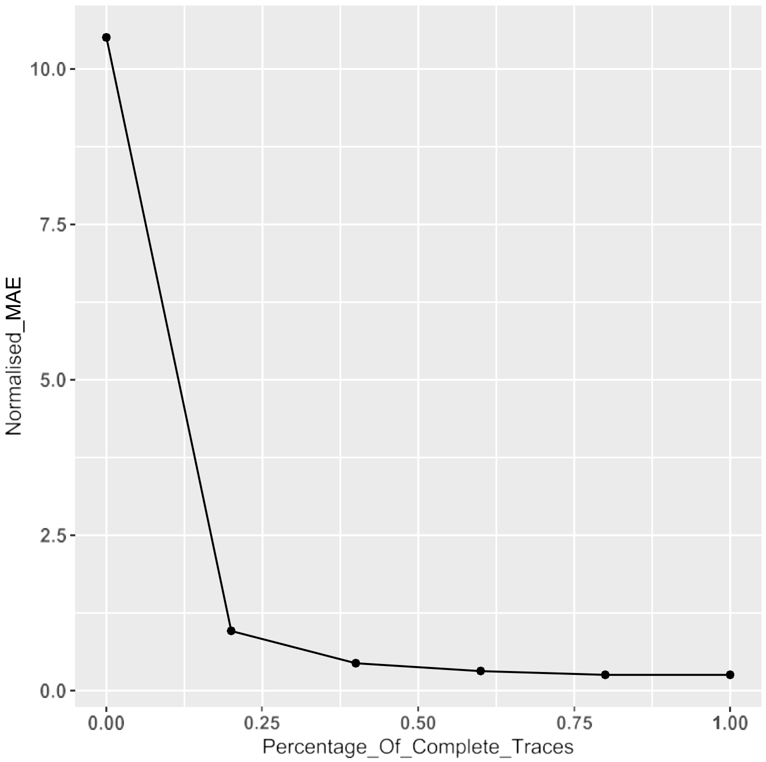


Figure 3.9 - BPIC 18 - Normalised MAE by proportion of incomplete traces

To test this effect, we utilise the non-parametric Kruskal Wallis to determine whether there is a significant difference in the MAE for each log. As expected, there is a significant difference in the MAE for both logs (For BPIC 12, p = 3.282e-09; for BPIC 18, p < 2.2e-16). We subsequently run pairwise comparisons using Wilcoxon rank-sum test to determine which proportions differ significantly from the baseline (i.e. the log with 100% complete traces)

Table 3.9 shows the results of the pairwise comparisons against the baseline

Table 3.9 - Pairwise Comparisons against the baseline using Wilcoxon rank-sum test

	% of Complete Traces in Event Log				
	0%	**20%**	**40%**	**60%**	**80%**
BPIC12	**1.2 x10^{-7}**	**0.0003**	0.1158	0.0737	0.0909
BPIC18	**<2x10^{-16}**	**<2x10^{-16}**	**<2x10^{-16}**	**2.5 x10^{-16}**	**6x10^{-12}**

For BPIC 12, we notice that there is a significant difference until the point at which there is 40% incomplete traces (see results in bold). However, with BPIC 18, we notice that there is a significant difference between the MAE for all logs with incomplete traces against the baseline. To understand the results, we consider the event logs metrics (see Table 3.5). We observe that, despite having roughly the same number of events, BPIC 18 is more complex than BPIC 12 particularly in terms of mean trace length (x4) and the number of distinct activities (x6). We postulate that for complex event logs, our approach delivers a significant difference compared to the baseline, even when there is a high proportion of complete traces. However, for simpler logs, the difference is less pronounced, levelling out when there the proportion of complete cases approaches c.40%.

3.8 Threats to Validity

We encoded the traces using an aggregation encoding technique to enable us to address the research question regarding which social contextual factors were the most important. However, we acknowledge that this encoding technique is quite lossy, which may adversely impact prediction accuracy. As such, we would recommend combining this with other encoding approaches for real-life use.

The final threat to validity is related to the choice of grouped centrality measures selected as social contextual. We selected the most widely used centrality measures in the literature (see van der Aalst, Reijers & Song, 2005; Everett & Borgatti, 1999). We, however, acknowledge that there are additional grouped centrality measures that we could have included (e.g., diffusion and fragmentation centrality) which may have shed further insight. We intend to explore the impact of these in future research studies

3.9 Summary

This chapter has proposed an approach to censor an event log to facilitate its use for building a survival function. We explored the impact of social contextual factors as covariates in the survival function. We found that group betweenness and closeness centrality were generally positively correlated with process cycle time while the group eigenvector centrality was generally negatively correlated. We also found that survival analysis approaches perform comparably with start-of-the-art predictive process monitoring techniques.

We conclude the chapter by briefly reflecting on the research choices made and the potential implications on the outcomes. We posit that different choices may likely have resulted in different outcomes. For example, we chose an aggregation and last state encoding technique to encode the traces. Whilst this approach was chosen as it captured the social relationships between the performers who were involved in executing a trace, we acknowledged that it was quite lossy compared to other encoding techniques such as index encoding. We believe a lossless encoding technique would have greater predictive power, though in this instance, it would not have enabled us to answer our proposed research questions. In real-world settings, a combination of our proposed aggregation techniques and lossless encoding techniques would achieve the twin objectives of incorporating the social context and maintaining predictive power as we as preserving as much information as possible in the encoded trace.

In a similar vein, we could have chosen the semi-parametric Cox model for performing survival analysis. This model differs from the fully parametric Weibull model (which we utilised) in that it is less strict in its assumptions of the time-to-event outcomes. This can be useful in certain scenarios and is a reason why the Cox model remains popular. Whether the less strict assumptions of the Cox model would result in comparable outcomes as our proposed predictive process monitoring workflow remains to be seen.

We conclude the reflection by highlighting several unanswered questions raised by our research findings. Firstly, there has been an increase in the number of automated agents in the workforce executing certain activities in the process and passing the work on to a human agent. Does the presence of an automated agent in the set of performers result in different group centrality measures? If so, what are the implications of this? In addition, non-group performers who interact with the group of performers may possess different sets of feature values (e.g., workload). What effect does this have on the performers? Addressing these questions potentially forms a basis for further research.

In the following chapter, we will examine the impact of spatial contextual factors (a type of external context) on the accuracy of remaining-time predictive process monitoring.

CHAPTER FOUR

4 INCORPORATING SPATIAL CONTEXT INTO REMAINING-TIME PREDICTIVE PROCESS MONITORING

4.1 Synopsis

Though the location of events is a crucial explanatory variable in many business processes, as yet there have been no studies which have incorporated spatial context into the predictive process monitoring framework. This chapter seeks to address this problem by introducing the concept of a spatial event log which records location details at a trace or event level.

The predictive utility of spatial contextual features is evaluated vis-à-vis other contextual features. An approach is proposed to predict the remaining time of an in-flight process instance by calculating the buffer distances between the location of events in a spatial event log to capture spatial proximity and connectedness. These distances are subsequently utilised to construct a regression model which is then used to predict the remaining time for events in the test dataset. The proposed approach is benchmarked against existing approaches using five real-life event logs and demonstrates that spatial features improve the predictive power of business process monitoring models.

This study formed the basis of a journal paper accepted for presentation at the Symposium of Applied Computing 2021 (SAC '21).

4.2 Introduction

Earlier in Section 3.2, we discussed the relationship between the four contextual types. In this chapter we focus on spatial context (a subset of external context).

Van der Aalst (2016:320) makes the point that "although ... external context can have a dramatic impact on the process being analysed; it is difficult to select the relevant variables." This chapter aims to address the problem of incorporating spatial context into the process monitoring workflow by introducing the idea of a spatial event log which includes the locations of process traces and events

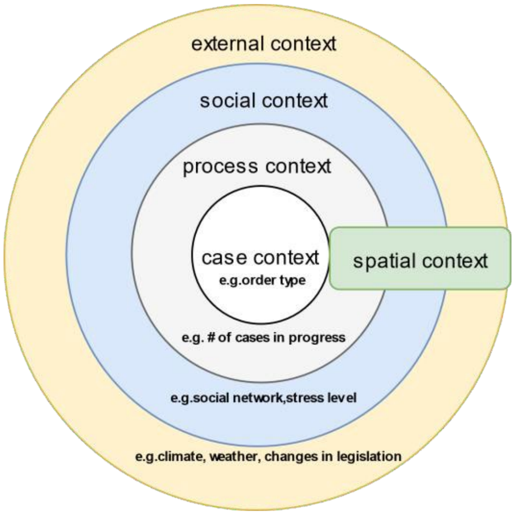

external context
social context
process context
case context
e.g.order type
spatial context
e.g. # of cases in progress
e.g.social network,stress level
e.g.climate, weather, changes in legislation

Figure 4.1 – Spatial Context Relationship with other Contextual Factors (Adapted from van der Aalst, 2016:320)

Even though every event occurs at a location, event logs do not typically capture spatial data. As shown by Figure 4.1, this contextual type overlaps with the other context types. For example, relevant process legislation (external context) and the manner process performers interact (social context) are both a function of location. Incorporating the spatial context enables process analysts to determine whether processes outcomes exhibit spatial patterns. This is a question of interest particularly with distributed processes and one that has increased in salience with the COVID-19 pandemic which has necessitated the distribution of process execution, for example, due to the requirement for process performers to work from home.

If it can be established that process outcomes display spatial pattern(s), location becomes a key explanatory variable. The concept of spatial autocorrelation, which attempts to "measure…simultaneously…the similarities in the location of spatial objects and their attributes", explains this relationship (Longley et al.,2015:34). Besides, incorporating the spatial dimension into event logs facilitates the discovery of the trajectory of process artefacts which could help detect motion waste.

Furthermore, it would be possible to construct a de jure process model for different locations (e.g., because of legislative requirements) and check whether discovered processes (stratified by location) conform. However, for this paper, the focus will be on utilising the spatial context to improve the prediction of the remaining time of process instances. In addition to a contribution to the knowledge base by proposing a novel way to incorporate the spatial context into the predictive process monitoring workflow, we demonstrate by empirical evaluation, the importance of these contextual features. We show that our proposed approach performs comparably with start-of-the-art predictive process monitoring techniques.

The remainder of the paper is structured as follows: Section 4.3 details preceding studies which have provided the motivation and methodological basis for this study. Section 4.4 defines vital terms built on throughout the paper. Section 4.5 describes the proposed approach, while Section 4.6 details the evaluation results. The penultimate section describes the threats to the validity of the study while the final section summarises the findings and proposes further research areas for extending these.

4.3 Related Works

With regards to spatial analysis, Tobler (1970) proposed the first law of geography (a.k.a. Tobler's Law) which states that "all objects are related, but nearer objects are more related than further objects". This law laid the foundation for spatial dependence and autocorrelation. Numerous studies have built on this foundation, and it is commonly accepted as a "reasonable regularity that generally holds true". Miller (2004) argues that rather than merely being a confounding factor, spatial autocorrelation "is information-bearing since it reveals the spatial association among geographic entities".

Hengl et al. (2018) proposes a framework for spatial prediction that utilises buffer distances from observation points as features to build a spatial machine learning model. Their approach offers advantages over traditional geostatistical techniques (e.g., kriging) because it makes "no rigid statistical assumptions about the distribution and stationarity of the target variable, it is more flexible towards incorporating, combining and extending covariates of different types, and it possibly yields more informative maps characterising the prediction error."

In this paper, we utilise the STEP approach combined with the framework proposed by Hengl et al. (2018) to build a spatial predictive process monitoring framework. We adopt this combination as it enables us to encode the event log into spatial features-outcomes pairs to address the relevant research questions. Besides, this avoids issues with generalisation and the "curse of dimensionality" which are associated with some of the other techniques (Senderovich et al., 2017)

4.4 Background

4.4.1 Definitions

Several key terms to be built on throughout this review are formally defined. We build on and extend the definitions from previous chapters.

4.4.1.1 Spatial Objects and Event Logs.

Definition 4.1 (Point) Let R^2 represent a two-dimensional Euclidean space. A point is a zero-dimensional geographical object used to indicate a spatial occurrence in R^2.

A point's coordinates can be specified as longitude, and latitude or Northing N and Easting E offsets relative to a specified origin, depending on the defined Coordinate Reference System (CRS - see Definition 4.4-iii)

Definition 4.2 (Spatial Point Process) Let $X \subseteq R^2$ for some distance d. A spatial point process is a stochastic model for a random scattering of points on X for d which describe the occurrence over time of points {#location$_{(x,y)}$ (e$_1$), #location$_{(x,y)}$(e$_2$)…#location$_{(x,y)}$(e$_n$)} over time {#completion_time (e$_1$), completion_time(e$_2$)…#completion_time(e$_n$)}

Definition 4.3 (Buffer Distances) Let #location$_{(x,y)}$(e$_z$) represent the location attribute for event Z, D_z = (d(#location$_{(x,y)}$(e$_1$), d(#location$_{(x,y)}$(e$_2$)… d(#location$_{(x,y)}$(e$_n$)) represents the buffer distance between #location$_{(x,y)}$(e$_z$) and the other events. It captures the spatial relationship between the location of events in the log.

Definition 4.4 (Spatial event log) A spatial event log is an event log where all events are associated with a location attribute (#location$_{(x,y)}$(e)). For example, we could define a function attr$_{location(x,y)}$$\in \varepsilon \rightarrow P$, to assign a location to each performer (or resource) who execute events. However, it could represent some other location that is meaningful to the process; e.g., for a process to report and track the resolution of a defect, the location could describe the location of the reported defect. We recommend providing the following attributes at the event log metadata level:

i. Location scope attribute (#location_scope(L)) to indicate whether the scope of the location attribute is trace- or event-wide.

ii. Location function (#location_function(L)) to describe the nature of the location attribute in the log.
iii. Coordinate Reference System (#CRS(L)) to indicate the Coordinate Reference System for the event location attribute

To illustrate the terms above, we extend the exemplar process for reporting and remediating defects to public goods, e.g., potholes, streetlight outages. As earlier stated, 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'}. This event log could be transformed into a spatial event log by, for example, associating the location of the appropriate performer with each event (see Table 4.1).

4.5 Approach

4.5.1 Overview

Figure 4.2 provides an overview of the proposed approach. The initial step is the creation of a spatial event log which associates the events in the log with the spatial context. Subsequently, we create measures of spatial proximity by calculating buffer distances for each point in the training data set to all the other points. These distances are used to build a spatial regression model. We improve runtime performance by performing these steps offline.

Subsequently, in the online phase, the remaining time for test data are predicted using the regression models based on the location of the event/trace

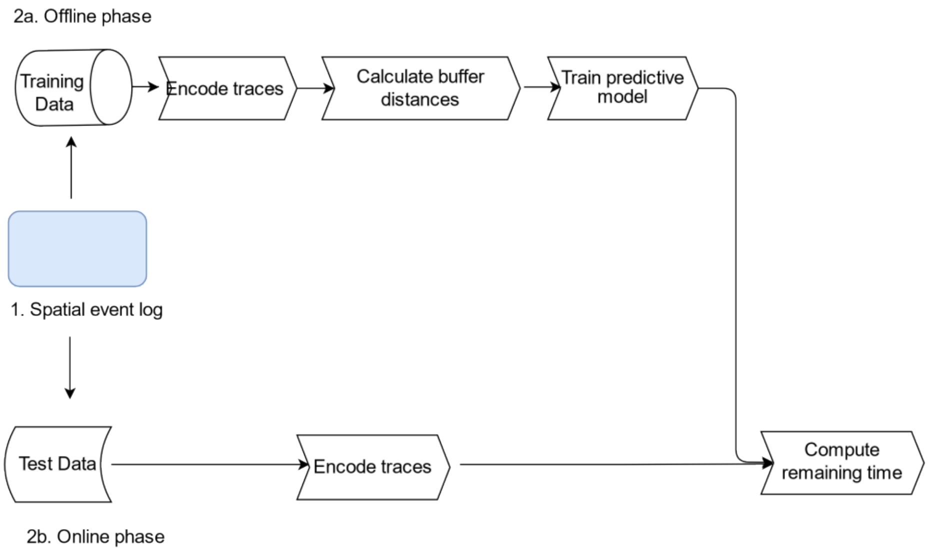


Figure 4.2 - Overview of the proposed approach

4.5.2 Spatial Event Log

As earlier mentioned, Table 4.1 shows a spatial event log. Each row in the log denotes an event with *Service Request ID* representing #case_identifier(e), *Activity* representing #activity(e), and *Completion Time* representing #completion_time(e), these being the triplet that minimally defines an event (see Definition 2.1). Although event logs do not typically contain spatial information, an event log can be transformed into a spatial event log by associating the coordinates of a meaningful location to each event in the log. A location is considered meaningful if it facilitates the discovery of spatial patterns in the event log. A typical choice is the location for the performer associated with each event.

Another example is the location of reported defects (coded as longitude and latitude) in a service management application. This approach is considered meaningful for these processes as the location of defects is expected to demonstrate evidence of a spatial point process. For example, for road defects, the spatial process will likely depend on the weather, maintenance schedule, organisational process, regulation, among others.

Table 4.1 - Spatial Event Log

Service Request ID	Service Category	Longitude	Latitude	Activity	Start Time	End Time
XY4567	Roads	51.3161	0.06047	Create Service Request	22/10/2017 18:34	22/10/2017 18:38
XY4567	Roads	51.2425	0.06132	Accept Ownership	25/10/2017 10:16	25/10/2017 10:17
XY4567	Roads	51.2557	0.06156	Assign Crew	25/10/2017 16:01	25/10/2017 16:22
XY4567	Roads	51.2557	0.06132	Contact Citizen	27/10/2017 11:04	27/10/2017 11:09
XY4567	Roads	51.2557	0.06114	Close Service Request	27/10/2017 11:45	27/10/2017 11:55

4.5.3 Predictive Modelling

The approach consists of two phases: offline (training) and online (testing). In the offline phase, the traces in the event log are encoded. In order to achieve this chapter’s research objectives (see section 4.6), we require an encoding method which enables us to capture the spatial relationships present in the spatial event log. As current predictive process monitoring encoding techniques are unable to facilitate this, we adopt a modification of the technique proposed by Hengl et al. (2018) as described below. Formally, for the training dataset, given a trace $\sigma_i = \{e_1 \ldots e_n, \#location_{(x,y)}(e_1) \ldots \#location_{(x,y)}(e_n)\}$ $\sigma \in \varepsilon*$, σ_i is transformed into a feature set $\{d(\#location_{(x,y)}(e_n), (\#location_{(x,y)}(e_i)) \ldots d(\#location_{(x,y)}(e_n), (\#location_{(x,y)}(e_j)) \} \rightarrow \tau_{rem} (\sigma_i)$ where $d(\#location_{(x,y)}(e_n), (\#location_{(x,y)}(e_i)) = \sqrt{}(\#location_{(x)}(e_n) - \#location_{(x)}(e_i))^2 + (\#location_{(x)}(e_n) - \#location_{(x)}(e_i))^2$

We subsequently utilised the framework proposed by Hengl et al. (2018) to build a Random Forest spatial predictive monitoring model as detailed below.

