University of Twente

Faculty Behavioural, Management and Social Sciences

Master of Science in Business Administration

Track Digital Business

Master Thesis

28/6/2021

Comparison between the predefined process and the actual process model

Digdem Ozturk – s2271400

Fist supervisor: Dr. M. De Visser

Second supervisor: Dr. M.L. Ehrenhard

Company supervisor: T. Genc

Table of Content

Acknowledgement

First of all, I want to express my special thanks of gratitude to my professor, dr. M. de Visser, for this valuable guidance. Secondly, would like to thank my second supervisor dr. M.L. Ehrenhard. Lastly, my company supervisor, T. Genc, allowed me to do this wonderful project on the topic process mining, which helped me learn a lot in this research field.

It would have been impossible to obtain this degree without my family's and friends unconditional love and support: my parents Kenan and Bilgin, my sister Betul, my friend Alice and my fiancé Enes, and all the other members of the family. You always backed me up along the course of my life and studies.

1

Introduction

Technical innovation and automatisation changed the accounting practices. With the integration of robotic process automation (RPA), the tasks that need to be done by humans have reduced (Knudsen, 2020). Considering the fast technological innovations that lead to the adaptation of current work practices, today's organisations constantly need to improve their process. Low-value work is standardised and replaced by software applications (Bhimani & Willcocks, 2014). For example, accounting software applications provide real-time and smart processing; scanning and matching information; and, approving and booking documents. While executing the tasks, accounting software applications generate data in the form of logs. For many organisations the challenge is to extract information from data stored in software applications.

This challenge is also the case for a global accounting, audit, and consulting firm. The company mentioned that there is a need to uncover the difference between the handmade process model of the invoice booking process and the event log generated from the accounting software application. Another challenge for them is to explore the potential of process mining, but they do not know how to approach this.

The invoice booking process starts at the moment a customer uploads an invoice in the software system. Accountants use the software system for digital invoice and document processing. The software automatically recognises the information on the submitted invoices and makes a booking proposal. Afterwards, the employer checks the booking proposal and books the invoice when the proposed information is correct. While doing this, the software system leaves digital traces at every activity which provides detailed information about the individual task. Examples of the detailed information about the individual task (also called the event or activity) might include - the way in which the invoice entered the accounting system (for example by mail, scan or app), the time and date of the entry of the invoice, and the person who checked the booking proposal. *Figure 1* illustrates the activities of invoice processing.

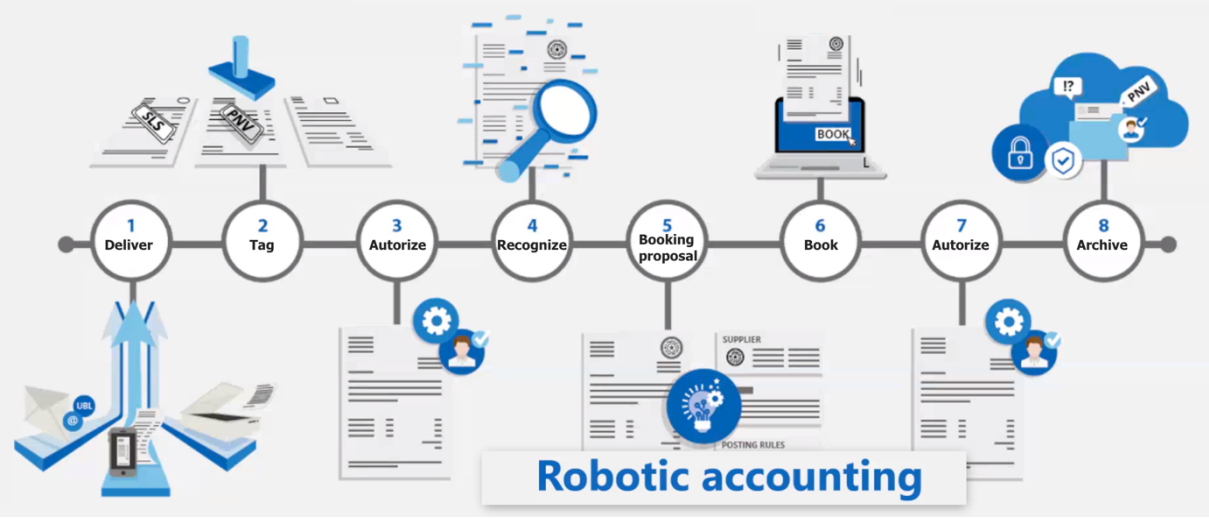


Figure 1 Steps of invoice processing

The activities mentioned in *figure 1* are expected to be executed for every invoice. But in reality, the firm experienced that the eight activities mentioned in *figure 1* do not always appear chronologically. They realised that some actions are skipped during the invoice processing. They also believe that some activities are not mentioned in this figure. This might lead to missing opportunity to be more efficient. The company believes that there can be more efficiency in the process by identifying the process variants that do not conform to the predefined process model. Therefore, the accounting firm would like to have fact-based insights about the alignment and deviation between the activities of the expected behaviour and the actual observed behaviour of the invoice processing.

Understanding the deviated paths of invoice booking processes is interesting for the accounting firm because they might know which customers do not comply to the predefined business model. This is interesting for the firm because they would be able to categorise the customers into "efficient" and "inefficient" customers. When an overview is created the company can pay special attention to the "inefficient" customers. The company might introduce initiatives that help the customers to improve their efficiency. The focus of this paper will not be on the cause of deviant behaviour and the initiatives that increase efficiency, but on the alignments and deviations between the reality and modelled invoice process.

Based on this focus, the master thesis seeks to answer the following research question: "*What is the difference between the predefined process and the actual process model?".* Understanding the invoice processing path might help the accounting firm to check the quality of the predefined model. With the results of this project, they would be able to analyse if the predefined model is accurate and correctly describing the reality.

This paper has been divided into four parts. The first section of this paper will provide a literature review. The chapter begins by laying out the theoretical dimensions of the research. The second section deals with the methodology used for this study, including data collection and data preparation. The third section presents the actual finding of the research. Finally, the paper will be concluded, the limitation of this study and suggestions for future research are proposed.

1.1 Management problem

The management notices that different customers use different billing processes. The accounting firm has included the different billing processes in their invoicing process, making it unclear which process flow variants exist and which one is the most efficient. For example, the firm experienced that some activities do not always appear chronologically in some process flows. Additionally, that some flows skip activities during the invoice processing. They also believe that there are some activities which they do not know exist.

This might lead to missing opportunity to be more efficient. The company believes that there can be more efficiency in the process by identifying the process variants that do not conform to the predefined process model. Therefore, the accounting firm would like to have fact-based insights about the alignment and deviation between the activities of the expected behaviour and the actual observed behaviour of the invoice processing.

1.2 Research design

This paper focuses on process mining. Through process mining businesses can verify whether firms follow the predefined business process and identify inefficiencies and effort drivers (Reinkemeyer, 2020). Process mining techniques provide fact-based insights and support process improvements (van der Aalst, 2016). The techniques analyse event logs from activities and provide insights into existing processes and complexities. Process mining techniques offer a thorough investigation and enable understanding how the process is being executed and provide the possibility to understand the level of resources and individual tasks (Dumas, La Rosa, Mendling, & Reijers, 2018).

According to van der Aalst (2016), process mining aims to analyse event data to give in-depth knowledge of the execution of the process in reality. It seeks to fill the gap between event data and process models. Process mining is frequently mixed with machine learning and data mining techniques to discover the root causes of deviations and inefficiencies. The observed behaviour (recognised from events) and modelled behaviour (recognised from process diagrams) are used to detect compliance and performance problems. Many process mining techniques can be applied to create insights.

This research aims to identify the differences between the predefined invoice booking process and the actual observed behaviour of the invoice booking process by using a process mining technique. The paper starts with a qualitative research to understand concepts and the process mining principle by examining various scholars' insights within the process mining domain. Combining process mining and qualitative research will give a theoretical background that will help this paper to find the right process mining technique to approach the research question. It concludes which process mining technique is the most applicable to answer the research question.

To answer this research question, the path of invoice processing must be described. Based on this description a model is designed. The designed model describes which actions must be performed until an invoice is booked. From now on, the designed model will be named the predefined process model. The predefined process model will be compared to the data generated from the accounting system.

This research will apply a quantitative research method because it will analyse data generated from the accounting system (ERP system). The quantitative data regarding invoice processing represent primary data which is requested by the database owner. The extracted excel file from the database is raw data. The raw is not suitable for process mining analysis therefore data cleaning and data integration has been applied to make it usable for processing mining analysis. The data is formalised into meaningful event logs in order to apply process mining analysis.

After the decision is made on which process mining technique is the most applicable to answer the research question, the predefined invoice booking process will be compared to the observed behaviour retrieved from the invoice booking process logs. To identify the difference between the predefined process model and the actual process model.

This thesis aims to solve a practical problem related to a new phenomenon. Process mining shows the alignment and deviation between the expected behaviour and the actual observed behaviour of the invoice processing. This information might be used for inspection and control purpose. When this project succeeds, the company will acquire knowledge of process mining to apply for other internal processes within the same context.

Some research has been carried out on process mining, but no single study has investigated logs from accounting systems to the best of my knowledge. This thesis's scope focuses on the real-life event logs received from an accounting system database with an event log's minimum requirements. This master thesis's learning objective is to analyse the research problem correctly and use scientific sources to create in-depth knowledge and understanding of the topic process mining to create a vivid research framework.