We start by converting the event log into the spatial data frame to efficiently handle the spatial data. In the offline phase, we calculated the Euclidean buffer distances to all locations in the training dataset as geographical covariates by generating multiple gridded maps. Figure 4.3 illustrates the calculation of the buffer distances for four events in the training dataset.

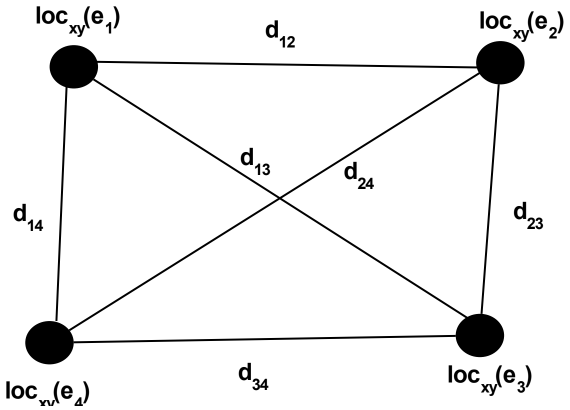


Figure 4.3 – Calculation of buffer distances for training dataset

The value of the target variable(s). i.e., remaining or cycle time, is then modelled as a function of the buffer distances. Table 4.2 illustrates the encoding of the events displayed in Figure 4.3

Table 4.2 – Encoding for training dataset

	e1	**e2**	**e3**	**e4**	**Remaining Time (Hours)**
e1	0	304105.2	760768.06	1070880	839
e2	304105.25	0	1070880	1296950.25	741
e3	760768.06	1030456.2	0	387091.72	322
e4	1070880	1296950.2	387091.72	0	0

In the online phase, the in-flight traces are encoded utilising the location of the final event in the trace, #location$_{(x,y)}$(e_n). The spatial model built in the training phase was used to estimate the remaining time directly for the event-level logs. However, for the trace-level log, the total cycle time for the trace was estimated and the remaining time for the trace is computed by subtracting the elapsed time from the estimated cycle time. Figure 4.4 illustrates how the gridded map enables prediction of the remaining time using the location of the last event in the trace

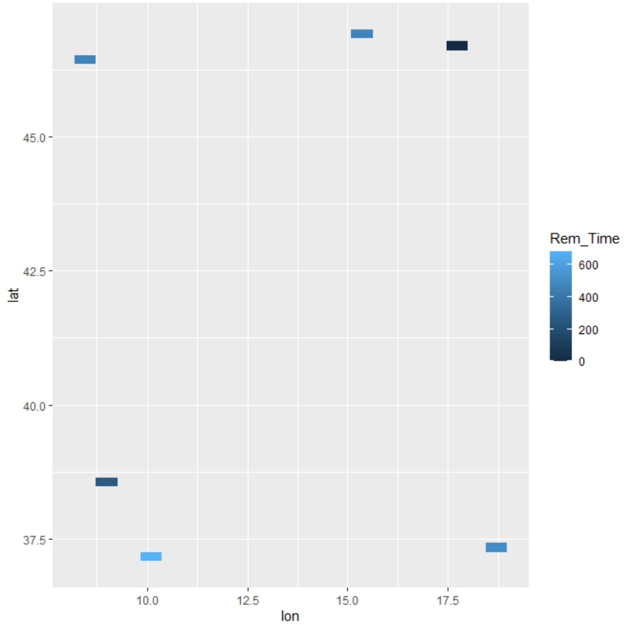


Figure 4.4 – Gridded Map Remaining Time Prediction

Algorithm 4.1 details the spatial predictive modelling algorithm.

Algorithm 4.1 - *S-PM* algorithm

Input: An event log *L* over some trace universe σ with a location scope attribute #location_scope(L), an associated target measure remaining time τ_{rem}, time τ_{ela}, cycle time τ_{cyc}, a spatial window B,a spatial overlay method O and a spatial regression method (REGR) method

Output: A spatial predictive model (*S-PM*) model for *L*

Method: Perform the following steps:

```
i.    Associate a point spatial object #location$_{(x,y)}$(e) with each trace $\sigma \in L$ (see definition 3.7)
ii.   Encode each trace using a suitable encoding function
iii.  For each #location$_{(x,y)}$(e$_i$), calculate D$_i$ =(d(#location$_{(x,y)}$(e$_1$), d(#location$_{(x,y)}$(e$_2$).... d(#location$_{(x,y)}$(e$_n$))
If attribute #location_scope(L) = 'event'
iv.   Overlay $\tau_{rem}$ over $B$ using method $O$ to return $b$
v.    Induce a regression model s-pm out of L using method REGR using {#location$_{(x,y)}$(e$_i$),{ D$_{i ...}$ D$_n$}, $b$} as input value and $\tau_{rem}(\sigma)$ as target value
vi.   Estimate the remaining time for each trace $\tau_{i.rem\_pred}$ : s-pm($\sigma_i$)
vii.  End
If attribute #location_scope(L) = 'trace',
iv.   Overlay $\tau_{cyc}$ over $B$ using method $O$ to return $b$
v.    Induce a regression model pst-pm out of L using method using {#location$_{(x,y)}$(e$_i$),{ D$_{i ...}$ D$_n$}, $b$} as input value and $\tau_{cyc}$ ($\sigma$) as target value
vi.   Estimate the cycle time for each trace $\tau_{i.cyc\_pred}$ : s-pm($\sigma_i$)
vii.  For each $\sigma_i$ do
          a. Estimate the remaining time for each trace $\tau_{i.rem\_pred}$ : $\tau_{i.cyc\_pred} - \tau_{ela}$
viii. End
ix.   Return c{$\tau_{1.rem\_pred}$....... $\tau_{n.rem\_pred}$ }
```

4.6 Evaluation

In this section, we detail our approach to evaluate the importance of spatial features in the predictive process monitoring workflow. We evaluated the proposed spatial predictive monitoring techniques against similar non-spatial predictive monitoring techniques. Specifically, we sought to address the following research questions:

RQ6. Do spatial features contribute to the predictive power of remaining-time predictive approaches vis-à-vis other features?

RQ7. How does spatial-based remaining-time predictive process monitoring approaches compare with existing approaches?

In the following section, we provide further details about the experimental setup and how we answer the research questions.

4.6.1 Datasets

We used five real-life events for our experiments (see Table 4.3). For four logs we enriched the event log with synthetic spatial data as follows: Traffic Fines (de Leoni & Mannhardt, 2015), BPI Challenge 2017 (van Dongen, 2017), BPI Challenge 2019 (van Dongen, 2019), BPI Challenge 2020 (van Dongen, 2020). We simulated the synthetic data to reflect as faithfully as possible the spatial patterns we expect to be present in the process. For example, all the event locations were simulated within the territory of the country where the event log was generated. Besides for each event, we approximated the expected distribution. To illustrate, the expectation for the traffic fine event log is that such fines are predominantly issued in urban areas; hence we simulated a spatially clustered pattern for these events modelled with the Thomas spatial process. For these logs, the location for each event is the simulated location of the performer executing each event. We subsequently refer to these logs as the event-level logs.

The fifth event log included real-life spatial data. This log is from a cloud-based request management platform currently used by public service providers (i.e., municipalities and regions) in Canada and the US. Citizens or service provider staff can raise service requests (i.e., requests for information or work to be carried out, application for permits, etc.) via an app on hand-held devices or through a web interface. Functionality exists for the public service provider (typically a municipal agency) to manage these requests through to completion as well as a suite of supporting functionality, e.g., analytics, work management, etc. The scope of the locations in this log is at a trace level. i.e., every event has the same location

Table 4.3 - Event Log Overview

	Traffic Fines	**BPIC 17**	**BPIC 19**	**BPIC 20**	**Road Defects**
# of events	149354	55358	140056	56437	9392
# of cases	26633	3084	306	10500	1324
# of traces	215	1126	305	99	413
# of distinct activities	11	25	34	17	29
Mean trace length	5.61	17.95	457.7	5.37	7.09
Mean throughput time (days)	528.96	21.87	156.78	11.53	82.3
Throughput time SD (days)	346.62	12.94	529.98	17.02	244.78
Domain	Public Admin	Financial Services	Manufact uring	Education	Public Admin
Location Scope	Event	Event	Event	Event	Trace

and the coordinates indicate the location of the reported defects; hence we hereafter refer to this as the trace-level log. We filter the log to extract defects related to road-related defects. However, we are unable to make the data available as doing so will create privacy concerns due to the location coordinates representing observed locations of real people. We considered robust anonymisation of the data; however, we concluded that doing so without loss of accuracy was not achievable

We added additional features such as elapsed time, remaining time, the number of requests raised on the same day as the service request (a measure of workload) and a couple of temporal features to each log.

4.6.2 Experimental Setup

For the evaluation, we implemented a function named *spatial* in R for the spatial algorithm described in section 4.5.3, respectively. This implementation enables assessment of the importance of the spatial features by building a predictive model from these features and evaluating them vis-à-vis predictive models based on non-spatial features. With regards to selecting the approaches to evaluate against, we were conscious that, unlike we had done with previous approaches, we would be evaluating spatial encoding and features against non-spatial ones. As such we decided to evaluate spatial approaches against a couple of non-spatial approaches which used a zero prefix-bucketing combined with a gradient boosting machine (*gbm*) and multilayer perceptron (*mlp*) neural network regressors respectively (as done in chapter 3) to predict the remaining time for each trace (see Verenich et al., 2019) to ensure consistency and parsimony. We blended each of these approaches with the spatial model using the arithmetic mean of the predictions to create a couple of ensemble models for evaluation purposes. To ensure completeness, we also create a blended ensemble of the non-spatial models. This combination of approaches would enable us to determine the contribution of spatial context to the predictive power of the model. The code and data for the experiments are located here: https://github.com/etioro/SpatialProcessMonitoring.

For the event-level logs, we encode the traces as described in Section 4.5.3. However, for the trace-level log, as there is a single location for each case, we utilise the location and cycle time for completed traces to build the gridded map from the training data set. Thereafter, for each inflight trace from the test dataset, the cycle time for the trace was estimated and the remaining time for the trace is computed by subtracting the elapsed time from the estimated cycle time.

We split each event log into test and training sets. We further subdivided the training set, using only the spatial features for 200 data points to build the spatial model and the non-spatial features for the remaining data points to construct the non-spatial models. We subsequently used the test set for making remaining-time predictions which are then evaluated.

As with the methodology used in (see Verenich et al., 2019), the training & test set were not temporally disjoint.

As earlier indicated in Section 2.7.3, we chose to utilise the Mean Absolute Error (MAE) to evaluate the accuracy as other measures such as the Root Mean Square Error (RMSE) are susceptible to outliers and Mean Absolute Percentage Error (MAPE) would be skewed towards the end of a case where remaining time tends towards zero.

To achieve the best performance from both the spatial and non-spatial models, we tuned the relevant model hyperparameters. For the spatial-based model, we utilise the techniques proposed in Hengl et al. (2018), while for the non-spatial methods, we use the tuning capabilities inbuilt into the caret package.

4.7 Results

Table 4.4 details the global MAE and Standard Deviation (SD) for each dataset/algorithm pair. The performance of the algorithms is visualised in Figure 4.5, which displays the average ranking of each algorithm over the datasets with associated error bars, calculated as the standard deviation of the rankings. Over the five datasets, the ensemble model *gbm+spat* performed best, though it had the third lowest error. In general, blending the spatial model with a non-spatial model improved the performance of the non-spatial model. This is explained by the fact that the spatial features explained as much as 30% of the dependent variable (i.e., remaining time) in the spatial models. It is also worth mentioning that the *spatial* model outperformed the ensemble non-spatial models (i.e., *gbm+mlp*). This confirms the valuable contributions of the spatial features. However, it is also worth noting that whilst *gbm+spat* performed best overall, its performance varied across the datasets. This appears to be related to the spatial variation in the event logs, with the algorithm performing better when there is more variation and worse when there is less.

Table 4.4 - Global MAE ± SD

	spatial	**mlp**	**Gbm**	**gbm+mlp**	**gbm+spat**	**mlp+spat**
Traffic Fines	183.09 ± 180.22	276.86 ± 173.20	255.96 ± 206.70	259.97 ± 115.24	216.52 ± 156.45	224.38 ± 151.92
BPIC 17	11.68 ± 10.71	14.62 ± 8.93	8.79 ± 9.51	11.44 ± 8.05	9.86 ± 9.72	12.62 ± 9.07

BPIC 19	81.47 ± 62.45	156.13 ± 86.91	69.29 ± 57.87	100.39 ± 62.67	60.92 ± 40.08	98.64 ± 69.37
BPIC 20	6.12 ± 22.95	6.55 ± 22.06	4.61 ± 21.94	5.38 ± 21.94	4.98 ± 21.94	6.07 ± 22.04
Road Defects	114.64 ± 214.17	109.06 ± 224.99	126.25 ± 208.65	113.36 ± 208.77	115.81 ± 203.14	111.21 ± 219.17

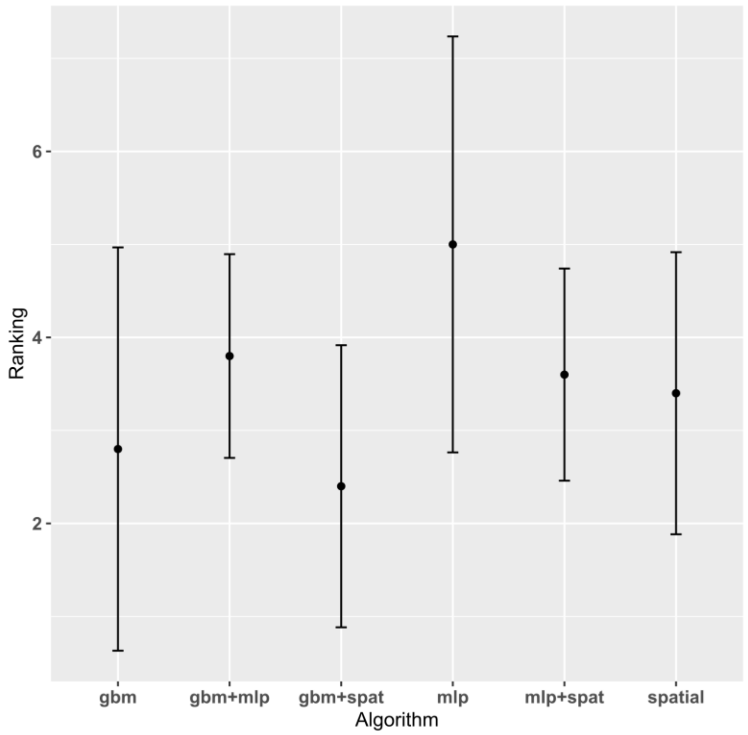


Figure 4.5 - Average Algorithm Ranking with associated error bars

Figures 4.6 show 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 Verenich et al., 2019). *gbm+spat* has the lowest normalised median and mean MAE (0.43 and 0.62 respectively)

To determine which algorithms, differ from the others, we utilise the Quade post-hoc test to perform a pair-wise comparison between the various algorithms. Table 4.5 shows the results of the pair-wise comparisons (with the value(s) statistically significant at the 95% confidence level in bold font). For most of the pairs, there is insufficient evidence to reject the null hypothesis that they are significantly different. However, the results indicate that the *gbm+spat* method significantly outperforming the existing method(s) (see results in bold).

Table 4.5 - Quade post-hoc test of approach rankings

	spatial	**mlp**	**mlp+spat**	**gbm**	**gbm+spat**
mlp	0.183				
mlp+spat	0.865	0.241			
gbm	0.61	0.072	0.498		
gbm+spat	0.399	**0.036**	0.313	0.734	
gbm+mlp	0.734	0.313	0.865	0.399	0.241

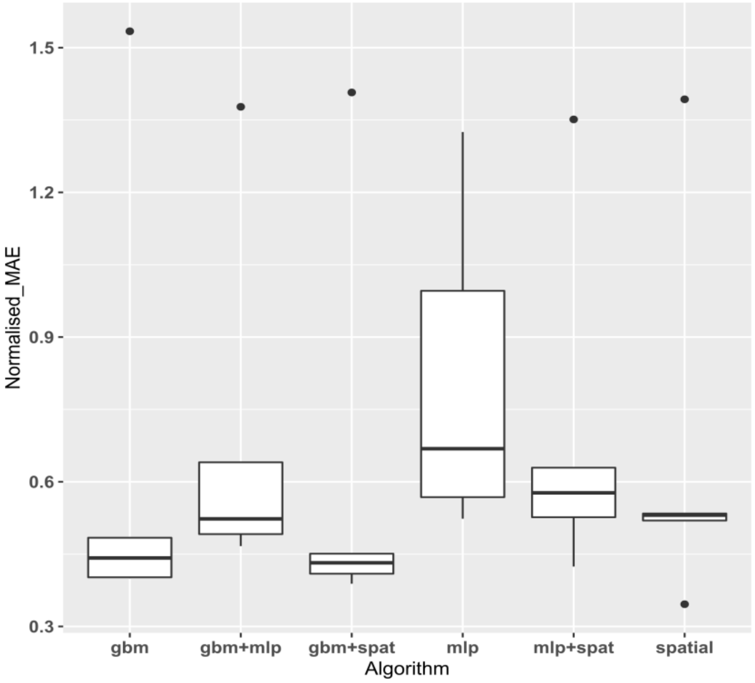


Figure 4.6(a) - Average Normalised MAE

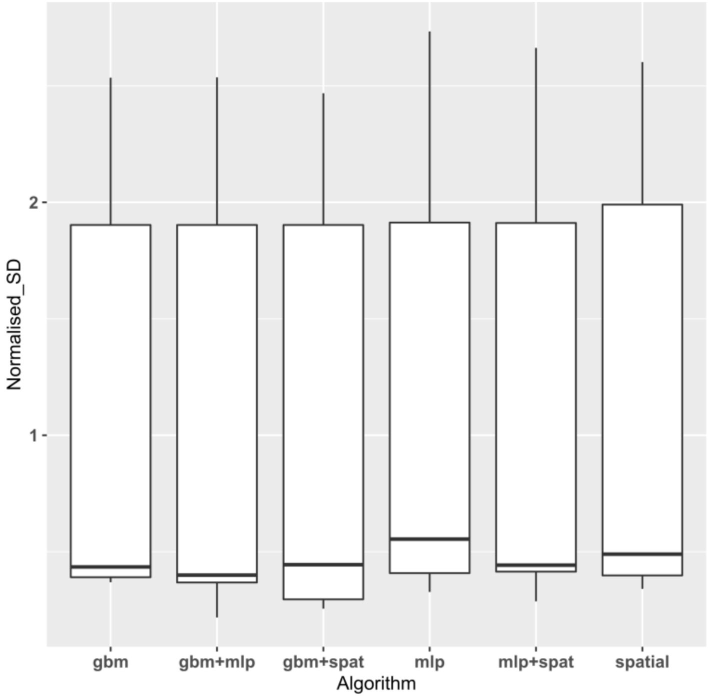


Figure 4.6(b) - Average Normalised Standard Deviation

4.8 Threats to Validity

The main threat to validity was the absence of real-life spatial data at the desired level of granularity. For the four event-level logs for which spatial data was simulated, even though care was taken to reflect the spatial distribution of the process in the simulated data, the spatial effect is likely under-estimated vis-à-vis real-life spatial data.