2

Literature review

In this chapter we introduced most of the basic definitions, concepts, and notation used throughout the rest of the chapters of this thesis. The theoretical base is set by examining various scholars' insights. This is done to create prior research knowledge, which will help this paper to find the right technique to approach the research question. Appendix 1 describes the literature research method used for this chapter.

2.1 Process model

In this part, an explanation of process models will be given because it is essential to understand the business process to carry out process analysis. Process models explain how things are executed and in what order they are being executed.

According to Carmona et al. (2018), a process is a collection of activities performed in a coordinated manner to achieve a specific goal. According to Kidler (2009), a business process requires a set of tasks performed in some administration or enterprise according to some rules to achieve specific goals. Process models describe how work is performed and map process properties into a model. Processes are modelled to understand the process and to discover and prevent issues (Dumas et al., 2018). Additionally, process models are used for documentation, animation, discussion, insights, verification, performance analysis, specification and configuration (van der Aalst, 2011)

Process models consist of tasks and each task represents an activity of the process and the execution dependencies of the process in a conceptual model. A process model aims to create an overarching view of the process and generalises individual cases (Carmona et al., 2018).

A process model might represent descriptive or prescriptive behaviour. The descriptive model behaviour shows reality, whereas the prescriptive model behaviour defines how reality should be. For both process model behaviours, it is crucial to relate the modelled action against the recorded behaviour to acquire insights on the model's capability to describe what is observed in the information system supporting a process (Munoz-Gama, 2016).

To avoid unnecessary details and remain the core of the process, some process properties might be ignored. This means that there might be a loss of information and can cause uncertainty in the relation between the process itself and the process model (Carmona et al., 2018). However, nowadays, footprints are left during the execution of a process and systems record these footprints. The recorded behaviour of a process, also called event logs, is an essential source of information that enables data-driven analysis (Carmona et al., 2018).

The next part will explain two important modelling languages, the Petri net and the BPMN, because the predefined process model and the actual observed behaviour will be explained using a modelling language.

2.1.1 Petri net

Petri net is a modelling language that can be used for modelling a business process. The Petri net explains the behaviour of systems recorded in the log (van der Aalst, Weijters & Maruster, 2004). Carl Adam Petri was the developer of Petri nets in 1962. Up until now, the Petri net has gone through many improvements and transactions. This modelling language is most studies and there are many publications on Petri net. It is applied in various computer science areas and other disciplines (Murata, 1989). The Petri net is a good model to use when studying distributed and concurrent systems.

It is essential to understand the business process to carry out process analysis, redesign, and automation. An example of deriving a process model from event data is the α-algorithm. The Alpha algorithm is the first algorithm capable of learning concurrent process models from event data while still providing formal guarantees (Knudsen, 2020). The α-algorithm takes the event log to create a Petri net by identifying process patterns in an event log.

Mans, Schonenberg, Song, van der Aalst and Bakker (2008) mention that a Petri net's structure consists of four parts. They categorise this as three static parts: arcs, places and transitions and the fourth part pass through other parts. The arcs in a process flow are represented as arrows that connect the places with transactions. The places are represented as circles, and these circles may have tokens (black dots) that passes to different places while executing an action. The transitions are represented as a box that indicates an action that is performed (Figure 2).

A Petri net process flow is characterised as process flows with arcs that go from places to transitions and another way around, other paths are excluded. The enabled transition can destroy one token by each incoming arc and produce one token by each outgoing arcs. Moreover, Petri net process flows have one starting place and one end (Leemans, Fahland & van der Aalst, 2013). The downside of Petri net process flow that it cannot differentiate short loops from true parallelism. Additionally, the Petri net cannot deal with noise (infrequent behaviour) and incompleteness in an event log (Dumas et al., 2018).

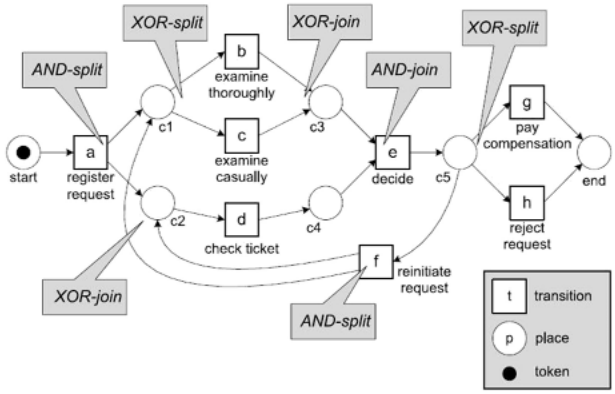


Figure 2 Apromore importer-exporter literature 1.0

2.1.2 BPMN

Another example of deriving a process model from event data is Business Process Model and Notation (BPMN). This is the most widely used modelling language to model business processes. It is developed under the coordination of and standardised by the Object Management Group (OMG). The OMG state the goal of the modelling language as follow: "The primary goal of BPMN is to provide a notation that is readily understandable by all business users, from the business analysts that create the initial drafts of the processes, to the technical developers responsible for implementing the technology that will perform those processes, and finally, to the business people who will manage and monitor those pro- cesses. Thus, BPMN creates a standardised bridge for the gap between the business process design and process implementation." (Object Management Group [OMG], 2014). Many tool vendors support this modelling language.

Appendix 2 shows the notational elements ("bpmn", 2020). The basic concepts of BPMN are events, activities, and arcs (Dumas et al., 2018). Those are represented in circles, rounded rectangles, and arrows. Events describe the things that happen immediately, for example, "receive an invoice", and are indicated by circles. Activities represent units of work having a duration, for example, "an invoice has been paid", and is characterised by rounded rectangles. Furthermore, arcs, also called sequence flows, are indicated by arrows with a full arrowhead (Dumas et al., 2018).

Activities and events can also be performed in illogical sequence, for example when two or more activities are alternative to one another. This situation is called mutually exclusive. Two activities that are independent of each other and are performed simultaneously (no sequence) can be performed in parallel. An activity is concurrent when two or more activities are interdependent (Dijkman, Dumas & Ouyang, 2008).

Gateways in a process model can be interpreted as a "door" that either able or disable to pass a gateway. In the BPMN language, this is represented as a diamond shape. A gateway can be categorised into the split and the join gateway. A split gateway illustrates the point where the process flow becomes different or follow a different direction. The split gateway has one incoming sequence flow and several outgoing sequences flow. This is referred to as an exclusive (XOR) split and represented as an "X" within the diamond shape in BPMN. The XOR-split gateway has mutually exclusive conditions, meaning that only one branch can be "true" or "chosen". *Figure 3* demonstrates an example of an XOR-split gateway (Figure 3). The model starts with a decision activity (decision activity is representing an activity that results in different outcomes), namely "Check documents tags for mismatches" following a start event with three possible outcomes. The outcomes are mutually exclusive and only one outgoing branch could be chosen every time. In the example, a document tag can be correct or false but can be corrected or false and cannot be corrected. Only one condition can be true per tagged document.

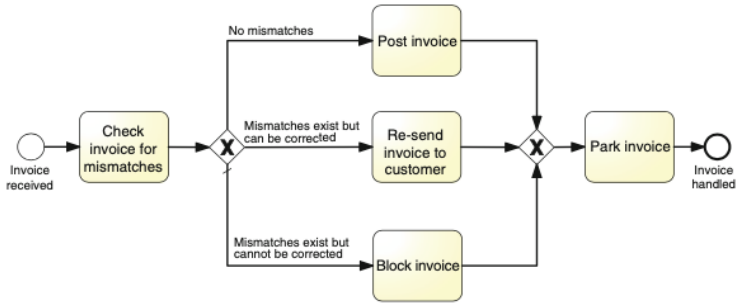


Figure 3 Example of XOR gateway (Carmona et al., 2018)

The join gateway illustrates the point where the process flow moves towards the same point where they join or meet. The join gateway has multiple incoming sequence flows and one outgoing sequence flow. When two or more activities do not need to follow or exclude the other, they can be performed parallelly or concurrently. In BPMN, this is referred as a parallel (AND) gateway and represented as an "+" within the diamond shape (Figure 4).

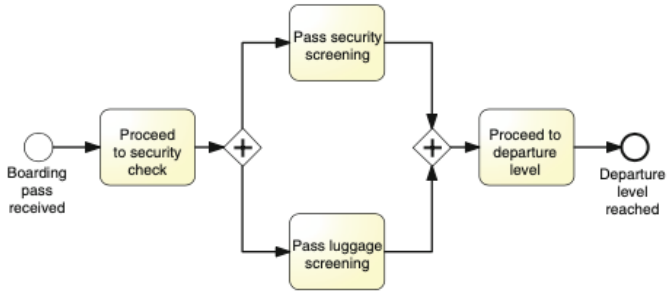


Figure 4 Example of AND gateway (Carmona et al., 2018)

However, two cases can omit a gateway. Firstly, omit XOR-join before an event or activity. The incoming arcs are straight connected to the event or activity. Secondly, omit AND-spit when it follows an event or activity. The outgoing arcs are straight connected to the event or activity.

There are models where one or more branches are needed after a decision activity depending on which conditions are true. One model is the inclusive (OR) split gateway. This model refers to a situation where a decision can lead to one or more options simultaneously. Or-split gateway is much like the XOR split however, the outgoing branches conditions do not need to be mutually exclusive. According

to Dumas (2018), the OR- join slit gateway is complex and can lead to more confusion for the reader therefore, he suggests using this model only if it is strictly required.