For the real-life spatial data, the available spatial data was at trace level. In other words, a single location (i.e., service request location) was associated with each completed trace. However, in reality, the location for events is typically dispersed, i.e., e_1 may occur at location A, e_2 at location B, etc. For example, a citizen may raise the service request at location A, reviewed by supervisor based in the field location (at location B) and assigned to a workcrew based at location C. Lower granularity of locations at event level is expected to produce better results as these captures more of the spatial variation present in the data

Another threat to validity is related to the real-life spatial data is geo-referencing uncertainty (Longley et al.,2015:81). For that dataset, the request creator may introduce uncertainty by specifying the incorrect location for the service request or by the service request submission platform. Hence a point may be incorrectly positioned. We assume that this uncertainty is minimal as the relevant public service provider was able to locate and complete all the service requests we selected for our experiment.

Finally, we recognise that not all processes will possess a significant amount of spatial variation. For example, for centralised processes, the process performers may all be co-located. For these processes, spatial features are not likely to significantly contribute to the accurate prediction of the remaining time

4.9 Summary

This study has proposed an approach to incorporate spatial context into event logs and performed a comparative analysis of spatial features against other contextual features. It found that spatial features improve the predictive power of the model and that spatial ensemble approaches yielded the best result for processes that are likely to exhibit spatial point processes.

In conclusion, we reflect on the potential impact on our choice of research methods. Firstly, the encoding technique for the test trace utilises the location of the last event in the trace for predicting the remaining time. However, as a trace is a sequence of events, we acknowledge that, a technique that incorporates the location of all the events in the trace, is likely to have a higher predictive power. In other words, we believe that utilising the spatial 'path' or 'trajectory' the case takes through to completion as the basis of prediction would result in more accurate predictions. This is especially true where a trace has rework loops, where the rework is executed at the same location as the earlier event. However, we were unable to locate a suitable spatial algorithm that utilised the trajectory hence our choice for using the last event.

Secondly, we separated the spatial from the non-spatial context in order to answer the research questions, i.e., determine the predictive power of spatial features. However, we recognise that, even when significant spatial variation exists in the event log, utilising only the spatial features is not likely to produce models with high predictive power. As such, in a real-world setting, we would expect both sets of features to be combined to increase the predictive power of the model.

Finally, it is worth pointing out that we utilise different techniques to encode the training and test data sets. As described in section 4.5.3, we utilised the buffer distance to every event in the training set to encode those traces and used the location of the last event in the trace for the test set. This difference is in our opinion necessary, and the right choice given the unique characteristic of spatial features. However, we call out this difference as other predictive process monitoring approaches typically utilise the same technique for encoding both the training and test datasets.

In the next chapter, rather than focus on the impact of contextual factors on remaining time prediction, we focus on the interplay between contextual factors. Specifically, we examine the relationship between workload (a typeof process context) and stress (a type of social context).

CHAPTER FIVE

5 INVESTIGATING THE DIFFUSION OF WORKLOAD-INDUCED STRESS - A SIMULATION APPROACH

5.1 Synopsis

As mentioned earlier, the target for a predictive process monitoring workflow is typically one of remaining time, outcome, or next step. Less frequently, cost, or other numeric process outcomes are predicted. However, we argue that the set of targets should be expanded to increase the scope and usefulness of predictive process monitoring workflows. In this chapter we attempt to predict a new target (stress - a social contextual factor) utilising workload (a process contextual factor). We thus add to the knowledge base by discovering the interaction between these process contextual factors.

Work-induced stress is widely acknowledged as harming physical and psychosocial health and has been linked with adverse outcomes such as a decrease in productivity. Recently, workplace stressors have increased due to the Covid-19 pandemic. This chapter aims to contribute to the literature base in a couple of areas.

First, it extends the current knowledge base by utilising Generative Additive Modelling (GAMs) to uncover the nature of the relationship between workload (a key workplace stressor) and productivity based on real-world event logs. Besides, it uses recursive partitioning modelling to shed light on the factors that drive the relationship between these variables.

Secondly, it utilises a simulation-based approach to investigate the diffusion of workload-induced stress in the workplace. Simulation is a valuable tool for exploring the effect of changes in a risk-free manner as it provides the ability to run multiple scenarios in a safe and virtual environment with a view to making recommendations to stakeholders.

However, there are several recognised issues with traditional simulation approaches, such as inadequate resource modelling and the limited use of simulations for operational decision making.

In this chapter, we propose an approach which extracts the required parameters from an event log and subsequently utilises them to initialise a workload-induced stress diffusion simulation model accurately. We also explore the effects of varying the parameters to control the spread of workload-induced stress within the network

With suitable amendments, this approach can be extended to model the spread of disease (e.g., Covid-19), diffusion of ideas, among others, in the workplace.

This chapter formed the basis of a journal paper published in *Information*

5.2 Introduction

Work-induced stress is defined as "the change in one's physical or mental state in response to workplaces that pose an appraised challenge or threat to that employee" (see Colligan & Higgins, 2006). The impact of workplace stress includes "increased absenteeism, organizational dysfunction, and decreased work productivity" (Colligan & Higgins, 2006). Workplace stress has also been linked to higher levels of alcohol consumption during retirement (see Richman et al., 2006). A key stressor in the workplace is the workload and pace of work (Bickford, 2005). Numerous studies have explored the relationship between workload and productivity. For example, Hebb (1955) proposed a quadratic relationship between arousal (a proxy for workload-induced stress) and performance (see Figure 5.1).

Other studies have built on and extended this relationship, referred to as the Yerkes-Dodson law. For example, Bertrand & Van Ooijen (2002) describe the widely accepted explanation of the relationship. The authors posit that when the workload is below the optimal level of arousal and performance, performers are not as alert and hence do not perform at the optimal level. However, as the workload increases, so does alertness until the optimum level of performance is reached. Any increase in workload past this point results in decreased performance as performers "need more time to process information, to take decisions and, due to the high level of arousal, might make more mistakes". They argue for a "load-based work order release" system which feeds work into the system based on the existing workload on the shop floor and posit that this has a positive impact on increasing work order throughput times. Nakatumba & van der Aalst (2009) also explores this relationship utilising a Process Mining approach. Whilst that paper concluded that "the relationship described by the Yerkes Dodson law of arousal really exists", the study stopped short of demonstrating the existence of the inverse U-shape relationship arguing that "more sophisticated…techniques" were required to confirm this.

That gap is what the first half of this paper attempts to address. Utilising a couple of real-world event logs, we build Generative Additive Models (GAMs) to uncover the nature of the relationship between workload and productivity. GAMs enable us to fit non-linear relationships to the data of interest and are relatively interpretable. In addition, we build a couple of recursive partitioning models to shed light on the factors that drive the relationship between these variables. Thus, this study contributes to the literature by uncovering the nature of the relationship between these variables and the factors that drive them.

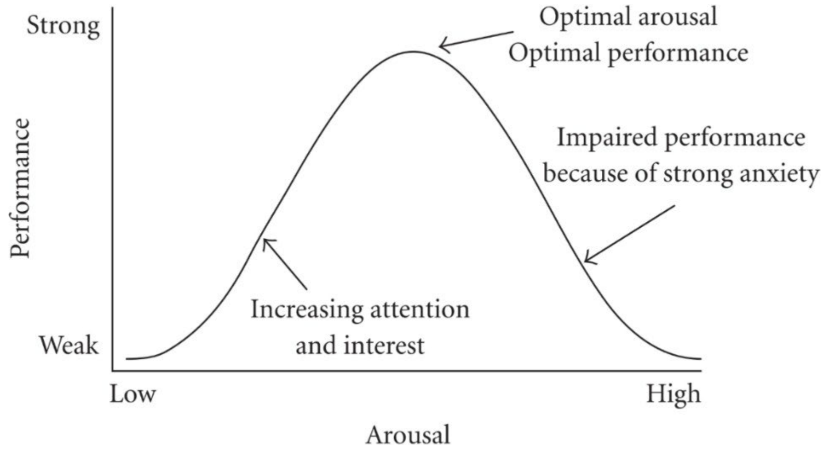


Figure 5.1 - Hebbian version of the Yerkes Dodson law (Source: Diamond et al., 2007)

The value of simulation to rapidly explore the effect of changes in a risk-free manner has long been understood. However, van der Aalst et al. (2008) highlights several issues with traditional simulation approaches. Apart from the limited use of "existing artifacts such as historical data and workflow schemas", the modelling of process performers is inadequate (e.g., the incorrect assumption that performers work at a constant speed or the assumption that performers immediately work on incoming tasks when they are available). To address these issues, van der Aalst (2010) argues that "to adequately set these parameters and make sure that processes are modeled accurately…the information available in event logs" needs to be exploited utilising process mining techniques. van der Aalst et al. (2008) highlights four pertinent types of data, namely: *event log* which describes historical information about recorded events, *process state* which represents information attached to cases, *process model* which describes the sequencing and routing of activities and the *resource model* which conveys information about performers, roles, departments, etc.

Extending the link earlier established between workload and stress, the principle of emotional contagion – the phenomena of having one person's emotions trigger emotions and related behaviours in others – has long been accepted (see Reik,1948; Jung, 1968). More recent studies in the field of neuroscience have established the neurological basis of these phenomena (Iacoboni et al., 2005; Rizzolatti, 2005). We posit that as co-workers interact as they execute common activities simultaneously, stressed workers “infect” non-stressed workers and thus diffuse stress across the workplace.

To identify a suitable simulation model for the spread of workload-induced stress, we delve into the field of epidemiological research to examine models for exploring the transmission of infectious diseases. Jenness, Goodreau & Morris (2018) proposes a “general stochastic framework for modelling the spread of epidemics on networks”. This approach is an ideal choice for combining simulation with processing mining as there exist several studies which have successfully discovered social networks from event logs. For example, van der Aalst, Reijers & Song (2005) proposes an approach for discovering social networks from an event log and several metrics based on potential causality, joint cases/activities and special event types. They also apply these concepts to a real-life event log. In Song & van der Aalst (2008), the authors build on these and extend the approach to discover organisational models from event logs.

In the second half of this study, we discover a social (co-worker) network from an event log and utilise the network properties to initiate a simulation model which explores the spread of workload-induced stress. We further contribute to the literature base by proposing a novel approach which investigates the diffusion of workload-induced stress utilising an epidemiological simulation model initialised with parameters extracted from an event log. Whilst the focus in this study is the diffusion of workload-induced stress, with suitable amendments, the model can also be used to explore the spread of disease (e.g., COVID-19) in the workplace or the diffusion of ideas, amongst others.

The remainder of the chapter is structured as follows. Section 5.3 defines vital terms built on throughout the paper and describes the proposed approach, while Section 5. 4 details the evaluation results of the proposed approach. The penultimate section describes the threats to the validity of the study, while the final section summarises the chapter.

5.3 Background

5.3.1 Definitions

5.3.1.1 Processing Time, Speed and Workload

Definition 5.1 *Processing time.* Let *e* represent an event, #start_time(*e*), the start time associated with the event and #completion_time(*e*) the completion time associated with the event. The processing time for *e,*

τ_{proc} = #completion_time(*e*) - #start_time(*e*). It indicates the time taken to complete processing the event

Definition 5.2 *Workload.* Let *A* represent the set of valid activity labels, *W* represent a time window with start, W_{start}, end, W_{end}, and event log *L*. The workload function is defined as:

$$workload(\#\text{activity}(e), W_{start}, W_{end}) \to \mathbb{N}$$

where ℕ is the set of natural numbers {0,1,2,3…}. This denotes the number of instances of a specific activity present in time window W

We further define $\text{attr}_{\text{completed}}$ $(p, a) \to \mathbb{N} \leq workload$ for each performer/activity pair to indicate the number of events (with activity label *a*) the performer *p* completed in the given time window

Definition 5.3 *Average Processing Speed.* Given a set of valid activity labels *A,* a time window *W* and a performer *p,* the processing speed is defined by:

$$\text{s} = \frac{\Sigma\ \tau_{proc}\text{: \#start\_time}(e_i) \wedge \text{ \#completion\_time}(e_i) \in W}{\text{completed (p,a)}} \tag{5.1}$$

This indicates the average processing speed for the performer/activity pair in the given time window

5.3.1.2 Social Networks and Network Models

Definition 5.4 *Co-worker network.* Let P represent the set of performers, E represents a set of undirected edges and ϕ: $E \rightarrow \{x, y$: $x, y \in P\}$ represent an incidence function mapping edges to vertices defined as follows:

$$\text{coworker}\ (x, y) = \begin{cases} 1, & \text{if } \#\text{activity\_label}(\mathrm{e}_i) = \#\text{activity\_label}(\mathrm{e}_j) \wedge \\ & [\#\text{start\_time}(\mathrm{e}_i), \#\text{completion\_time}(\mathrm{e}_i)] \cap [\#\text{start\_time}(\mathrm{e}_j), \#\text{completion\_time}(\mathrm{e}_j)] > 0 \\ 0, & \text{Otherwise} \end{cases} \tag{5.2}$$

A co-worker graph is an undirected multigraph G = (P, E, ϕ). For our study, the incidence function maps an edge when two performers are co-workers as well as the duration of each interaction. Two performer *x and y* are considered co-workers if *x* completes e_i, *y* completes e_j, with both events having identical activity labels and the processing time interval for both events overlap (see Figure 5.2).

Definition 5.5 *Network Model*. Let G denote an undirected graph representing the co-worker network, P represent the set of performers. The partnership formation process for the network simulation model is defined by:

$$\text{logit}\ [\mathrm{P}(\mathrm{G}_{xy,\mathrm{t+1}} = 1 |\ \mathrm{G}_{\mathrm{xy,t}} = 0, \mathrm{G}^{\mathrm{c}})] = \theta_{+}^{\mathrm{T}}\ \delta(\mathrm{g}_{+}(\mathrm{z})) \tag{5.3}$$

where G_{xy} denotes the edge between vertices $x,y \in P$, P(G) denotes the probability distribution of the network, z denotes the observed network, G^{c} denotes the rest of the network, θ denotes the conditional log-odds of G_{xy} as a function of the number of configurations it creates and δ(g) denotes change statistics that indicate how the count of configurations change when G_{xy} is toggled from 0 to 1. Note that G_{xy} is indexed by time and formation at time *t+1* is conditional on G_{xy} existing by time *t* (see Jenness et al., 2018).

The complimentary edge dissolution process is defined as follows:

$$\text{logit}\,[P(G_{xy,t+1} = 1 \mid G_{xy,t} = 0, G^c)] = \theta^{\underline{T}}\,\delta(g_{-}(z)) \qquad (5.4)$$

To illustrate the terms above, we again extend our exemplar process for reporting and remediating defects to public goods. A snippet of the event log is shown in Table 5.1. The set of valid activity labels is as follows: {'Create Service Request', 'Initial Review', 'Assign Service Request', 'Assign Crew', 'Contact Citizen', 'Put Service Request On Hold', 'Close Service Request'}.

Table 5.1 – Illustrative Event Log

Service Request ID	Activity	Start Time	End Time	Performer
XY4567	Create Service Request	22/10/2017 18:34	22/10/2017 18:38	Citizen1
XY4567	Initial Review	25/10/2017 10:12	25/10/2017 10:14	Resource1
XY4567	Accept Ownership	25/10/2017 10:16	25/10/2017 10:17	Resource1
XY4567	Assign Service Request	25/10/2017 11:26	25/10/2017 11:29	Resource1
XY4567	**Assign Crew**	**25/10/2017 16:01**	**25/10/2017 16:22**	**Resource2**
XY4567	Contact Citizen	27/10/2017 11:04	27/10/2017 11:09	Resource2
XY4567	Close Service Request	27/10/2017 11:45	27/10/2017 11:55	Resource2
XY8910	Create Service Request	21/10/2017 15:12	22/10/2017 15:20	Citizen2
XY8910	Accept Ownership	22/10/2017 11:22	25/10/2017 11:25	Resource3
XY8910	**Assign Crew**	**25/10/2017 16:12**	**25/10/2017 16:32**	**Resource4**
XY8910	Close Service Request	26/10/2017 12:23	26/10/2017 12:55	Resource4

Figure 5.2 illustrates the concept of "co-workers" (see Definition 5.4). We observe that the processing time interval for #case_identifier (XY4567), #activity (Assign Crew) executed by Resource2 overlaps by 10 minutes with the processing time interval for #case_identifier (XY8910), #activity (Assign Crew) executed by Resource4 (see bold font). Thus, an edge is formed between Resource2 and Resource4 in the co-worker network, and the

duration of interaction (or exposure) is 10 minutes (see dashed lines).

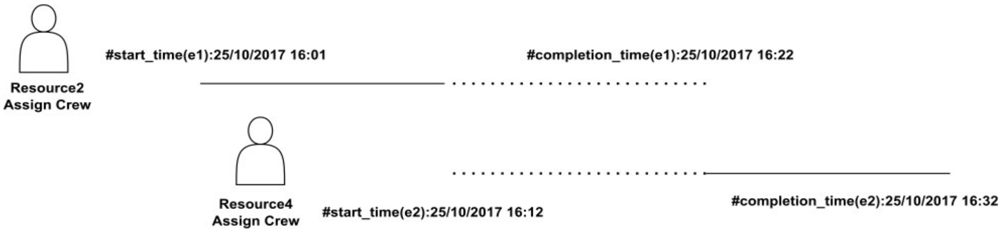

#start_time(e1):25/10/2017 16:01
#completion_time(e1):25/10/2017 16:22
Resource2
Assign Crew
Resource4
Assign Crew
#start_time(e2):25/10/2017 16:12
#completion_time(e2):25/10/2017 16:32

Figure 5.2 - Co-worker network

5.4 Evaluation

In this section, we describe the two sets of analyses performed to address the research questions of interest in this study. In the first set of analysis, we evaluate the relationship between the workload and processing speed to determine whether it displays a quadratic relationship (as predicted by the Yerke-Dodson law) and if so, under which conditions. In the second set of analysis, we simulate a network model to investigate the diffusion of workload-induced stress in a co-worker network. Specifically, we seek to address the following research questions:

RQ8: Does the relationship between workload and processing speed exhibit a quadratic relationship as proposed by the Yerkes-Dodson law?

RQ9: If so, when does this relationship hold and when not?

RQ10: Do network simulation approaches facilitate the discovery of successful interventions to mitigate the diffusion of workload-induced stress?

In the following section, we provide further details about the setup and how we answer the research questions.

5.4.1 Datasets

Two real-life event logs from the Business Process Intelligence Challenge (BPIC) were used as follows: BPIC12(W) (van Dongen, 2012), BPIC17(W) (van Dongen, 2017). BPIC 12 contains event log data for a credit (i.e., personal loan or overdraft) application process at a Dutch financial institution. BPIC 17 contains data from the same process and institution, however from a different supporting system. These logs were selected as they contained a significant proportion of cases with both event start (#start_time(e)) and completion (#completion_time(e)) timestamps. This enabled us to calculate the processing speed for these events

For the simulation exercise, we used a synthetic event log (FutureLearn, n.d.) which contains the details for a repair process. This log was selected as, not only did it contain data which enables calculation of processing speed, but also information about the performer role which was used in the initialisation of the network simulation model.

See Table 5.2 for a summary of the logs used for the experiments.