Additionally, some models repeat one or several activities for example because they failed to check activity. A model rework or repletion uses an XOR-join gateway to reconnect to the point before the repetition block (Figure 5).

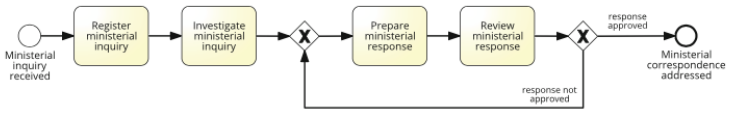


Figure 5 Example of the repeated process (Carmona et al., 2018)

Furthermore, the desired feature of process models is that they are block-structured. Block structure is a model fragment with a single- entry and single- exit. The entry and exit points are two gateways (one split and one join) and every route from one gateway directs to the other gateway. Dumas et al. (2018) state that block-structured process models are easier to understand than unstructured ones. The unstructured models have one entry and two or more exit points.

2.2 Event logs

In this part, an explanation of event logs will be given because process mining of the event log is a technique that analyses business processes using the information in the event log (van der Aalst, 2011). Event logs are used to enrich and learn the process model. Through repeating history and using the model, it is possible to determine the exact relationship between event and model elements. This relationship could be used to analyse performance and check conformance (van der Aalst et al., 2012).

According to Dumas et al. (2018), an event log is a group of timestamped events. Event log reports the performance of a task (activity) from the process, the event's message, and other valuable information within a business process's context.

Additionally, Jans, Alles, and Vasarhelyi (2013) write that an event log is a set of digital traces that automatically and chronologically register the system's actions. In the corporate environment, these are stored for each business activity in databases, information systems, and enterprise systems, such as Customer Relationship Management (CRM), Enterprise Resource Planning (ERP), Supplier Relationship Management (SRM), and many more systems. Each task (or action) is processed in a database where it leaves digital traces. To "mine" a process, it needs transparency in the order of activities that have taken place. This can be acquired by identifying the digital traces, extracting the digital traces, and visualising the digital traces to demonstrate the actual process flow. The sequence and the processing time of the events provides an overview of each case. In this way, process flows can be traced and enable to understand delays, detect complexity drivers, and separate loops (Reinkemeyer, 2020).

In brief, event logs refer to activities performed by resources at a specific time and for a particular case. The majority of databases of enterprise systems, information systems, business process management systems, and other sources register events based on a task's execution. These event data provide the possibility to analyse what happened, when it happened, and how many times in what context it happened. The extracted event records can be represented as an event log.

2.2.1 Log format

To analyse information from event logs, three necessary attributes must be extracted: (1) the case identifier, (2) activity, and (3) timestamp. Table 1 represents the minimum required attributes of an event log for process mining (Dumas et al., 2018). In practice, there might be other event attributes. Further attributes can provide additional information over specific activities, such as the resource and variant. The resource tells who performed a particular task, also known as the action owner. The variant is a single path followed by one or more case identifiers with identical routings. For instance, if the case identifier one and two have both the same routing "Tag document – Insert supplier – Create invoice –

Make booking– Archive invoice", then the case identifier one and two will be grouped into one variant (Dumas et al., 2018).

TABLE 1
Minimum requirement of an event log

Content		Description
Case identifier		• The case identifier tells in which case the event occurred • Example: Case ID or invoice number
Activity		• The activity provides a specification of what kind of activity was performed • Example: "book" or "tag" by invoice processing
Timestamp		• The timestamp points out when the event occurred • Example: time, day, month, and year of the event (08:00:38 20-10-2020)

Table 2 shows a detailed example of an event log of an invoice processing system. As shown, several events are corresponding to actions performed by the system, for example, classification of the document ("Tag document"), supplier insertions ("Insert supplier"), invoices ("Create invoice"), bookings ("Make booking"), and archived documents ("Archive"). Additionally, each event has a timestamp.

TABLE 2
Example of an event log for invoice processing

Case identifier	Variant	Activity	Resource	Timestamp
20208001	1	Tag document	Nick@mz.staff.com	07/08/2020 09:13:00
20208001	1	Insert supplier	Hans@mz.staff.com	12/08/2020 09:30:32
20208001	1	Create invoice	Hans@mz.staff.com	14/08/2020 09:31:07
20208001	1	Make booking	Hans@mz.staff.com	14/08/2020 14:28:55
20208001	1	Archive	SYS	14/08/2020 14:50:18
20208002	1	Tag document	Nick@mz.staff.com	07/08/2020 09:15:01
20208002	1	Insert supplier	Sara@mz.staff.com	11/08/2020 09:55:52
20208002	1	Create invoice	Sara@mz.staff.com	11/08/2020 10:20:36
20208002	1	Make booking	Bob@mz.staff.com	17/08/2020 10:50:27
20208002	1	Archive	SYS	17/08/2020 10:51:11
20208003	1	Tag document	Lisa@mz.staff.com	13/08/2020 09:05:37
20208003	1	Insert supplier	Lisa@mz.staff.com	13/08/2020 09:08:06
20208003	1	Create invoice	Hans@mz.staff.com	14/08/2020 09:30:58
20208003	1	Make booking	Bob@mz.staff.com	17/08/2020 11:00:00
20208003	1	Archive	SYS	17/08/2020 11:01:48

Table 2 also shows an example of a variant and three matching process paths. As mentioned before, all process paths that have the same routine will be grouped into one variant. In contrast, processes with different routings paths will be grouped into different variants. The standard and non-standard routings can be identified by studying the variants. When the group of process instances confirms the firm's standard business processes, it is called a standard variant. When the process instances include deviated paths from the standard business process, it is called a non-standard variant (Chiu & Jans, 2019). For example, a standard invoice processing is "Tag document – Insert supplier – Create invoice – Make booking – Archive", yet a non-standard invoice processing could be ""Tag document –Create invoice – Make booking – Archive". Among these two variants there is a missing activity, namely "Insert supplier" this can cause potential risks.

With the understanding of standard variants and non-standard variants, organisations can detect the most common paths, types of deviations, undesirable performance variation, inefficient processes, and potential risks.

2.3 Process mining – an overview

In the previous parts, two essentials perspectives have been explained. First, the process model is presented as the conceptual description of the underlying process. Second, event logs are introduced as footprints recorded by information systems during process execution. This part will introduce the three main types of process mining. Combining process mining and qualitative research will give a theoretical background that will help this paper find the right process mining technique to approach the research question.

Process mining has its roots in the business process management (BPM) discipline (Dumas et al., 2018; van der Aalst, 2013). In the 1990s, Cook et al. were one of the first who measured the relationship between process model and event logs. They compared the t-streams created from the event log with the model's event streams (Cook & Wolf, 1999; Cook, He, & Ma, 2001). Over the last two decades, many process mining techniques have been proposed and refined over time. The first process mining techniques could not handle infrequent behaviours and made strong assumptions about the event logs' completeness (van der Aalst, 2018). These approaches contained fuzzy mining, heuristic mining, and diverse generic approaches (van der Aalst, 2018; Weijters & van der Aalst, 2003). Since 2010, there is an increase in new process mining techniques. More recent attention is on event logs with noise and infrequent behaviours. Besides, event data have become readily available for analysis, and process mining techniques have matured.

Dumas et al. (2018) use the term process mining to refer to a broad collection of techniques to obtain insights from event logs created while executing the business process. Also, Dumas et al. (2018) found that some of the process mining techniques focus on discovering a process model, and others on analysing the process. Additionally, according to van der Aalst, van Dongen, Herbst, Maruster, Schimm, and Weijters (2003), the term process mining refers to "methods for distilling a structured process description from a set of real executions." (p.241). Van der Aalst (2011) mentions that there are two main drivers for process mining. The first driver is the increase in recorded event data that provides in-depth information about the history of processes. Despite the presentence of event data, many organisations identify problems based on fiction rather than facts. The second driver is the promising miracle made by vendors of Business Intelligence (BI) and Business Process Management (BPM), despite that they did not reach the expectations of consultants, software vendors, and academics.

Process mining aims to analyse business processes' event data to understand the processes' execution and behaviours, for example, undesirable performance variation, most common paths, bottlenecks, deviations, and the frequent resource of a defect. As mentioned in the definition of Dumas et al. (2018) process mining spectrum is broad, and many process mining techniques can be used, deciding which technique to use depends on the desired insight and answer. Examples of process mining techniques are process discovery, conformance checking, compliance checking, process enhancement, process prediction, process monitoring, and operational support.

According to van der Aalst (2016), process mining aims to analyse event data to give in-depth knowledge of the execution of the process in reality. It seeks to fill the gap between event data and process models. Process mining is frequently mixed with machine learning and data mining techniques to discover the root causes of deviations and inefficiencies. The observed behaviour (recognised from events) and modelled behaviour (recognised from process diagrams) are used to detect compliance and performance problems. Van der Aalst (2016) mentions that process mining techniques can be categorised based on their relation to three core tasks: process discovery, conformance checking, and process enhancement (Figure 6).