Table 5.2 - Event Log Overview

	BPIC 17(W)	**BPIC 12(W)**	**Repairs Log**
Number of events	768,823	170,107	15,486
Number of cases	31,509	9,658	1,104
Number of traces	10,701	2,643	80

Number of distinct activities	8	7	8
Mean trace length	24.40	17.61	14.03
Mean throughput time (days)	21.89	11.68	0.05
Throughput time SD (days)	13.17	12.79	0.01
Domain	Financial services	Financial services	IT Support

5.4.2 Experimental Setup

To investigate the first two research questions, we implemented a function in R to calculate the daily workload and average processing speed for each performer (*time window start* =00:00:00; *time window end*=23:59:59). We selected this window for the sake of parsimony and due to the presence of activities in the log which complete late in the day (e.g., after 23:00). Hence, we decided not to filter the log to a typical workday (. i.e., 08:00 – 18:00) as the observed work pattern did not fit this. We considered calculating the daily workload and processing speed per performer better to capture the true nature of the demand on performers. However, we realised that the mean for different activities differed based on activity complexity, as such combining all the activities performers had completed each day was likely to distort the average processing speed. As such, we adopted the methodology used in Nakatumba & van der Aalst (2009) and calculated the total daily workload for each activity, the number of activities each performer completed daily and the average processing speed per performer/activity. We also calculate the cumulative workload for each activity and the number of activities each performer completed over the event log. We subsequently fitted a Generalized Additive Model (GAM) to uncover the relationship between total daily workload and processing speed. We smooth the GAMs with the restricted maximum likelihood method as this is widely acknowledged as most likely to produce stable and reliable results. We create a GAM model for each performer/activity combination with the average processing speed as the dependent variable and the total daily workload as the independent variable. In addition, we extract relevant statistics from each model such as the expected degrees of freedom (edf) which indicates the complexity of the model's smooth, and the p-value, amongst others.

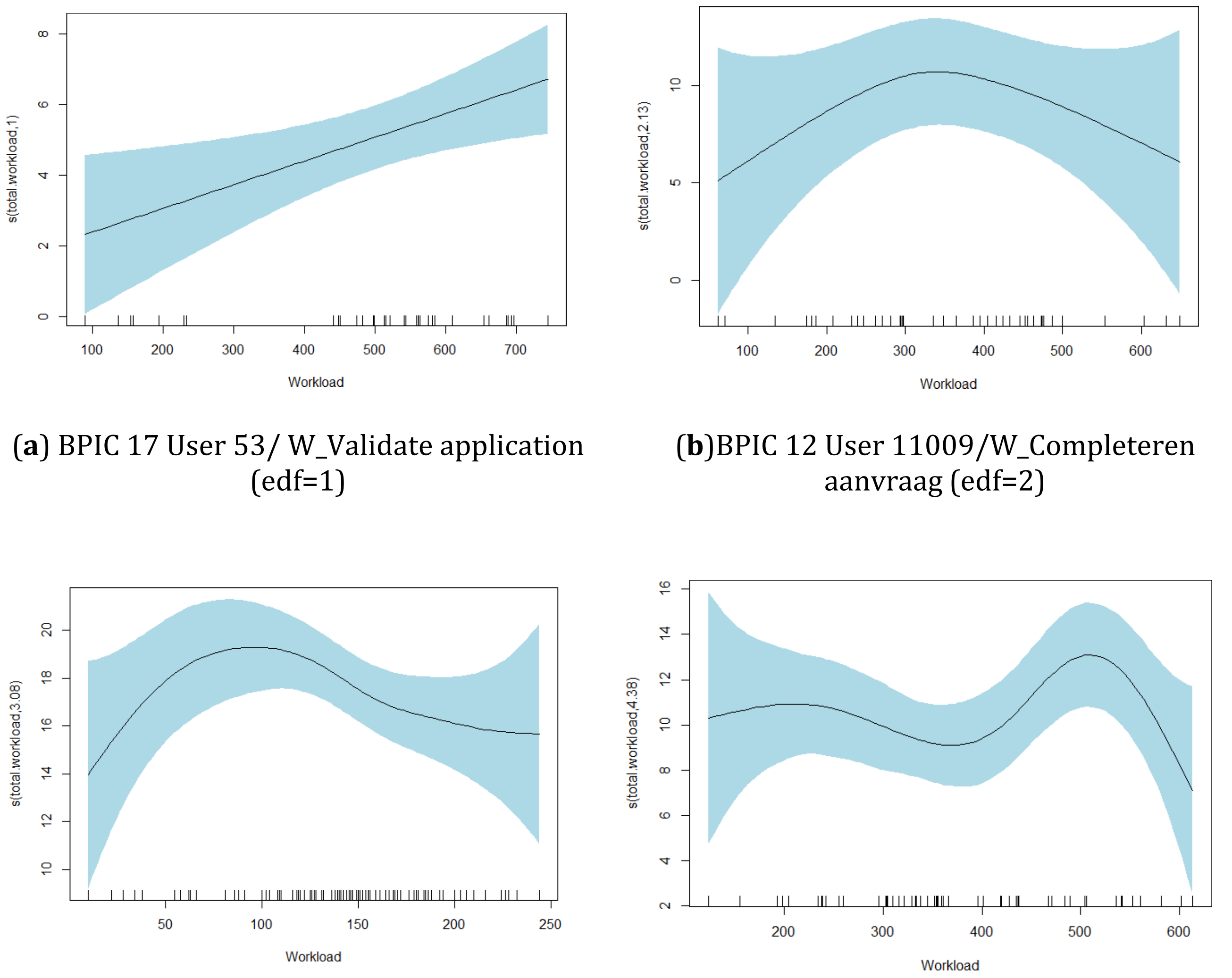
To answer the last research question, we created a social network from the co-worker network of performers in the Repairs event log. We subsequently extracted the following network properties from the co-worker network: the number of edges and vertices in the network, the number of stressed edges, the number of homogenous edges, the number of concurrent interactions and the mean duration of interactions. Each performer is assigned an appropriate state (stressed/not stressed) based on whether their daily workload completed falls within the final quartile. As earlier established, based on findings in the literature, we posit that a stressed performer (. i.e., infected) can spread stress through the co-worker network by "infecting" non-stressed performers via the process of emotional contagion. Borrowing from the field of epidemiology, we create a Susceptible-Infected-Susceptible (SIS) model to simulate the diffusion of stress across the worker network. We chose this model (as opposed to a Susceptible-Infected-Recovered (SIR) model which assumes immunity once recovered, for example) as a performer is again susceptible to workload-induced stress after recovery. The extracted co-worker network properties were used to initiate the SIS model. We concluded by executing multiple runs of the simulation to determine the effect that varying the infection probability and recovery rate had on the number of performers who were stressed at the end of the simulation run.

5.5 Results

For the first set of results, to explore the relationship between workload and performance, we filter for the GAM models which are significant at the 90% confidence level. Table 5.3 shows the distribution of the edf for the models. We observe that across both datasets, 43% of models have an edf of 1 indicating a linear relationship, 18% an edf of 2 indicating a quadratic relationship and 40% an edf greater than or equal to 3, indicating a more complex smooth (see Figures 5.3a-d for example plots). We note that there is partial support for the inverse U shape in the literature as even the more complex smooths (i.e. edf ≥ 3) demonstrate this relationship. Note that the scale of the plots is shifted by the value of the intercept to aid interpretability. Hence, we can predict the output assuming other variables are held at their average value. For example, for plot 5.3b, the predicted productivity for User 11009 performing activity “W_Completeren aanvraag” at the daily workload of 300 cases is 10 activities per day.

Table 5.3 - Distribution of Effective Degrees of Freedom (edf) for GAM Models

	1	2	≥ 3
BPIC 12 (W)	44%	17%	39%
BPIC 17 (W)	41%	18%	41%



(**a**) BPIC 17 User 53/ W_Validate application (edf=1)

(**b**)BPIC 12 User 11009/W_Completeren aanvraag (edf=2)

(**c**) BPIC 12 User 10972/W_Valideren aanvraag (edf=3)

(**d**) BPIC 17 User 24/ W_Complete application (edf=4)

Figure 5.3 - GAM Plots of workload against processing speed

We explore further to uncover the factors which drive the nature of the relationship between the average processing speed and the daily workload. Borrowing from the approach adopted by Hunsicker et al. (2016), we build a couple of recursive partitioning (*rpart*) models from the GAM model data. The rounded edf for each model was selected as the classification target and the cumulative workload, activity label and average processing speed were the independent variables. Figures 5.4 and 5.5 shows the binary tree representation of the model. We expanded the GAM models to include all those significant at the 80% confident level to broaden the dataset.

We observe that for both datasets, the cumulative total workload or the cumulative number of cases completed by a performer were the factors that influenced whether the Yerkes Dodson law is obeyed. Examining the rules that determine the classification for both datasets, it appears that that the Yerkes Dodson law is obeyed when a threshold value is surpassed; otherwise, it is not. For the cumulative number of cases worked, the threshold value was the 28th percentile. Given that this attribute is a proxy for the individual experience of the performer, the results would seem to suggest that less experienced performers tend to obey the Yerkes Dodson law. The other attribute is a proxy for the collective experience of the performers. Given that the threshold value is 32nd percentile, the results would seem to suggest that a less experienced workforce tend to obey the Yerkes Dodson law.

These finding potentially have theoretical and practical implications. From a theoretical perspective, it potentially sheds light on the conditions under which the Yerkes-Dodson law applies and adds to the empirical basis on which the law is built. From a practitioner perspective, the findings have implications on work design, for example, in the design and implementation of an effective load-based work order release system. However, we recommend further research is undertaken to validate this with additional datasets and to test the generalisability of these results.

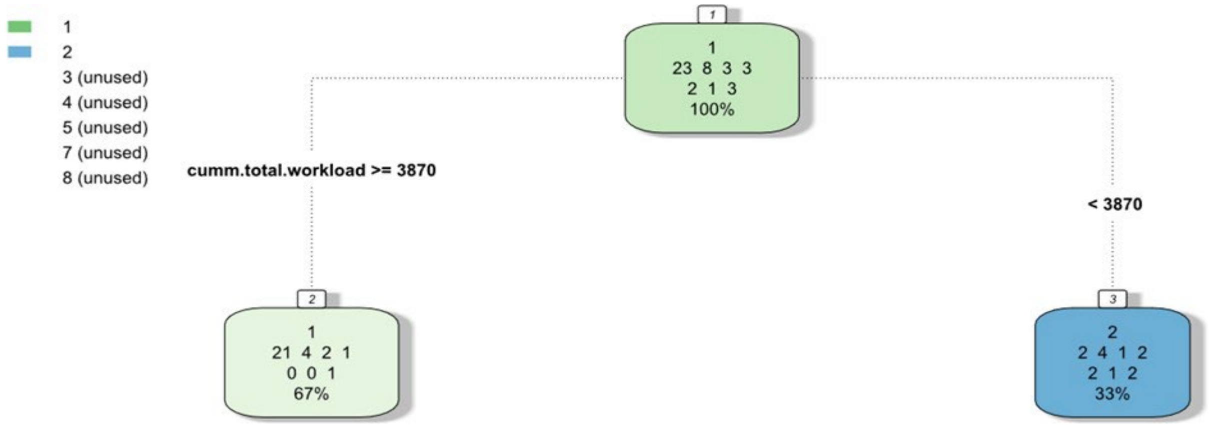


Figure 5.4 - BPIC 12 - Classification tree indicating factors driving degree of non-linearity(edf)

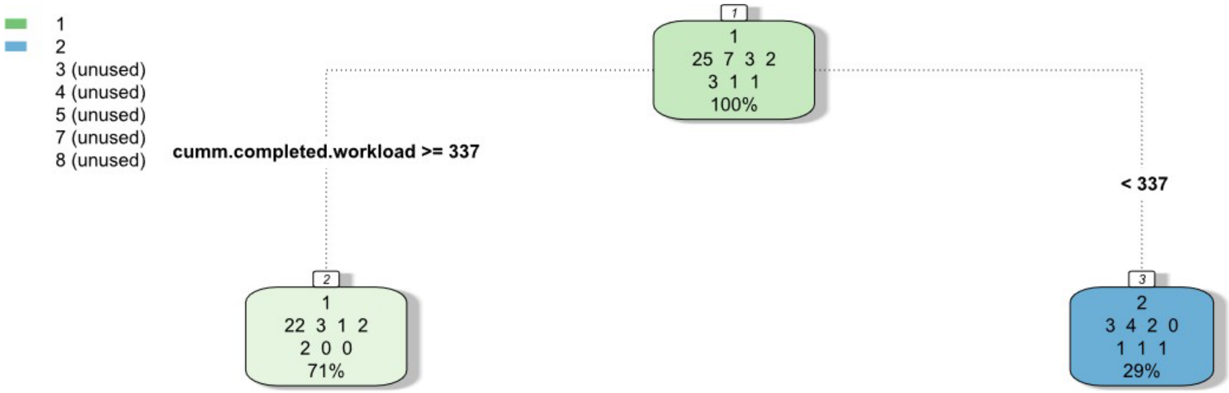


Figure 5.5 - BPIC 17 - Classification tree indicating factors driving degree of non-linearity(edf)

For the second set of results, we examine the effect of varying the infection probability and recovery rate on the number of performers who are stressed (*i.num*) at the end of the simulation run (time step 500). Figure 5.6 shows the plot of incidence and recoveries for infection probability = 0.75 and recovery rate = 0.5. Table 5.4 shows the percentage of the workforce who are stressed at the end of the simulation run as the infection probability and recovery rate are varied.

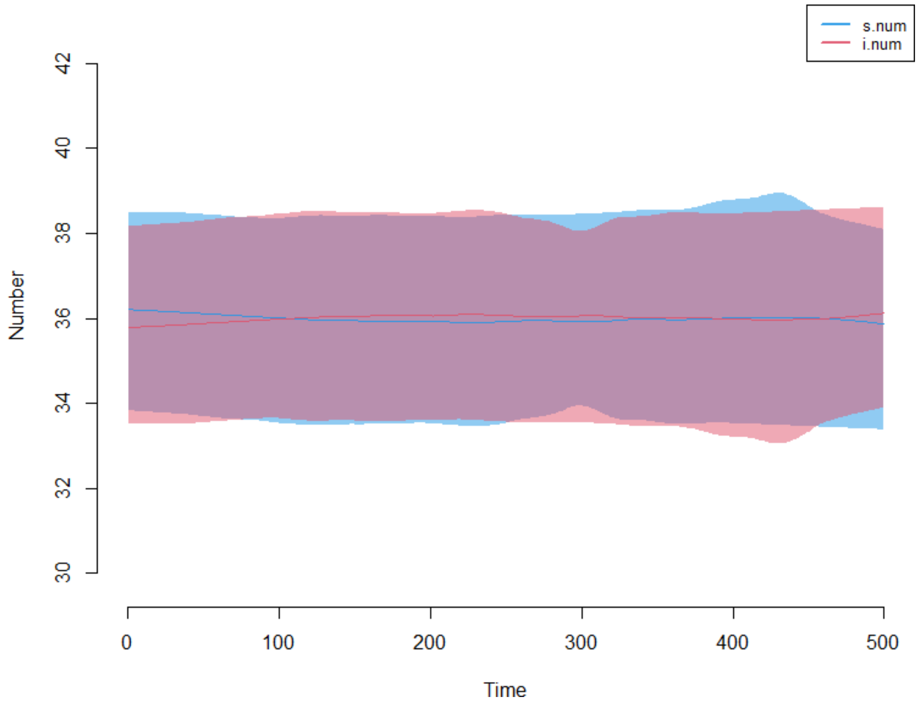


Figure 5.6 - Stress Simulation Model – Incidence and recoveries

Table 5.4 - Stress prevalence as a function of infection probability and recovery rate

Infection Probability	0.25	0.25	0.25	0.25	0.5	0.5	0.5	0.5	0.75	0.75	0.75	0.75	1	1	1	1
Recovery Rate	0.25	0.5	0.75	1	0.25	0.5	0.75	1	0.25	0.5	0.75	1	0.25	0.5	0.75	1
% stressed (at t = 500)	76	50	25	0	74	50	25	0	78	49	28	0	75	50	25	0

We observe that the recovery rate has a more significant effect on reducing the prevalence of stress in the workplace, such that for all values of infection probability, at recovery rate=1, the prevalence of stress in the workplace is eliminated. From a policy perspective, this finding potentially has implications for the allocation of stress management and intervention resources. Whilst intuitive reasoning might indicate that allocating more resources to interventions designed to reduce the infection probability are best (e.g., by making the workforce more resilient to "infection" by stress), the results would appear to indicate that interventions designed to increase the recovery rate (e.g. by engaging in moderate exercise, incorporating mindfulness techniques, etc) are more effective in reducing the prevalence of stress in the workplace (see Burton, Hoobler & Scheuer,2012; Tetrick & Winslow, 2015).

It is worth noting that while, there is a third model parameter that can be varied (*act rate* which measures the mean number of interactions between co-workers), we chose not to vary this as we believe the adverse impact of reducing interaction (resulting in increased isolation and reduction in knowledge diffusion) outweighs any advantages gained by reducing stress prevalence

5.6 Threats to Validity

With regards to the first part of the study, we utilised two real-world datasets. These were the only real-world data that we were able to identify which contained the start and end timestamps for each activity, which was required to calculate the processing time. However, the relatively low number of datasets adversely impacts the ability to generalise these results. We would recommend repeating these experiments with additional real-world data set to validate the results further

In addition, we highlight the propensity of GAMs and recursive partitioning models to overfit data as a limitation to these models as well as the added limitation that GAMs lose predictability when the independent variable is from a range outside of the observed data.

For our simulation model, we utilised a constant quantile applied against the total completed caseload for each performer to determine the stress status for the sake of parsimony. However, based on the results on the first part of the study, we realise that the determination of the stress status of each performer may differ (i.e. linear, quadratic,etc) with different inflection points. We do not believe this significantly affected the results of the simulation as we visually examined the GAM plots for the performer/activity pairs to determine the optimal value of the appropriate quantile. However, we recognise that dynamically determining the optimal value for each performer/activity pair would be best.

Finally, our simulation model explored the effects of a single stressor (. i.e., workload) in isolation. We recognise that in the real-world, multiple stressors exist in the workplace and they are likely to be in play simultaneously (Bickford, 2005). Our model does not consider these non-workload stressors and the interrelationships between them which is likely to impact the performance of the simulation model in a real-world setting (a known limitation of simulation models).

5.7 Summary

This chapter has attempted to uncover the nature of the relationship between workload (a key workplace stressor) and productivity from a couple of real-world event log utilising GAMs. We further explored the factors which drive this relationship. Whilst we found partial evidence for this law in the event log, this was in the minority, with most of the relationships being linear. We also found that the cumulative total workload or the cumulative number of cases completed by a performer are factors that influence whether the Yerkes Dodson law is obeyed and that this happened when a threshold value was surpassed; otherwise, it is not.

In the second part of the paper, we utilised a simulation-based approach to investigate the diffusion of workload-induced stress in the workplace. We found that in terms of stress management intervention, increasing the

recovery rate yields better results vis-à-vis reducing the exposure of the workforce to stress.
As usual, we conclude with some reflections on some of our research choices and their implications. Firstly, we chose to calculate the total daily workload for each activity, the number of activities each performer completed daily and the average processing speed per performer/activity. This was a necessary choice as the activities in the event log had differing mean execution times due to varying complexity. However, we acknowledge that our approach does not capture the true nature of a performer's workload as this typically comprises a mixture of different activities. This is an issue that will have to be addressed prior to utilising our approach in a real-world setting.