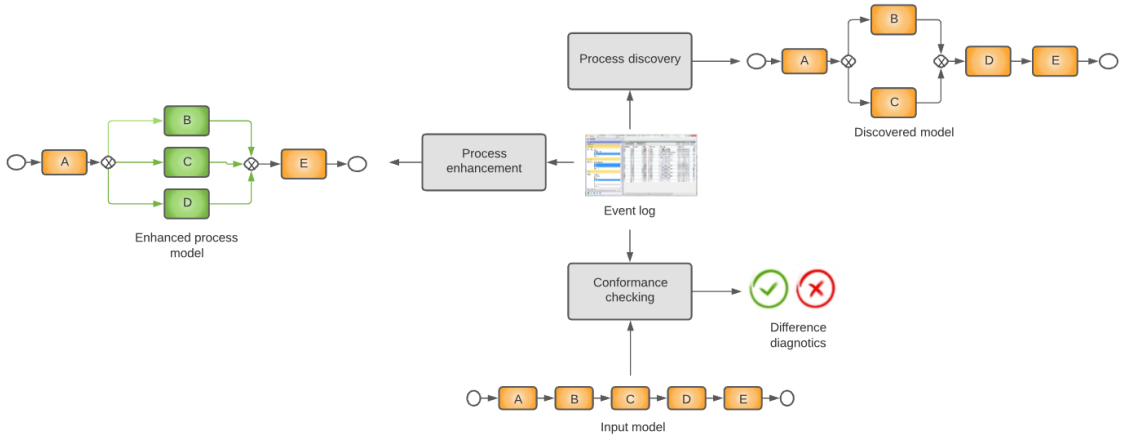


Figure 6 Categories of process mining techniques

2.3.1 Process discovery

The process discovery technique takes an event log and produces a process without using any additional and prior information. Over time, it became clear that the process discovery techniques are the starting point to process improvements and other types of analysis. An example of deriving a process model from event data is the α-algorithm. This algorithm takes the event log to create a Petri net by identifying process patterns in an event log. The Petri net explains the behaviour of systems recorded in the log (van der Aalst, Weijters & Maruster, 2004). The firm's business process is captured in a process model with the process discovery technique. The business process map visualises a model that shows the process flow of a firm's occurring activities (Chiu & Jans, 2019).

Even though the alpha algorithm still provides formal guarantees, it cannot differentiate short loops from true parallelism. Additionally, the α-algorithm cannot deal with noise (infrequent behaviour) and incompleteness in an event log (Dumas et al., 2018). For instance, event logs might have cases where the head is missing, or with a missing intermediate, or missing tail because the events in the cases are not recorded. For example, an employee might have forgotten to tag a document as "invoice" and therefore the document cannot be processed and is missing the corresponding trace. Also, there might be cases where events are recorded in an incorrect order or twice. These noises should not bias the process model created by a process discovery technique.

More algorithms use different representations such as heuristics miner, inductive miner and split miner. Contrary to the α-algorithm, the heuristics miner can handle noisy and incomplete event logs and discover self-loops and short- loops. Although the heuristics miner is applied to a large real-life event log, it often produces process models that are too large, behaviourally incorrect, and spaghetti-like. The experiments reported by Augusto et al. (2017) suggest that the split miner and inductive miner are among the most robust algorithms for automated process discovery. However, some tuning is required and the relative performance on a given log can change. Process discovery techniques are essential however, the attention is switching to the steps after process discovery using machine learning, optimisation, and simulation.

2.3.2 Conformance checking

With conformance checking, a process model is compared with the logs of the same process model. With this technique, the event log and a process model's commonalities and differences can be detected and diagnosed (van der Aalst et al., 2012). Both representations, the event log and model, mention the same thing, the real process. However, creating a relationship between them is essential for understanding how the process is executed, and how far apart the described model is from the recorded reality. According to Carmona et al. (2018), conformance checking refers to the analysis of "the relation between the intended behaviour of a process as described in a process model and event logs that have been recorded during the execution of the process." (p. 3). Conformance checking techniques are applied to compute the event log and model relation automatically.

The input of conformance checking techniques is event logs and process models. The output of conformance checking techniques is a list of differences between the process model and the event log. The confrontation between process model (discovers automatically or handmade) and event data (recorded behaviour) might touch on interesting and relevant questions. For instance, a process describes that after executing task A, another task must be completed called B. However, from the logs, it is recognised that sometimes after task A, task B is not performed. This might be due to an exception that had happened, which is not recorded in the process model or an error. Conformance checking techniques might also take other inputs, such as a set of business rules and event logs. In this way, organisations can detect if the log fulfils the business rules and possible cases that violate laws (Dumas et al., 2018).

Detecting the deviations between actual process models and predefined process models could be essential because it explains the validity of a process model and warns for unusable behaviours in a case (Leemans, van der Aalst, Brockhoff, & Polyvyanyy, 2021). For example, deviations can stress the quality of systems that control the process or emphasise the quality of the process's progress. Additionally, it can also highlight that the model is inaccurate or outdated because of new evolutions or pathways that are not incorporated into the process model and is therefore not correctly describing the reality. An organisation can continuously develop its operations by analysing the deviation that expresses weaknesses in the recorded process or the process model.