Secondly, our findings raise further questions on why the level of experience impacts the Yerkes Dodson law. We also observed that there were different inflection points for each performer leading to the questions: Which factors impact the inflection point? Does the inflection point change over time? These unanswered questions potentially form the basis for future research.

Thirdly, we chose not to vary *act rate* parameter (which measures the mean number of interactions between co-workers). Intuition would seem to suggest that this parameter would be positively correlated with the diffusion of stress; however, it is recommended that this hypothesis is tested. It is worth mentioning that the restriction imposed in certain jurisdictions due to the Covid pandemic (e.g., Work-From-Home directives) has reduced the number of interactions between co-workers.

Finally, we conclude our reflection by observing that there are currently no benchmarks for our predictions, hence it is difficult to assess the validity of the results. It would be worth utilising mixed mode research methods (e.g., ethnographic research, sensors such as smartwatches or trackers to collect physiological data from performers) which can be used to event log data with a view to assessing the validity of the results.

In the next chapter, we consider the ethical dimensions of these predictive workflows. Often human agents use these to make decisions. However, there is a risk that any algorithmic bias present in the models might influence their decisions. We consider the ethical decision-making process when human agents use AI tools (including predictive process monitoring tools) to make decisions and discuss how to design tools which facilitate ethical decision-making.

CHAPTER SIX

6 AN EXPLORATION OF ETHICAL DECISION MAKING WITH AI AUGMENTATION

6.1 Synopsis

The predictive process models (PPMs) discussed thus far are designed to offer operation support to users in real world settings (Van der Aalst, 2016:34). However, the fact that a PPM has been built and deployed does not necessarily translate to adoption by users. A significant barrier to adoption of PPM tools is a lack of understanding by users of the factors that drove the prediction. This is a problem mainly in outcome-based prediction but is also relevant for remaining-time prediction. As a result, there has recently been a focus on explainability in PPM (see Rizzi, Di Francescomarino & Maggi (2020); Galanti et al (2020); Pasquadibisceglie et al (2021)). As the issues addressed in this chapter applies not only to PPM but also to Artificial Intelligence (AI) tools in general, we will use the terms AI and PPM interchangeably.

In recent years, the use of Artificial Intelligence agents to augment and enhance the operational decision-making of human agents has increased. This has delivered real benefits in terms of improved service quality, delivery of more personalised services, reduction in processing time and more efficient allocation of resources, amongst others. However, it has also raised issues which have real-world ethical implications such as predicting different credit outcomes for individuals who have an identical financial profile but different characteristics (e.g., gender, race). The popular press has highlighted several high-profile cases of algorithmic discrimination, and the issue has gained traction.

While both the fields of ethical decision making (in a business context) and

Explainable AI (XAI) have been extensively researched, as yet we are not aware of any studies which have examined the process of ethical decision making with AI augmentation. We aim to address that gap with this study. We amalgamate the literature in both fields of research and propose, but not attempt to validate empirically, propositions and belief statements based on the synthesis of the existing literature, observation, logic and empirical analogy.We aim to test these propositions in future studies.

This chapter formed the basis of a journal paper published in *Social Sciences*

6.2 Background

The use of Artificial Intelligence (AI) agents has gained widespread attention in the last few years (Science and Technology Committee, 2018). As used in this paper, AI refers to "a set of statistical tools and algorithms that combine to form, in part intelligent software enabling computers to simulate elements of human behaviour such as learning, reasoning and classification" (Science and Technology Committee, 2018). These include the predictive process monitoring models we have discussed thus far, though it can be argued that the following discussion is more applicable to outcome-based prediction (as opposed to remaining-time prediction).

One of the prominent uses of AI is to assist human stakeholders in decision making (Abdul, Vermeulen, Wang, Lim & Kankanhalli, 2018). This has been described as 'AI augmentation', as AI models are used to augment the judgement of human agents (S. Miller, 2018). As highlighted by the Academy of Medical Science (2017), AI augmentation has been used in healthcare to enable "clinicians work more efficiently and better handle complex information". It has also been utilised in the criminal justice system to detect crime hotspots and decide whether a suspect could be eligible for deferred prosecution (Oxford Internet Institute,2017), and by financial services providers to determine the outcome of a credit application (Financial Service Consumer Panel, 2017), amongst others. AI augmentation has resulted in significant benefits including improved quality, more personalised service, reduced processing time, and more efficient allocation of resources.

However, several issues have arisen that have raised a cause for concern. For example, several high-profile instances have been highlighted where similar individuals with identical financial data, but different gender have had different outcomes to credit applications (Peachey, 2019). Allegations that AI algorithms used in the criminal justice system discriminated against defendants based on race have also been raised (Maybin, 2016). This algorithmic bias is attributed to unrepresentative or insufficient training data, sophisticated pattern learning which can discover proxies for protected characteristics (e.g., gender, race, sexual orientation, and religious beliefs) -even when these are explicitly moved from the data, amongst others (see Bell, 2016; Murgia,2019). The issue has gained such attention that the UK Parliament Select Committee on Science and Technology commissioned an enquiry to investigate accountability and transparency in algorithmic decision making (see Science and Technology Committee, 2018) The IEEE Standards Association also introduced a global initiative for ethical considerations in the design of autonomous systems (see IEEE,2016). The Association for The Advancement of Artificial Intelligence (AAAI) in its code of conduct acknowledged that "the use of information and technology may cause new or enhance existing inequalities" and urges "AI professional...to avoid creating systems or technologies that disenfranchise or oppress people" (see AAAI, 2019).

In terms of positioning this study, we briefly discuss related studies. The study by Paradice and Dejoie (1991) established that "the presence of a computer-based information system may influence ethical decision making". However, we presume that given that this study predates the recent exponential growth in capability and ubiquity of AI tools, it does not address the peculiar challenges of AI tools in ethical decision making.

Johnson (2015) advances the topic to include artificial agents, highlighting the "push in the direction of programming artificial agents to be more ethical". Martin (2019) extends the discussion further positing that algorithms are "not neutral but value-laden in that they…. reinforce or undercut ethical principles" and highlights that "algorithms are…an important part of a larger decision and influence the delegation of roles within an ethical decision". Martin, Shilton and Smith (2019) argue that "ethical biases in technology might take the form of …biases or values accidentally or purposely built into a product's design assumptions".

This paper aims to contribute to the literature base by synthesising the fields of ethical decision making and Explainable AI (XAI) and proposing, but not attempt to validate empirically, propositions, and belief statements that can be subsequently tested in future studies. These propositions are based on conclusions derived from the existing literature, observation, logic, and empirical analogy. The scope of the study is *AI augmentation* where an PPM tool makes a prediction to a user (who makes the final decision) as opposed to *automation* where autonomous machines make decisions previously entrusted to humans.

A better understanding of how users navigate these ethical issues is of interest in evaluating decisions made by human agents using PPM models regardless of the degree of transparency of the model. Martin (2019) argues that "responsibility for…design decisions [which allow users to take responsibility for algorithmic decisions] is on knowledgeable and uniquely positioned developers". By shedding light on how human agents make decisions with AI models, it would also assist developers with the design of explainable AI (XAI) systems that would assist human agents in identifying ethical issues and dealing with them appropriately. This will serve to improve AI augmentation, which will only increase as more AI tools are deployed in "the wild" (Rubeiro, Singh & Guestrin, 2016).

The issue is relevant and salient as it assists with answering questions about accountability. i.e., who is responsible when a human agent accepts an unethical prediction made by a PPM tool: Is it the human decision-maker or the AI agent? The UK Parliamentary Select Committee report recommends exploring "the scope for individuals" ... "where appropriate, to seek redress for the impacts of such decisions". Some experts are "wary of placing full responsibility on the user of an algorithm" (Klimov, 2019). That would suggest that a degree of responsibility (however small) rests with the user. Other experts suggest that "we may want to assign strictliability [to the user of the algorithm] in certain settings" (see Weller, 2017).

A couple of factors further compounds this issue:

Human users tend to assign traits typically associated with other humans (e.g., intentionality, beliefs, desires) to AI tools (de Graaf & Malle, 2017).

The acknowledgement that these models can process vast amounts of data effectively and discover interactions in the data far beyond a typical human's comprehension (Amoore, 2017).

The combination of these factors increases the likelihood that an unethical prediction by an AI model will be accepted as it is regarded as a trusted expert.

The remainder of the paper is structured as follows: Section 6.3 defines vital terms built on throughout the paper. Section 6.4 discusses the basis for the findings and propositions from the literature synthesis while the final section summarises recommendations and proposes further research areasfor extending these.

6.3 Definitions

6.3.1 Moral agent

A person who makes a moral decision regardless of how the issue is constructed (Sonenshein, 2007). In the context of this study, the moral agent is the stakeholder who decides with the aid of a PPM tool. For example, the Human Resources (HR) officer who determines that a job application should not proceed based on the prediction of a model.The moral agent is also referred to in this chapter as 'the user' of the PPM tool.

6.3.2 Ethical decision

Several studies have highlighted the lack of a widely accepted definition of ethical behaviour (see Cavanagh, Moberg & Velasquez, 1981; Bechamp & Bowie & Arnold, 2004). Rather than base our definition of an ethical decision on consensus (e.g., see Jones, 1991; Trevino, Weaver & Reynolds, 2006) we adopt definitions based on a priori principles, e.g., Kant's (1785/1964) respect principle (see Tenbrunsel & Smith-Crowe, 2008). For example, it is unethical to disrespectfully discriminate against a person based on their ethnicity or gender, while the converse is also true. Smith-Crowe (2004) and Bowie (1999) provides further examples of how Kant's principle is applied in business.

Tenbrusel & Smith-Crowe (2008) argue that unethical decisions could be made intentionally or unintentionally (intended and unintended unethicality). We posit that the use of PPM models has the potential to significantly increase instances of unintended unethicality where a human agent accepts the prediction of a model without realising it may be flawed.

6.3.3 Explainability

The ability of a PPMmodel to summarise the reason for its behaviour or produce insights about the causes of its decisions. Explainable models are also described as "transparent" models. Closely associated with explainability is the quality of explanation, i.e., is the explanation "good" enough? Gilpin, Bau, Yuan, Bajwa, Specter and Kagal (2018) posit that the quality of the explanation can be evaluated by its degree of interpretability and completeness.

6.3.4 Explainer

An agent who supplies an explanation for the recommendation made by itself or another PPM model.

6.3.5 Explainee

A person to whom an explanation is supplied, often in response to a request for an explanation. In this context, the explainee is usually the moral agent who makes the final decision based on the prediction provided by the PPM model.

6.3.6 Interpretability

The ability to describe what the PPM model did (or did not do) in a manner that is understandable to users. As users vary in their level of skills and expertise, interpretability requires the ability to describe in a flexible and versatile manner, tailored to the user's particular mental model. Ribeiro, Singh & Guestrin (2016) make a connection between a user's "trust" in the system and the likelihood of accepting the prediction of the model. They make a distinction between trusting a prediction and the model as a whole. Trust at both levels is predicated on how much the user understands the model's behaviour.

6.3.7 Completeness

The ability to describe the operation of a PPM model accurately. An explanation is more complete if it allows the behaviour of the model to be anticipated in more situations (Gilpin et al.,2018).

6.4 Literature Synthesis

We commence by examining rationalist (or reason-based) models of ethical decision. Numerous models have been proposed that view ethical decision making as a rational process (Ferrell & Gresham, 1985; Rest, 1986; Trevino, 1986; Hunt & Vitell,1986). Perhaps the best-known of these models is that proposed by Rest (1986). This model argues that ethical decision-making progresses through a four-stage process from recognising a moral issue ending with engaging in moral behaviour. Jones (1991) extended this model further to develop an issue-contingent model which argued that the moral intensity of an issue influenced ethical decision making and behaviour. However, Sonenshein (2007) highlights four key limitations of rationalist models as follows: (i) they fail to adequately address the presence of equivocality (i.e., the existence of multiple interpretations) and uncertainty (i.e. lack of complete information) that are present in many real-world scenarios (ii) they presume that ethical behaviour is preceded by deliberate reasoning (iii) they fail to fully emphasize the construction of ethical issues and, (iv) they assume a strong causal link between moral reasoning and judgement. We recognise that we must build on a model that addresses these limitations. For example, due to varying degrees of transparency, equivocality and uncertainty are ubiquitous in decision-making making with PPM tools. Hence, we adopt the Sensemaking Intuition Model (see Sonenshein, 2007), which addresses these limitations, as the foundation on which we build our synthesis.

However, we will first consider how a human user interacts with an explainable PPM agent to obtain explanations and subsequently arrive at a decision (ethical or otherwise - see Figure 6.1). The process commences when the user detects that a prediction it has received from the PPM agent is abnormal. The user subsequently evaluates the explanations (if any) provided by the explainer and selects a subset of it. Depending on how comprehensible or plausible the explanations are, the user may request clarification, which is in turn, evaluated. The user may conclude the explanation/clarification are plausible and accept the prediction or conversely may conclude that the prediction provides evidence of algorithmic bias and reject the prediction.

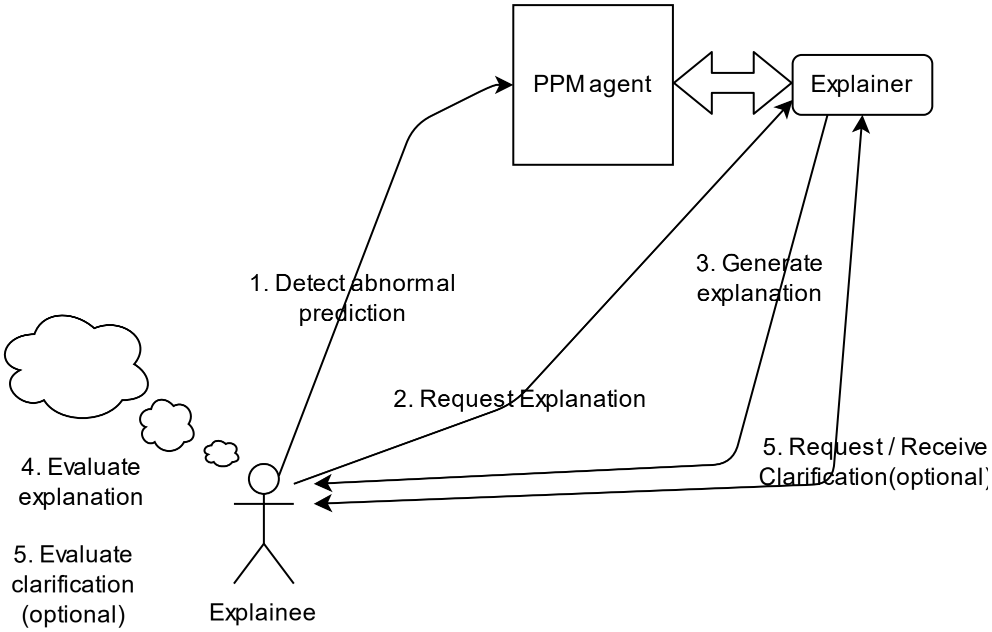

PPM agent
Explainer
1. Detect abnormal prediction
2. Request Explanation
3. Generate explanation
4. Evaluate explanation
5. Request / Receive Clarification(optional)
5. Evaluate clarification (optional)
Explainee

Figure 6.1 - The PPM explanation cycle

Figure 6.2 maps the PPM explanation cycle to the various stages of the ethical decision-making model. Though this is not a precise mapping - for example, the user may request clarification after making an intuitive judgement - the mapping is useful for designing interventions that will increase the likelihood that a user will detect algorithmic discrimination and behave ethically.

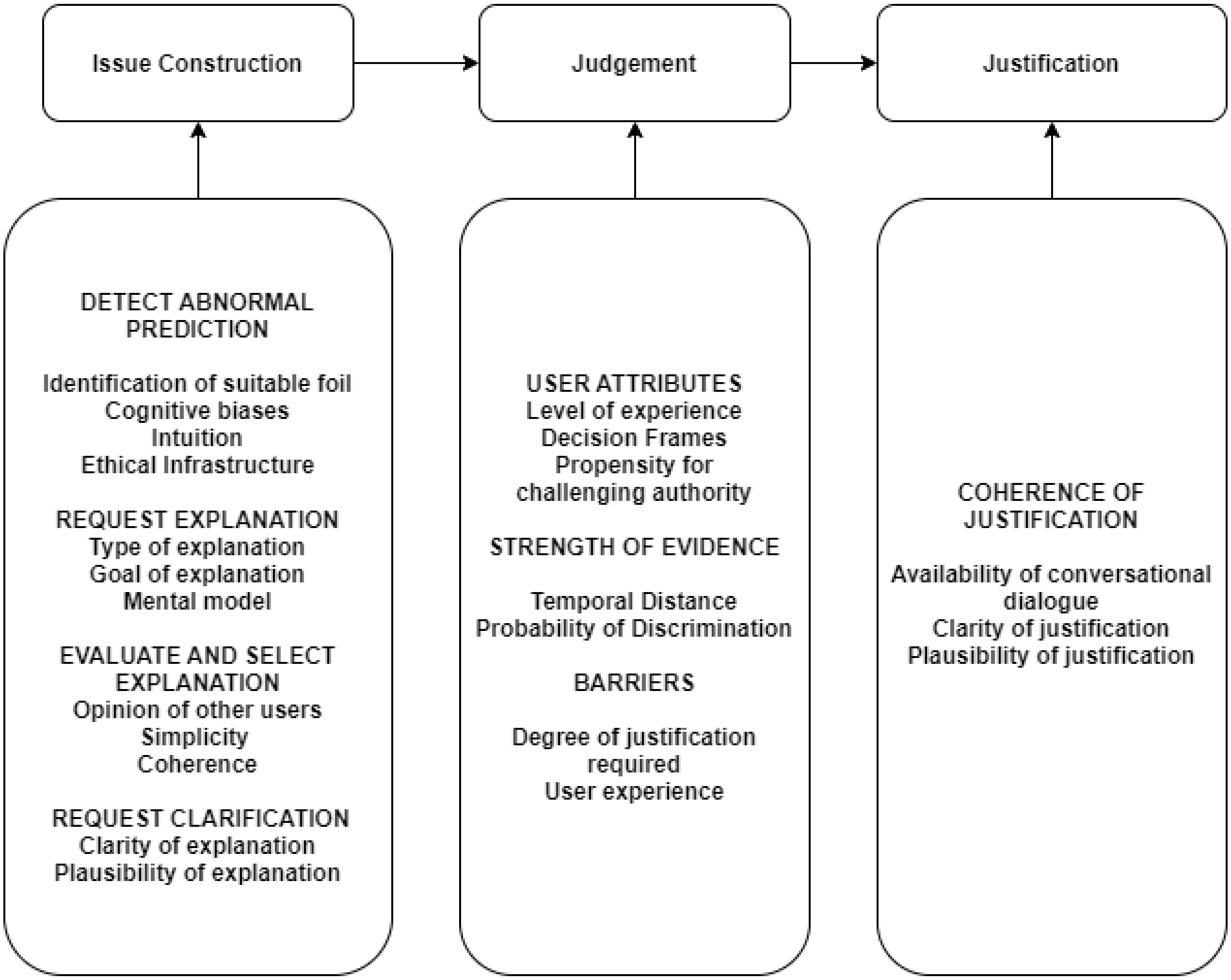


Figure 6.2 - Synthesised Model of Ethical Decision Making with AI Augmentation

In the following section, we describe each part of the model in detail and describe how the explanation provided (or lack of such), impacts ethical decision making.

6.4.1 Issue Construction

6.4.1.1 Expectations

Sonenshein (2007) posits that the ethical decision-making process commences with issue construction where “individuals create their own meaning from a set of stimuli in the environment”.

We argue that in the context of PPM tools augmenting human judgement, the issue construction process is influenced by the degree of transparency of the PPM model. Sonenshein argues that “individuals’ expectation affect how they construct meaning”. In the case of an opaque (or black box) model, whether the user recognises the prediction as “abnormal” will depend on the ***identification of a suitable “foil”***. As established by several studies, people tend to request clarification about observations that they consider unusual or abnormal from their current perspective (see Hilton & Slugoski,1986; Hilton, 1996). Van Bouwel & Weber (2002) argue that establishing abnormality is often done using a contrastive case (also referred to as a foil).Of particular interest in this regard is what they label the O-contract of the form: *why does object a have Property P, while object b has property Q*. To bemore precise; we consider the case where object *a* has Property P=X while an identical object *b* has Property P=Y. To illustrate, consider two individuals, *a* and *b*, identical in all respects except for gender, who both submit a loan application around the same time. However, the model predicts that one individual will default, while the other will not. If the user is aware of the predicted outcomes in both cases, one of the cases will serve as a foil (or counterfactual). The user will utilise abductive reasoning to attempt to determine the cause of the observed prediction (See Peirce,1997). To accomplish this, the user will generate several hypotheses as to the likelycauses of the prediction (one of which is likely to be that there is algorithmic discrimination at play), assess the plausibility of these hypotheses and select the “best” hypothesis. Harman (1965) describes this process as “inference to the best explanation”. If algorithmic discrimination is thought to be the best hypothesis (regardless of whether or not it is the real cause), the user will construct the issue as ethical. However, there may not be a foil readily available, or the user may not be aware of it, in which case the user may not construct the issue as an ethical one.

In the case of a PPM model with any degree of transparency, the trigger for detecting abnormality typically starts with a request for an explanation. Though Miller (2019) posits that curiosity is the primary reason an explanation is requested, we argue that in this context, an explanation is more likely to be requested for regulatory or customer relations management reasons i.e., to justify the decision made to a regulator or customer respectively. However, the issue construction process is dependent on how the system presents the reasons for the prediction, as well on how the user selects and evaluates the explanation. Though most transparent models present their explanation as causal chains or probabilistic models, Miller (2019) argues that "whilst a person could use a causal chain to obtain their own explanation "…" this does not constitute giving an explanation". In terms of explanation evaluation, he argues that "whilst likely causes are good causes, they do not correlate with explanations people find useful". He posits from his review of the literature that there are three criteria people find useful in evaluating explanations: simplicity, generality, and coherence. To illustrate this, consider the case of a user requesting an explanation for a prediction from an explainer such as LIME (Rubeiro, Singh & Guestrin, 2016). The user is presented witha list of features that contributed to the prediction in the order of magnitude of their contribution. The user subsequently assesses whether or not the features that drove the prediction are plausible based on their subject matter expertise. If some unexpected features are driving the prediction, the user may view this as an abnormal prediction. The user may search for a foil (i.e., a similar case), examine whether a similar prediction was made and whether similar unexpected features drove this. If the identical cases have different outcomes, the user may be alerted to the existence of an ethical issue.

Given the contrast between opaque and transparent models as described above, we argue that the more transparent a model is, the more likely it is that the agent will construct the issue as an ethical one. For example, if the user can understand which features made the most significant contribution to a prediction, they are more likely to detect if algorithmic discrimination exists (interpretability). By the same token, because a more complete explanation is likely to shed light on of the system's behaviour, it is likely to make the user recognise the existence of ethical issues than a less complete system.

Proposition 1: The more explainable a PPM model is, the more likely it is that the human agent will construct the issue as ethical as compared to less explainable models.

We argue that the ***type and goal of the explanation*** requested also influences the issue construction process. Initially, the user may request an explanation for the abnormal prediction vis-a-vis the foil. However, where the issue is of high moral intensity (e.g., see Maybin, 2016) or there have been repeated instances of abnormal predictions, the user startsto question the credibility of the entire system. Rather than request an explanation for a specific prediction, they start to ask for explanationsabout the model itself and its learning configuration - a "global perspective" which explains the model (see Ribeiro, Singh & Guestrin,2016). This requires a "model of self" which approximates the original model and exists primarily for an explanation (see Miller, 2019). That paper highlights an example of such an explanatory modification of self from a study by Hayes & Shah (2017).

Numerous studies have demonstrated that the user evaluates and selects a subset of explanations provided by the explainer as relevant based on factors such as abnormality, the contrast between the fact (i.e., observed prediction) and the foil, and robustness, amongst others (see Miller, 2019). Also related to issue construction, the study by Kulesza, Stumpf, Burnett, Yang, Kwan and Wong (2013) explored the link between the soundness (or correctness) and completeness of the explanation. They recommended that while completeness was more critical than soundness; itwas important not to overwhelm the user. Miller (2019) also argues that when the entire causal chain is presented to the user, there is a risk that the less relevant parts of the chain will dilute the crucial parts that are important to explain the prediction. This recommendation runs contrary to the intuitive view that more information is better than less.

Proposition 2: The more interpretability (as opposed to complete) the supplied explanations are, the more likely it is that the human agent will correctly construct the issue as a moral one.

6.4.1.2 Motivational drive

Tenbrusel and Smith-Crowe (2008) argue that "***biases, intuition and emotion*** .. must be considered" in the ethical decision-making process. This aligns with the position put forward by Sonenshein (2007) that "individuals see what they expect to see, but …also see what they want to see". We consider the implications of these biases on ethical decision-making using PPM tools.

Messick and Bazeman (1996) postulate that internal theories influence the way we make decisions. Strudler and Warren (2001) provide an example of one such bias (authority heuristics) which describes the trust we place in the expertise of authority figures which may be misplaced. As we argued earlier, humans tend to view AI models as authority figures. However, as highlighted earlier, the user is likely to bring biases into the evaluation of explanations provided based on their perceived intention of the explainer (see Dodd & Bradshaw, 1980).

Abnormality is a critical factor in ethical decision making as it triggers the request for an explanation regarding the basis for the model's prediction. (Miller, 2019). However, what is viewed as abnormal is subject to cognitive biases held by the user. For example, Gilbert and Malone (1995) highlight correspondence bias due to which people tend to explain other people's behaviour based on traits. In other words, a user may not view a model's prediction as abnormal due to discriminatory tendencies they may harbour or may give higher weight to unimportant causal features that support biases. It is also possible that the user may select a conjunction of facts in the causal chain and assign them higher weighting that they deserve because it aligns with their preconceptions (see Tversky & Kahneman, 1983).

Proposition 3: The more aligned a human agent's biases are with the PPM model's recommendations, the less likely they are to construct the issue as a moral one.

6.4.1.3 Social anchors

We posit that the existing "***ethical infrastructure***" in the organisation is another important factor that impacts the issue construction process when making decisions with AI tools. Ethical infrastructure refers to “organisational climate, informal systems and formal systems relevant to ethics” (Tenbrunsel, Smith-Crowe & Umphress, 2003). Where the ethical infrastructure supports constructing moral issues regarding the existence of algorithmic discrimination, the user is likely to do so; otherwise, they will not.

Sonenshein (2007) argues that “employee’s goals …will affect how they construct an issue”. They may view unethical behaviour as consistent with “rules of business” if it enables them to achieve their goals. This is consistent with results from Schweitzer, Ordonez and Douma (2004), which concludes that goal setting is negatively associated with ethical behaviour. This conclusion aligns with the findings by Hegarty and Sims (1978) and Tenbrunsel (1998), which discovered a positive correlation between incentives and unethical behaviour. In terms of decision making with PPM tools, this suggests that if an organisation sets goals that encourage specific outcomes based on PPM tools without taking appropriate action to manage undesirable side effects, users are less likely to construct ethical issues appropriately.

Proposition 4: Users working with PPM models in organisations with more supportive ethical infrastructures are more likely to challenge the model’s recommendation compared to users in organisations with less supportive ethical infrastructures.

6.4.1.4 Representation

Sonenshein (2007) posits that the user’s representation – their “***mental model***…of how others see a situation” – is also an important moderator of issue construction. We argue that a significant “other” in decision making with PPM tools is the explainer which can shed light on the factors that drove the prediction made by the PPM agent. Sonenshein quotes a study by Weick and Roberts (1993), which found that people engage in representation through communication with others. This highlights the need, not only for the user to be able to request an explanation, but also about the ***nature of the explanation*** provided. For example, though the explanation provided could include details such as the training data, optimisation cost function, hyperparameters, etc., these are not likely to be useful to the typical user. It is worth noting that the user is likely to be distrustful of the explanation offered. This conclusion is based on research which indicates that individuals tend to prejudge the intention of the explainer and filter out information that supports the prejudgement (see Dodd & Bradshaw, 1980). As a result, the user is likely to request clarification and additional explanation of any explanations provided. This supports the requirement for an explanation as dialogue, which facilitates challenges from the explainee ("I do not accept your explanation or parts of it").

The other form of representation that is relevant, is the ***opinion of other users*** of the PPM tool. Sonenshein (2007) refers to these as "social anchors…interlocutors who help an actor test his or her interpretation of social stimuli". This suggests that the way a user constructs the ethical issue in isolation is likely to be different from the manner it will be constructed if done collaboratively with other users. In the latter scenario, additional users are likely to be able to provide additional examples of foils which will broaden the initial user's frame of reference and enable them to construct the issue in a broader manner.

Proposition 5: The more collaborative the issue construction process is, the more likely it is that the user will correctly construct an issue as ethical as compared to issue construction done in isolation.

6.4.2 Judgement

6.4.2.1 User Attributes

Sonenshein (2007) posits that the intuitive judgement stage directly follows the end of the issue construction stage with the user reaching a plausible interpretation. At this stage, the actor responds to the issue (as constructed in the initial stage) using intuition - an "automatic, affective reaction". With regards to ethical decision making with PPM tools, the agent makes an intuitive judgement which rules the prediction as discriminatory or non-discriminatory.

Sonenshein (2007) further argues that an individual's ***level of experience*** is a key factor that influences their judgement, stating that "as individuals develop experience, they can internalise that experience into intuitions. Regarding ethical decision making with PPM tools, this implies that a less experienced user is likely to challenge the prediction of the PPM tool. They are likely to have less experience of abnormal predictions and as such are more likely to view the PPM tool as an expert, the converse of whichis true of more experienced users.

Also relevant at this stage is the work of Tenbrunsel and Smith-Crowe (2008). The authors introduce the concept of ***decision frames*** that illuminate the perspective of the decision-maker and is moderated by previous experience. For example, suppose the decision-maker primarily adopts a business or legal frame. In that case, they are less likely to judge a model's prediction as discriminatory (even if there exists evidence to the contrary). Though Sonenshein suggests that individuals infrequently alter their initial judgement after it has been made, we argue that exposing the basis on which a PPMmodel made a prediction can shift the user's decision frame such that what may have been perceived at the outset as a business/legal decision is transformed into an ethical one. We believe this is especially the case for more morally intense issues. e.g. if the PPM model is being used to determine the risk of reoffending for an offender (Maybin, 2016).

We argue that the degree of interpretability also influences the user's judgement. For example, given an opaque model and the absence of a suitable foil, the user is unlikely to judge the model's prediction as discriminatory. This is backed up by findings which relate the degree of perceived control over an event with attribution of responsibility (Fiske & Taylor,1991). In other words, a user deciding based on a prediction from a black box model is likely to attribute the decision to the model ("Computer says 'No'") as opposed to a prediction based on an interpretable model where the user is more likely to perceive that they have more control.

Kelman and Hamilton (1989) argue that an individual's ***propensity for challenging authority*** (i.e., the model's prediction) depends on which is the more powerful of two opposing forces in tension - binding and opposing forces. Binding forces strengthen authority existing structures while opposing forces intensify resistance to authority. As earlier stated, human users tend to view the model as an 'authority' due to their data computation ability. We argue that interpretability is likely to heighten the opposing force and make the user more likely to challenge the prediction where it is abnormal. In addition, it will also make it easier for the user to justify their rationale for disregarding the model'srecommendation.

Proposition 6: The more experienced a human agent is, the more likely they are to correctly judge a model's prediction as discriminatory.

6.4.2.2 Strength of Evidence

In addition, we argue that for an explainable system, the strength of the evidence provided to support the explanation will influence the judgement the user reaches. Below we highlight a couple of factors in this regard.

Miller and Gunasegaram (1990) argue that the ***temporal distance*** of events is an important moderating factor, specifically that people tend to "undo" more recent events. As it pertains to ethical decision making with AI, this would suggest that the user is unlikely to recognise the foil as valid if it is sufficiently temporally distant. Even if the case currently being assessed and the identified foil have identical properties, the user is likely to intuitively feel that due to the passage of time, changes to legislation, policies, and procedures, etc., treating the case as identical is not feasible. As a result, they may not dismiss evidence that points towards algorithmic discrimination, even if the user has some suspicion about the unethicality of previous decisions.

We posit that the use of ***probability*** in explanation, primarily when used to explain the causes of the prediction will increase the likelihood of correctly judging an algorithm's prediction as discriminatory (see Josephson & Josephson, 1996). The study by Eynon, Hills and Stevens (1997) would also appear to indicate that where ethical training is available, especially when it is tailored to the use of AI tools, users are likely to correctly judge a model's prediction as discriminatory.

Proposition 7: The more substantial the evidence presented to support an explanation, the more likely it is that the human agent will correctly judge the model's prediction as discriminatory as compared to weaker evidence.

6.4.2.3 Social pressures

Regarding the influence of social pressure on forming ethical judgements, Sonenshein argues that “organizations strongly influence how their members behave and what they believe”. We posit that in terms of ethical decision making, a key influencing factor is the ***design*** of the AI tool and associated processes. Martin (2019) refers to these as “affordances – properties of technologies that make some actions easier than others”. The higher the technological hurdle the user must clear, the less likely they are to adjudge the model’s prediction as discriminatory. For example, if the user has to perform more operations (e.g., navigate to different screens, click multiple buttons, etc.) in order to reject the AI tool’s prediction and provide a significant amount of mandatory justification (vis-a-vis accepting the recommendation), then the design of the tool or process is likely to influence their judgement. This concept has been acknowledged in“Values in Design” (ViD) which describes the field of research that investigates how “individually and organizationally held values become translated into design features” (Martin, Shilton and Smith, 2019).

Proposition 8: The more complicated the design of the tool and associated processes make it to judge the AI tool’s prediction as incorrect, the less likely it is that the human agent will do so.

6.4.3 Explanation and justification

Sonenshein (2007) posits that the judgement phase is followed by the explanation and justification phase, where the moral agent attempts to explain and justify their reaction to the constructed issue. We refer to this stage simply as the justification stage to avoid any confusion with the point(s) in the construction stage where the explainer provides an explanation for the prediction. Sonenshein (2007) further argues thatmoral agents “employ the rules of rational analysis” to “bolster their confidence in the decision” as well as that of others. This reinforces the recommendation by Miller (2019) for the adoption of a ***conversational mode*** of explanation. The dialogue between the user and the explainer preserves an audit trail of the process by which the user constructed the issue and reached their judgement. Making this conversation readily available to the user for review also has the added advantage of highlighting inconsistencies in the issue construction process (e.g., implausible arguments). This highlights the requirement for social interaction between the explainee (i.e. the user) and the explainer.

Proposition 9: The more conversational the dialogue between the explainer and the user, the better the quality of justification the human agent can provide for their judgement.

6.5 Summary

Based on the preceding, we conclude with several non-exhaustive recommendations to assist users in making more ethically sound decisions when using PPM tools. First, we recommend that the explanation provided by the explainer pre-empt the user's request for an explanation for abnormal predictions and make that available. Though Hilton (1990) recommends providing a contrastive explanation vis-a-vis a 'typical' case, we suggest thatthe explainer should select an appropriate foil with identical properties anddifferent predictions and explain why the predictions were different. The user should also have the ability to replace the system-selected foil with another they choose and request a contrastive explanation for this. Though Miller (2019) suggests that an unprompted explanation could prove superfluous and distracting over time, we recommend that these explanations could be presented as 'hints' where the details remain hidden, but which can be readily accessed as required by theuser. We argue that if a user correctly constructs and judges an issue as ethical early on, they are more likely to engage in moral behaviour and as such, investing in the design to highlight such issues is likely to drive desired behaviour.

Secondly, we propose the provision of explanations at different levels to facilitate ethical decision making. Apart from the explanations for each prediction, the tool should be able to explain its model of self (see section 6.4.1.1). Besides, the tool should also support the ability to request clarification on any section of the causal chain.

Thirdly, based on the conversation model of explanation (see Hilton,1990), the explanation should be presented in a manner that follows the "basic rules of conversation". This would include only presenting information relevant to the user based on their mental model (see also Jaspars & Hilton, 1988), keeping track of which information has already been shared (based on the premise that once something has been learned, it should not need to be explained again) and whether the user accepted it as credible or not, etc. It could also support presentation modes such as chatbots which facilitate conversational dialogue.

Fourthly, we recommend providing the capability for the user to collaborate with other users of the tool in the decision-making process. The user could share details of the case and the model's prediction with one or moreusers and request their opinion(s). This would help to widen the initial user's frame of reference and could make them aware of more suitable foils. It would also assist in raising their experience level as the collective experience of all the collaborators will be utilised in the decision-making process.

Finally, we recommend that the process for rejecting the PPM model's prediction and highlighting the potential existence of ethical issues should be streamlined as possible and should not be more complicated than the process for accepting the model's prediction. This will increase the likelihood that the user will follow through on any moral intent they had previously established.

In this paper, we have synthesised the literature on ethical decision making and explainable AI and proposed several testable belief statements. Though we do not present empirical evidence for these in this paper, we expect that these will be tested and empirically validated in future studies utilising a variety of appropriate methodologies. For example, there exist opportunities to utilise technology such as functional Magnetic Resonance Imaging (fMRI) to monitor the brain activity of users as they make decisions using AI tools, undertake a phenomenological study to gather rich data among real-life practitioners, amongst others. In future work, we intend to attempt to tackle a number of these opportunities.

In the concluding chapter we summarise the key findings from the thesis and propose future work to build on the research we have done.

7 CONCLUSION AND FUTURE WORK

7.1 Conclusion

In this thesis we have discussed contextual and ethical issues with the predictive process monitoring framework and proposed ways to address these.

We started by describing the motivation for the thesis, i.e., a dearth of research that assesses the relative importance of contextual factors in the predictive process monitoring workflow. We also mentioned our objective to shed light on the nature of the relationship between contextual types, specifically process and social context. Finally, we expressed our aim to uncover the manner human agents use the predictions generated by process monitoring and other Artificial Intelligence (AI) workflows.

We subsequently detailed the findings of a systematic literature review we undertook to identify existing studies which adopted a clustering-based remaining-time predictive process monitoring approach and a comparative analysis to compare and benchmark the output of the identified studies using 5 real-life event logs. We identified five pertinent approaches and found that the approach that clustered traces based on activities (similar to state-based clustering) performed best (See Section 2.8). We also detailed the process for performing a Systematic Pre-Review Mapping (subsequently referred to as SPRM) process to determine the degree of overlap between existing studies and this review. Whilst the literature base recommends performing this process, we were unable to identify any papers in the computing science field which had implemented it.

The findings from this chapter are important in providing guidance for researchers and practitioners for selecting a clustering approach for their predictive process monitoring workflow. It may be argued that clustering-based approaches are outdated and have been overshadowed by neural network-based approaches (which have become very popular in recent years). Whilst the latter make it relatively easy to include additional attributes into the prediction model and generally provide more accurate predictions, they are also less explainable and interpretable than clustering-based approaches. As highlighted in chapter 6, predictive models which are less explainable are generally less trusted. We argue that as it is essential that the predictive process monitoring tools deployed in real world setting are trusted, clustering-based approaches still have an important role to play in the family of approaches.

However, accuracy is also an important feature to consider as an explainable model with poor accuracy will not be of much use to a user. Interestingly, the best performing algorithm (i.e., *data driven*) was based on association rule clustering which clustered traces based on the activities executed. This technique treats the set of activities as a "basket" and does not consider the order or sequence of the activities. In other words, σ_i={a,b} is equivalent to σ_j={b,a}. The same is true, though, of the frequency encoding technique required for the other approaches, which does not capture sequence information in the feature vector used for clustering. Saxena et al (2017) discussing choosing appropriate clustering techniques, argue that "no algorithm can be uniformly good under all circumstances…. each algorithm has its merit (strength) under some specific nature of data but fails on other type of data." It appears that, for the encoded traces, the association rules technique does a better job of minimising the distance between traces in the same cluster and maximising the distance between clusters, hence grouping similar traces together and resulting in better accuracy.

Subsequently, we investigated the impact of social contextual factors in the predictive process monitoring framework and proposed a survival analysis approach for predictive process monitoring. We found that group betweenness and closeness centrality were generally positively correlated with process cycle time while the group eigenvector centrality was generally negatively correlated. We also found that survival analysis approach performed comparably with start-of-the-art predictive process monitoring techniques.

The findings from this chapter have implication for team formation and management. Firstly, it invites a rethink of the definition of a team. Returning to the public service example we have utilised in earlier chapters, rather than simply adopting a functional perspective (e.g., Road Maintenance, Parks, etc), we adopt a case (or process instance) perspective and define our team as the group of performers involved in executing the activities on the case (including the citizen who raised the initial request). This will result in the creation of numerous temporary, transient teams working together to complete a particular case and which "disband" when the case completes. The concept of temporary teams is well established and is a mature field of research. For example, criminal investigations are often undertaken by temporary law enforcement teams set up with the sole aim of completing that investigation. However, unlike temporary law enforcement teams, with our approach, team members not typically preselected. Often the performer is selected just before the activity is due to be executed with availability of the performer being the key factor influencing selection. Our research findings would suggest that the performers for each case should be preassigned with the group centrality measure for the group of performers being the key factor for selection. In other words, given a network of handover-of-work ties, how can we design a team to achieve the desired process outcome(s)?

The main advantage of this approach is that it increases the likelihood that the desired process outcomes will be achieved due to the intentionality of the team's selection. On the other hand, we should bear in mind that, rather than being based on actual historic data, the handover-of-work graph is forward-looking as it is based on proposed (rather than actual) handovers of work. If a performer is unavailable because of sickness, resignation, etc, the proposed handover-of-work graph will have to be adjusted to reflect reality.

It is also worth calling out that the ties between team member is not only weak but also temporal in nature as a performer is an "active" team member when they are working on the case but is inactive before and thereafter. Often a performer is a member of multiple teams at any point in time as they are working on multiple cases. The implication of this is that there may be a requirement for a case manager role who manages the case teams in collaboration with functional managers. The case manager would be responsible for team formation based on group centrality measures.

Geletkanycz and Hambrick (1997) posit that "strategic choices are affected by the external ties of top management team members" and "the informational and social influences arising from external interactions will be reflected in strategic profiles". By extension, we argue that the decision and choices performers make as they work on cases is likely to be affected by ties with non-group members. However, there is a recognition that our approach only recognises handover-of-work. There are other interactions between performers that are just as important but are not captured (e.g., watercooler chats, email exchanges, etc). We acknowledge this as a limitation, in that our approach only considers handover-of-work which is captured as digital traces in a Process-Aware-Information-System (PAIS) but not other interactions which may be just as important, but not captured or analysed.

Next, we explored the impact of spatial context (a type of external context) on the predictive process monitoring workflow. We introduce the concept of a spatial event log and propose an approach for incorporating the spatial context into the predictive workflow. We demonstrate that the spatial context improves the predictive power of business process monitoring models.

In our opinion, the main challenge with the adoption of spatial context in the predictive process monitoring workflow is increasing the collection of real-world spatial data. The technology for collecting spatial data is readily available, as devices such as IoT sensors, RFID tags and location trackers, among others, have become more mainstream. The capacity to collect spatial data is ubiquitous as most basic smartphones are enabled to collect spatial data. As a result, there may currently be suitable spatial data available which could be combined with existing event logs to create spatial event logs. However, incorporating spatial context into the process mining / monitoring workflow will require an appreciation by the process mining community (both academic and practitioners) of the utility of spatial data, not only for prediction, but for the other use cases mentioned in section 4.2. We acknowledge that the utility of spatial features *in isolation* is limited; however, we believe that in conjunction with other features it can be transformational.

As this section concludes the use of contextual features to predict remaining time, we acknowledge that we have studied contextual types in isolation. However, we realise that there are opportunities to study contextual types collectively, especially the interaction between them. To illustrate, if the location of process performers is recorded and one of them tests positive for Covid, it would be possible to determine which performers had been within a certain threshold distance (e.g., 2 m) of that performer and require them to isolate, rather than require all performers to isolate (i.e., interaction between social and spatial context). In that scenario, it would also be possible to determine the availability of performers for working on inflight cases (i.e., process context). However, ethical factors should be taken into consideration in the collection and management of this data.

We subsequently examined the nature of the relationship between workload (a process contextual factor) and stress (a social contextual factor). We found partial evidence for the Yerkes-Dodson law; however, this was in the minority in the event logs we examined, with the majority of the relationships being linear. We also found that the cumulative total workload or the cumulative number of cases completed by a performer are factors that influence whether the Yerkes Dodson law is obeyed and that this happened when a threshold value was surpassed; otherwise, it is not. Recently there has been an increased focus on mental well-being. Studies such as Williams, Michie & Pattani (1998) and Pines & Aronson (1988) highlight workload as a leading cause of stress in the workplace. Our findings suggest that in terms of stress management intervention, increasing the recovery rate yields better results vis-à-vis reducing the exposure of the workforce to stress. It is worth clarifying that this finding does not suggest that workplace stress prevention efforts should be eliminated or reduced. Rather, it recommends that in allocating resources, more resources should be allocated to stress recovery efforts than to prevention. However, we acknowledge that this finding may have ethical implications as it may be perceived as an increased acceptance or tolerance for workplace stress.

Our findings also suggests the implementation of a sophisticated workload management system that monitors each performer's workload and processing speed, detects their inflection point, and provides them the right quantity of work to manage their stress level. However, the impact of holding back work on process & business outcomes should also be considered, especially in terms of increased delays and wait time. If the implementation of the proposed workload management system is proven to add delays to processing time, a trade-off may have to be struck between a healthier and more productive workforce, on the one hand, and accepting a certain amount of delay to processing time.

Finally, we investigated ethical decision making with AI tools. We proposed a model of ethical decision making with AI augmentation and made a number of recommendations to enable the design of AI tools which facilitate ethical decision making.

As the recommendations in this chapter are based on yet-to-be tested propositions, there is a need to test these with a view to validating them. We believe that there is scope to test these using multi-disciplinary approaches. For example, experiments could be designed utilising brain monitoring devices such as electroencephalography (EEG)) to measure brain activity and other parameters (such as where the gaze is focused) when a user is presented with a contrasting foil with a view to determining their thoughts and emotions as they cycle through issue construction, judgment, and justification. These experiments could provide very rich insight into the ethical decision-making process with PPM tools.

However, once these propositions are validated, consideration should be given to developing and rolling out industry-wide frameworks based on them with a view to assisting users to identify algorithmic discrimination. As highlighted in section 6.2, there is an increasing acknowledgement by legislators, the academic and practitioner community, among others, of the harm caused by algorithmic discrimination, and this has resulted in focused attention to limit the harm. We argue that, in the same way that standards, frameworks and guiding principles, etc have been established to facilitate good design in industries such as software and construction, similar scheme(s) could be set up to facilitate the design of PPM tools that adhere to these recommendations. This could be taken further and incorporate licensing or certification such as those utilised for building safety-critical systems (e.g., used in aircrafts or nuclear power plants). It could be argued that the latter position is extreme, however so is harm caused to a person given a more severe prison sentence or denied an urgently needed loan because a user acted on the incorrect prediction of a PPM tool.

7.2 Future Work

In this concluding section, we recommend future research that builds on and extends the work done in this research project.

We propose a couple of areas for further research to extend the work done on the role of social context in predictive monitoring (Chapter 3). Firstly, as an increasing number of automated agents augment the human workforce, we believe there is value in exploring how social networks between these categories of workers differ and the implications of these. Secondly, we propose an exploration of the effect the feature values of neighbouring nodes have on behaviour. For example, we could explore the temporal effect of a high workload on workers in an individual or group's neighbourhood. This would enable a better understanding of workload distribution in the social network over time.

With regards to the role of spatial context in predictive monitoring (Chapter 4), as mentioned in Section 4.2, incorporating the spatial context into the event log facilitates research opportunities which extend beyond predictive process monitoring. Referencing the refined process mining framework earlier mentioned (see Section 1.1), it 'opens the door' to performing spatial process discovery (process models by location) and conformance testing. For 'Recommend', it would be possible to incorporate spatial context into the recommendation (i.e., the model recommends a user in location A performs activity X; however, suggests a user in location B performs activity Y).

Besides, a spatio-temporal extension to Tobler's law is proposed as follows: "everything is related to everything else but near and recent things are more related than distant things" (Bennet & Vale, 2018). As a result, we expect that a spatio-temporal model will make a more significant contribution to remaining-time predictive monitoring.

Finally, we utilised the Euclidean buffer distance as the geographic covariate in this study. In subsequent studies, we recommend investigating whether other distance measures (e.g., the sum of distances between e_i and e_{i+1}) will yield better results.

Considering the relationship between workload and stress (Chapter 5), wepropose several areas for further exploration. Firstly, we recommend that the study is repeated with different datasets and methodologies with a view to replicating the results and triangulating the conclusions. This would shed some more light on the generalisability of the results.

Secondly, if the thresholds we identified are replicated in further studies, we propose an exploration of why the identified factors (cumulative total workload or the cumulative number of cases completedby a performer) impact the Yerkes Dodson law and why these thresholdsoccur where they do.

Thirdly, we propose the development of more sophisticated stress simulation models. For example, if the GAM model indicates a non-linear relationship between the workload and productivity, the simulation model could dynamically determine the inflection point at which productivity reduces. In addition, the model could simulate the impact of the prevalence of stress on variables of interest (e.g., on overall and individual productivity). Additionally, the model could factor in additional stressors such as the pace and variety of work and shift patterns – all of which can be derived from the event log – and model the interrelationships between them. Finally, we propose adding spatial context to the event log. This would enable us to calculate the distance between performers and better model the interaction between them. We can subsequently utilise these interactions to model the spread of information, disease, etc in the workplace.

Finally, with regards to the examination of ethical decision-making using AI augmentation, we proposed several belief statements but did not present empirical evidence for these. We expect that these will be tested and empirically validated in future studies utilising a variety of appropriate methodologies. For example, there exist opportunities to utilise technology such as functional Magnetic Resonance Imaging (fMRI) to monitor the brain activity of users as they make decisions using AI tools, undertake a phenomenological study to gather rich data among real-life practitioners, amongst others. In future work, we intend to attempt to tackle a number of these opportunities.

8 GLOSSARY

Term **Definition**

BPM

Business Process Management is the discipline that combines approaches for the design, execution, control, measurement and optimization of business processes. (Van der Aalst, 2016:44)

BPMS

Business Process Management System. A platform or system that facilitates management of business processes (see above)

Case

A process instance. For example, for the process 'Request Waste Collection' each individual requests corresponds to a case. (Van der Aalst, 2016:32)

Data Mining

The analysis of (often large) data sets to find unsuspected relationships and to summarize the data in novel waysthat are both understandable and useful to the data owner (Van der Aalst, 2016:12)

de jure model

A normative model which specifies how things should be done or handled (Van der Aalst, 2016:303)

de facto model

A descriptive model whose goal is not to steer or control reality. Instead, de facto models aim to capture reality (Van der Aalst, 2016:303)

Event

An activity (i.e., a well-defined step in the process) and is related to a particular case (i.e., a process instance) (Van der Aalst, 2016:32)

LSS

Lean Six Sigma is a methodology that combines ideas from lean manufacturing and Six Sigma to improve performance by systematically removing waste and reduce variation in process outcomes. (Van der Aalst, 2016:46)

Post-mortem

Information about cases that have completed, i.e., these data can be used for process improvement and auditing, but not for influencing the cases they refer to. (Van der Aalst, 2016:302)

Predictive power

A measure of the ability of a model to correctly predict a variable of interest

Pre-mortem

Event data for cases that have not yet completed. As such information in the event log about this case (i.e., current data) can be exploited to ensure the correct or efficient handling of this case. (Van der Aalst, 2016:302)

Process Aware' Information System

All software systems that support processes and not just isolated activities. There is a process notion present in the software (e.g., the completion of one activity triggers another activity) and that the information system is aware of the processes it supports (Van der Aalst, 2016:27)

Process Model

A description of the activities and ordering of these activities required to achieve process outcomes. The process model may also describe temporal properties, specify the creation and use of data and stipulate the way thatresources interact with the process (Van der Aalst, 2016:26)

Social Network Analysis

The application of the broader network of network science to the study of human relationships and connections (Hansen, Shneiderman and Smith, 2011)

Subcontracting

A subset of Social Network Analysis concerned with the transfer of work. The main idea behind the subcontracting metric (which measures this concept) is to count the number of times individual *j* executed an activity in between two activities executed by individual *i* (Van der Aalst et al,2007)

Transition System

The most basic process modelling notation. A transition system consists of states and transitions. Each transition connects two states and is labelled with the name of an activity (Van der Aalst, 2016:58)

9 BIBLIOGRAPHY

AAAI (2019). AAAI Code of Professional Ethics and Conduct. https://aaai.org/Conferences/code-of-ethics-and-conduct.php [Accessed 11 April 2020]

Abdul, A., Vermeulen, J., Wang, D., Lim, B.Y. & Kankanhalli, M. (2018) April. Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. *Proceedings of the 2018 CHI conference on human factors in computing systems*:1-18.

Academy of Medical Science (2017) *Algorithms in decision-making*, 2018, May 23. HC 351, 2017-2019, ALG0055

Akritas, M.G. (2004) Non-parametric survival analysis. Statistical Science, pp.615-623.

Amoore, L, (2017) *Algorithms in decision-making* (2018), May 23. HC 351, 2017-2019, Q4

Ancona, D.G. (1990) Outward bound: strategic for team survival in an organization. *Academy of Management journal*, *33*(2), pp.334-365.

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

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

Aslan, A. (2017) Combining Process Mining and Queueing Theory for the ICT Ticket Resolution Process at LUMC (Master's thesis, University of Twente).

Beauchamp, T.L., Bowie, N.E. & Arnold, D.G. eds. (2004) *Ethical theory andbusiness*. London, UK: Pearson Education.

Bell, E (2016) Dec 27. Controlling the Unaccountable Algorithm, BBC. https://www.bbc.co.uk/sounds/play/b085wj18 [Accessed 23 March 2020]

Bennett L., D'Acosta J. and Vale F. (2018) in 'Spatial Data Mining II: A Deep Dive Into Space-Time Analysis' [Online] Available at https://www.youtube.com/watch?v=0aV6HHwJuo4&t=364s [Accessed 30 May 2019]

Bertrand, J.M. and Van Ooijen, H.P.G. (2002) Workload based order release and productivity: a missing link. *Production Planning & Control,* 13(7), pp.665-678.

Bevacqua A., Carnuccio M., Folino F., Guarascio M., Pontieri L. (2014) A Data-Driven Prediction Framework for Analysing 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

Bickford, M. (2005). Stress in the Workplace: A General Overview of the Causes, the Effects, and the Solutions. *Canadian Mental Health Association Newfoundland and Labrador Division, 44.*

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

Bowie, N.E. (2017) *Business ethics: A Kantian perspective*. Cambridge University Press.

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

Breuker, D., Matzner, M., Delfmann, P. and Becker, J. (2016) Comprehensible Predictive Models for Business Processes. Mis Quarterly, 40(4), pp.1009-1034

Burton, J.P., Hoobler, J.M. and Scheuer, M.L. (2012) Supervisor workplace stress and abusive supervision: The buffering effect of exercise. *Journal of Business and Psychology*, 27(3), pp.271-279.

Carroll, K.J. (2003) On the use and utility of the Weibull model in the analysis of survival data. Controlled clinical trials, 24(6), pp.682-701.

Cavanagh, G.F., Moberg, D.J. and Velasquez, M (1981) The ethics of organisationalpolitics. *Academy of Management Review*, 6(3): 363-374.

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

Colligan, T.W. and Higgins, E.M. (2006) Workplace stress: Etiology and consequences. *Journal of workplace behavioral health, 21(2),* pp.89-97.

De Graaf, M.M. and Malle, B.F, (2017) October. How people explain action (and autonomous intelligent systems should too). In 2017 *AAAI Fall Symposium Series*.
de Leoni, M. (Massimiliano); Mannhardt, Felix (2015): Road Traffic Fine Management Process. 4TU.ResearchData. Dataset. https://doi.org/10.4121/uuid:270fd440-1057-4fb9-89a9-b699b47990f5

Demšar, J. (2006) Statistical comparisons of classifiers over multiple data sets. Journal of Machine learning research, 7(Jan), pp.1-30.

Denisov, V., Fahland, D. and van der Aalst, W.M. (2019) Predictive performance monitoring of material handling systems using the performance spectrum. In *2019 International Conference on Process Mining (ICPM)* (pp. 137-144). IEEE.

Diamond, D.M., Campbell, A.M., Park, C.R., Halonen, J. and Zoladz, P.R. (2007) The temporal dynamics model of emotional memory processing: a synthesis on the neurobiological basis of stress-induced amnesia, flashbulb and traumatic memories, and the Yerkes-Dodson law. *Neural plasticity*, 2007.

Di Francescomarino, C., Dumas, M., Maggi, F. M., & Teinemaa, I. (2016) Clustering-based predictive process monitoring. *IEEE transactions on services computing*, *12*(6), 896-909.

Di Francescomarino, C., Ghidini, C., Maggi, F. M., & Milani, F. (2018) Predictive process monitoring methods: Which one suits me best?. In *International Conference on Business Process Management* (pp. 462-479). Springer, Cham.

Dirick, L., Claeskens, G. and Baesens, B. (2017) Time to default in credit scoring using survival analysis: a benchmark study. Journal of the Operational Research Society, 68(6), pp.652-665.

Dodd, D.H. & Bradshaw, J.M. (1980) Leading questions and memory: Pragmatic constraints. *Journal of Verbal Learning and Verbal Behavior*, 19(6):695-704.

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

Eder, J., Panagos, E., Pozewaunig, H. and Rabinovich, M. (1999). *Time management in workflow systems* (pp. 265-280).

Everett, M.G. and Borgatti, S.P. (1999) The centrality of groups and classes. The Journal of mathematical sociology, 23(3), pp.181-201.

Evermann, J., Rehse, J.R. and Fettke, P. (2017) Predicting process behaviour using deep learning. Decision Support Systems, 100, pp.129-140.

Eynon, G., Hills, N.T. & Stevens, K.T. (1997) Factors that influence the moral reasoning abilities of accountants: Implications for universities and the profession. *Journal of Business ethics*, 16(12-13): 1297-1309.

Ferrell, O.C. & Gresham, L.G. (1985) A Contingency Framework for Understanding Ethical Decision Making in Marketing. Journal of marketing 1985, 49, 87–96.

Financial Service Consumer Panel. (2017) *Algorithms in decision-making.* 2018, May 23. HC 351, 2017-2019, ALG0039.

Fiske, S.T. & Taylor, S.E. (1991) *Social cognition*. Mcgraw-Hill Book Company.

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

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

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

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

FutureLearn. (n.d.). Event logs used in this course. Available online: http://www.promtools.org/prom6/downloads/FutureLearn%20-%20Process%20mining%20with%20ProM%20-%20Event%20logs.zip (accessed on 10 Sept 2020)

Galanti, R., Coma-Puig, B., de Leoni, M., Carmona, J., & Navarin, N. (2020) Explainable predictive process monitoring. In *2020 2nd International Conference on Process Mining (ICPM)* (pp. 1-8). IEEE.

Geletkanycz, M.A. and Hambrick, D.C. (1997) The external ties of top executives: Implications for strategic choice and performance. Administrative science quarterly, pp.654-681.

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

Gilbert, D.T. & Malone, P.S. (1995) The correspondence bias. *Psychological Bulletin*, 117(1):21.

Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M. & Kagal, L. (2018) October. Explaining explanations: An overview of interpretability of machine learning. *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*: 80-89. IEEE.

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].

Grigori, D., Casati, F., Castellanos, M., Dayal, U., Sayal, M. and Shan, M.C. (2004) Business process intelligence. *Computers in industry*, *53*(3), pp.321-343.

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

Haidt, J. (2001) The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological Review*, 108(4):814.

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

Hansen, D., Smith, M.A. and Shneiderman, B. (2011) Eventgraphs: charting collections of conference connections. In *2011 44th Hawaii International Conference on System Sciences* (pp. 1-10). IEEE.

Harman, G.H. (1965) The inference to the best explanation. *The philosophical review*, 74(1) :88-95.

Hayes, B. & Shah, J.A. (2017) March. Improving robot controller transparency through autonomous policy explanation. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*:303-312. IEEE.

Hebb, D.O, (1955) Drives and the CNS (conceptual nervous system). *Psychological review*, 62(4), p.243.

Hegarty, W.H. & Sims, H.P. (1978) Some determinants of unethical decision behavior: An experiment. *Journal of Applied Psychology*, 63(4): 451.

Hengl, T., Nussbaum, M., Wright, M.N., Heuvelink, G.B. and Gräler, B. (2018) Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*, *6*, p.e5518.

Hilton, D.J. & Slugoski, B.R. (1986) Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological Review*, 93(1): 75.

Hilton, D.J. (1990) Conversational processes and causal explanation. *Psychological Bulletin*, 107(1): 65.

Hilton, D.J. (1996) Mental models and causal explanation: Judgements of probable cause and explanatory relevance. *Thinking & Reasoning*, 2(4): 273-308.

Hunsicker, M.E., Kappel, C.V., Selkoe, K.A., Halpern, B.S., Scarborough, C., Mease, L. and Amrhein, A. (2016) Characterizing driver–response relationships in marine pelagic ecosystems for improved ocean management. *Ecological applications*, 26(3), pp.651-663.

Hunt, S.D. & Vitell, S. (1986) A general theory of marketing ethics. *Journal of macromarketing*, 6(1):5-16.

Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J.C. and Rizzolatti, G. (2005) Grasping the intentions of others with one's own mirror neuron system. *PLoS Biol*, 3(3), p.e79.

IEEE (2016) IEEE Standards Association Introduces Global Initiative for Ethical Considerations in the Design of Autonomous Systems
https://standards.ieee.org/news/2016/ieee_autonomous_systems.html
[Accessed 25 March 2020]

Jaspars, J.M. & Hilton, D.J. (1988) Mental models of causal reasoning.

Jenness, S.M., Goodreau, S.M. and Morris, M. (2018) EpiModel: an R package for mathematical modeling of infectious disease over networks. *Journal of statistical software, 84*

Johnson, D. G. (2015) Technology with no human responsibility? *Journal of Business Ethics*, *127*(4): 707-715.

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

Jones, T.M. (1991) Ethical decision making by individuals in organisations: Anissue-contingent model. *Academy of management review*, 16(2):366-395.

Josephson, J.R. & Josephson, S.G. eds. (1996) *Abductive inference: Computation, philosophy, technology*. Cambridge University Press.

Jung, C.G. (1968) Lecture five: Analytical psychology: Its theory and practice (151-160).

Kelman, H.C. & Hamilton, V.L. (1989) *Crimes of obedience: Toward a social psychology of authority and responsibility*. Yale University Press.

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]

Klimov, P. (2017) *Algorithms in decision-making*. 2018, May 23. HC 351, 2017-2019, Q83

Kulesza, T., Stumpf, S., Burnett, M., Yang, S., Kwan, I. & Wong, W.K. (2013) September. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on Visual Languages and HumanCentric Computing*: 3-10. IEEE.

Larivière, B. and Van den Poel, D. (2004) Investigating the role of product featuresin preventing customer churn, by using survival analysis and choice modeling:

The case of financial services. Expert Systems with Applications, 27(2), pp.277-285.

Linden, A. and Yarnold, P.R. (2017) Modeling time-to-event (survival) data using classification tree analysis. Journal of evaluation in clinical practice, 23(6), pp.1299-1308.

Longley, P.A., Goodchild, M.F., Maguire, D.J. and Rhind, D.W. (2015) *Geographic information science and systems*. John Wiley & Sons.

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

Martin, K (2019). Ethical implications and accountability of algorithms. *Journal of Business Ethics*, *160*(4): 835-850.

Martin, K., Shilton, K. & Smith, J. (2019) Business and the Ethical Implications of Technology: Introduction to the Symposium: 1-11

Maybin, S. (2016) October 17. How maths can get you locked up. BBC. https://www.bbc.co.uk/news/magazine-37658374 [Accessed 2 April 2020]

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.

Messick, D.M. & Bazerman, M.H. (1996) Ethical leadership and the psychology of decision making. *MIT Sloan Management Review*, 37(2): 9.

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

Miller, D.T. and Gunasegaram, S. (1990). Temporal order and the perceived mutability of events: Implications for blame assignment. *Journal of personality and social psychology*, 59(6): 1111.

Miller, H.J. (2004) Tobler's first law and spatial analysis. *Annals of the Association of American Geographers*, *94*(2), pp.284-289.

Miller, S.M. (2018). AI: Augmentation, more so than automation. Singapore Management University. https://ink.library.smu.edu.sg/cgi/viewcontent.cgi?article=1081&context=ami [Accessed 12 April 2020]

Miller, T. (2019) Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267: 1-38.

Murgia, M. (2019) December 12. Algorithms drive online discrimination, academic warns. Financial Times. https://www.ft.com/content/bc959e8c-1b67-11ea-97df-cc63de1d73f4 [Accessed 10 March 2020]

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

Oxford Internet Institute (2017) *Algorithms in decision-making.* 2018, May 23. HC 351, 2017-2019, ALG0031.

Panagos, E. and Rabinovich, M. (1996) November. Escalations in workflow management systems. In *Proceedings of the workshop on Databases: active and real-time* (pp. 25-28). ACM

Paradice, D. B., & Dejoie, R. M. (1991) The ethical decision-making processes of information systems workers. *Journal of Business Ethics*, *10*(1): 1-21.

Pasquadibisceglie, V., Appice, A., Castellano, G. and Malerba, D. (2019) June. Using Convolutional Neural Networks for Predictive Process Analytics. In 2019 International Conference on Process Mining (ICPM) (pp. 129-136). IEEE.

Pasquadibisceglie, V., Castellano, G., Appice, A., & Malerba, D. (2021). FOX: a neuro-Fuzzy model for process Outcome prediction and eXplanation. In *2021 3rd International Conference on Process Mining (ICPM)* (pp. 112-119). IEEE.

Peachey, K. (2019) November 18. Sexist and biased? How credit firms make decisions. BBC. https://www.bbc.co.uk/news/business-50432634b [Accessed 15 March 2020]

Peirce, C.S. (1997) *Pragmatism as a principle and method of right thinking: The 1903 Harvard lectures on pragmatism*. SUNY Press.

Pines, A., & Aronson, E. (1988). *Career burnout: Causes and cures*. Free press.

Reijers, H.A. (2007). Case prediction in BPM systems: a research challenge. *Journal of Korean Institute of Industrial Engineers*, 33(1), pp.1-10.

Reik, T. (1948) Listening with the Third Ear: The Inner Experience of a Psychoanalyst. Farrar, Straus and Cudahy. Inc., NY.

Rest, J.R. (1986) Moral development: Advances in research and theory. New York: Praeger

Ribeiro, M.T., Singh, S. & Guestrin, C. (2016) "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* :1135-1144.

Richman, J.A., Zlatoper, K.W., Ehmke, J.L.Z. and Rospenda, K.M. (2006) Retirement and drinking outcomes: Lingering effects of workplace stress*?. Addictive Behaviors*, 31(5), pp.767-776.

Rizzi, W., Di Francescomarino, C., & Maggi, F. M. (2020). Explainability in predictive process monitoring: When understanding helps improving. In *International Conference on Business Process Management* (pp. 141-158). Springer, Cham.

Rizzolatti, G. (2005) The mirror neuron system and imitation. *Perspectives on imitation: From neuroscience to social science*, 1, pp.55-76.

Rogge-Solti, A. and Weske, M. (2013) December. Prediction of remaining service execution time using stochastic petri nets with arbitrary firing delays. In International Conference on Service-Oriented Computing (pp. 389-403). Springer, Berlin, Heidelberg.

Rozinat, A., Wynn, M.T., van der Aalst, W.M., ter Hofstede, A.H. and Fidge, C.J. (2009) 'Workflow simulation for operational decision support.' *Data & Knowledge Engineering*, 68(9), pp.834-850.

Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., Er, M.J., Ding, W. & Lin, C. T. (2017). A review of clustering techniques and developments. *Neurocomputing*, *267*, 664-681.

Schellekens,B. (2009) *Cycle time prediction in Staffware*.Masters Thesis, Eindhoven University of Technology, Eindhoven.

Schweitzer, M.E., Ordóñez, L. & Douma, B (2004) Goal setting as a motivator of unethical behavior. *Academy of Management Journal*, 47(3): 422-432.

Science and Technology Committee (2018) *Algorithms in decision-making*. 2018, May 23. HC 351, 2017-2019.

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

Senderovich A., Beck, C., Gal, A., and Weidlich, M. (2019) *"Congestion Graphs for Automated Time Predictions"* in *. Proceedings of the* 33rd AAAI Conference on Artificial Intelligence (AAAI), 2019

Smith-Crowe, K. (2004) *An interactionist perspective on ethical decision-making: Integrative complexity and the case of worker safety* (Doctoral dissertation, Tulane University).

Somers, M.J. (1996) Modelling employee withdrawal behaviour over time: A study of turnover using survival analysis. Journal of Occupational and Organizational Psychology, 69(4), pp.315-326.

Sonenshein, S (2007) The role of construction, intuition, and justification in responding to ethical issues at work: The sensemaking-intuition model. *Academy of Management Review*, *32*(4):1022-1040.

Song, M. and Van der Aalst, W.M. (2008) Towards comprehensive support for organisational mining. Decision support systems, 46(1), pp.300-317.

Strudler, A. & Warren, D. (2001) Authority, heuristics, and the structure of excuses. In *Next Phase of Business Ethics: Integrating Psychology and Ethics*: 355-375. JAI Press.

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

Tax, N., Verenich, I., La Rosa, M. and Dumas, M. (2017) June. Predictive business process monitoring with LSTM neural networks. In International Conference on Advanced Information Systems Engineering (pp. 477-492). Springer, Cham.

Taylor, S.E (1975) On inferring one's attitudes from one's behavior: Some delimiting conditions. *Journal of Personality and Social Psychology*, 31(1):126.

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]

Tenbrunsel, A.E. & Smith-Crowe, K. (2008) 13 ethical decision making: Where we've been and where we're going. *The Academy of Management Annals*, 2(1): 545-607.

Tenbrunsel, A.E. (1998) Misrepresentation and expectations of misrepresentation in an ethical dilemma: The role of incentives and temptation. *Academy of Management Journal*, 41(3): 330-339.

Tenbrunsel, A.E., Smith-Crowe, K. & Umphress, E.E (2003) Building houses on rocks: The role of the ethical infrastructure in organisations*. Social justice research*, 16(3): 285-307.

Tetrick, L.E. and Winslow, C.J. (2015). Workplace stress management interventions and health promotion. *Annual Review of Organizational Psychology and Organizational Behavior*, 2(1), pp.583-603.

Tobler, W.R. (1970) A computer movie simulating urban growth in the Detroit region. *Economic geography*, *46*(sup1), pp.234-240.

Treviño, L. K., Weaver, G. R., & Reynolds, S. J. (2006) Behavioral ethics in organizations: A review. *Journal of management*, *32*(6): 951-990.

Treviño, L.K. (1986) Ethical decision making in organisations: A person-situation interactionist model. *Academy of Management Review*, 11(3): 601-617.

Tversky, A. & Kahneman, D (1983) Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90(4): 293.

Van Bouwel, J. & Weber, E (2002) Remote causes, bad explanations? *Journal for the Theory of Social Behaviour*, 32(4): 437-449.

van der Aalst, W.M. (2016) Process Mining: Data Science in Action. 2nd edition. Springer Berlin Heidelberg.

van der Aalst, W.M. (2013). Business process management: a comprehensive survey. *International Scholarly Research Notices*, *2013*.

van der Aalst, W.M. (2010) June. Business process simulation revisited. In *Workshop on Enterprise and Organizational Modeling and Simulation* (pp. 1-14). Springer, Berlin, Heidelberg.

van der Aalst, W.M., Nakatumba, J., Rozinat, A. and Russell, N., (2008). Business Process Simulation: How to get it right. *BPM Center Report BPM-08-07, BPMcenter. org*, 285, pp.286-291.

van der Aalst, W.M., Reijers, H.A. and Song, M., (2005) Discovering social networks from event logs. Computer Supported Cooperative Work (CSCW), 14(6), pp.549-593.

van der Aalst, W.M., Schonenberg, M.H. and Song, M. (2011)' Time prediction based on process mining.' *Information Systems*, 36(2), pp.450-475.

van der Aalst, W.M., Reijers, H.A., Weijters, A.J., van Dongen, B.F., De Medeiros, A.A., Song, M. and Verbeek, H.M.W. (2007) Business process mining: An industrial application. *Information Systems*, *32*(5), pp.713-732.

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 van Dongen, B.F. (Boudewijn) (2012) BPI Challenge 2012. 4TU. Centre for Research Data. Dataset. https://doi.org/10.4121/uuid:3926db30-f712-4394-aebc-75976070e91f

van Dongen, B.F. (Boudewijn) (2014) BPI Challenge 2014. 4TU.Centre for Research Data. Dataset. https://doi.org/10.4121/uuid:c3e5d162-0cfd-4bb0-bd82-af5268819c35

van Dongen, B.F. (Boudewijn) (2015) BPI Challenge 2015 Municipality 3. Eindhoven University of Technology. Dataset. https://doi.org/10.4121/uuid:ed445cdd-27d5-4d77-a1f7-59fe7360cfbe

van Dongen, B.F. (Boudewijn) (2017) BPI Challenge 2017. Eindhoven University of Technology. Dataset. https://doi.org/10.4121/uuid:5f3067df-f10b-45da-b98b-86ae4c7a310b

van Dongen, B.F. (Boudewijn) (2017) BPI Challenge 2017. Eindhoven University of Technology. Dataset. https://doi.org/10.4121/uuid:5f3067df-f10b-45da-b98b-86ae4c7a310b

van Dongen, B.F. (Boudewijn); Borchert, F. (Florian) (2018) BPI Challenge 2018. Eindhoven University of Technology. Dataset. https://doi.org/10.4121/uuid:3301445f-95e8-4ff0-98a4-901f1f204972

van Dongen, Boudewijn (2019): BPI Challenge 2019. 4TU.ResearchData. Dataset. https://doi.org/10.4121/uuid:d06aff4b-79f0-45e6-8ec8-e19730c248f1

van Dongen, B.F. (Boudewijn) (2020) BPI Challenge 2020. 4TU.Centre for Research Data. Dataset. https://doi.org/10.4121/uuid:52fb97d4-4588-43c9-9d04-3604d4613b51

Veldhoen, J. (2011) *The Applicability of Short-term Simulation of Business Processes for the Support of Operational Decisions*, Masters Thesis, Technische Universiteit Eindhoven, Available at**:** http://alexandria.tue.nl/extra2/afstversl/tm/Veldhoen%202011.pdf

Verenich, I., Dumas, M., Rosa, M.L., Maggi, F.M. and Teinemaa, I. (2019) Survey and cross-benchmark comparison of remaining time prediction methods in business process monitoring. ACM Transactions on Intelligent Systems and Technology (TIST), 10(4), pp.1-34.

Verenich, I., Nguyen, H., La Rosa, M., & Dumas, M. (2017) White-box prediction of process performance indicators via flow analysis. In *Proceedings of the 2017 International Conference on Software and System Process Pages*, ACM, Paris, France, pp. 85-94.

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

Weick, K.E. (1993) The Collapse of Sensemaking: The Mann Gulch Disaster. Administrative Science Quarterly 1993, 38.

Weller, P (2017) *Algorithms in decision-making*. 2018, May 23. HC 351, 2017-2019, Q30

Williams, S., Michie, S., & Pattani, S. (1998). *Improving the health of the NHS workforce: Report of the partnership on the health of the NHS workforce*. Nuffield Trust.

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

Yokuma, J.T. and Armstrong, J.S., 1995. Beyond accuracy: Comparison of criteria used to select forecasting methods. *International Journal of Forecasting*, 11(4), pp.591-597.

1