2.3.3 Process enhancement

When the differences from the To-Be process is identified with the conformance checking techniques, the process enhancement phase follows. Within this phase, the earlier identified improvement potentials are touched. Hassani, van Zelst and van der Aalst (2019) mention that process enhancement aims at increasing the overall view of the process. For instance, by discovering the causal relationships among data attributes in the data and decision points within the process (Hassani et al., 2019). According to van der Aalst et al. (2012), process enhancement aims to extend or improve a- priori process model by using the information of the actual process recorded in some event log.

There are two types of process enhancement: repair and extension (van der Aalst, 2011). A process enhancement can be used to repair the reference model to conform better to the observed behaviour. This reflects better and creates alignment with reality (van der Aalst, 2011). For example, if two activities are modelled in any order, but in reality, this happens in sequence, then the model should be modified to reflect this.

The other type of enhancement is the extension. A process enhancement can be used to extend the reference model with additional information or new perspectives (van der Aalst, 2011). The extension type can add a new perspective to the process mode by cross-correlating it with the log. A process model can be extended with, for example, additional information about recourses, quality metrics and decision rules. Detecting and including data dependencies can affect the routing of process execution (van der Aalst, 2011).

3
Methodology

This paper starts with a problem description to understanding the current problem. The current problem is the need to uncover the difference between the handmade process model of the invoice booking process and the event log generated from the accounting software. And the challenge is to explore the potential of process mining, but the company do not know how to approach this. Therefore, this paper will give particular attention to the methods and outcomes of a process mining technique.

The literature review aimed to understand the fundamental concepts of process mining and the process mining principle by examining various scholars' insights within the process mining domain. The qualitative research provides a theoretical background that helps this paper find the right technique to approach the research question.

To answer the research question, "*What is the difference between the predefined process and the actual process model?"* the company must describe the path of the current invoice processing. Based on this description a model is designed. The modelled behaviour describes which actions must be performed until an invoice is booked. From now on, the designed model behaviour will be named the predefined process model. The predefined process model will be compared to the data generated from the accounting software system.

This research will apply a quantitative research method because it will analyse data generated from the accounting system (ERP system). The quantitative data regarding invoice processing represent primary data which is requested by the database owner. The extracted xlsx file from the database is raw data. The raw data is not suitable for process mining analysis therefore some data integration has been applied to make it usable for processing mining analysis. The data is formalised into meaningful event logs in order to apply process mining analysis.

After the decision is made on which process mining technique is the most applicable to answer the research question, the predefined invoice booking process will be compared to the observed behaviour retrieved from the invoice booking process logs. By doing so, the difference between the predefined process model and the actual process model will be identified. At the end the results will be presented. *Figure 7* visualise the roadmap of the implemented methodology.

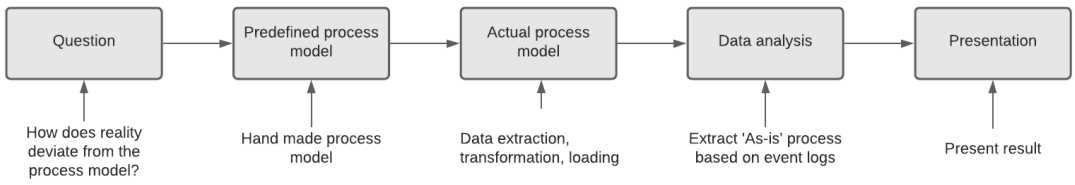


Figure 7 Roadmap

3.1 Predefined process model

The predefined process model describes the invoice booking process that is designed together with the management. This model will be compared against the data generated from the accounting system to identify the differences. The path of invoice processing is designed based on the description of the management. The predefined process model visualises the order of process tasks and describes the actions that must be performed until an invoice is booked.

Section 2.1 provides a detailed description of process model languages used for modelling a process model. Petri net and BMPN are both widely used languages to model business processes. The Petri net is used when studying distributed and concurrent systems. Simultaneously, BPMN is mostly used to communicate the internal processes in a simple way (Burattin, 2015). The BPMN modelling language is compatible with the open-source platform Apromore, while Petri net is not. Apromore will be explained in section 3.3.1.3. Models should be understandable for all concerned, therefore the predefined process model will be expressed using BPMN.

3.2 The actual process

This part will explain how the data is extracted, transformed and loaded in the process mining software to detect deviations between process models. Extract, transform, and load (ETL) is usually carried out when working with databases (Caserta and Kimball, 2013). It refers to the steps involved in the preparation of data from databases for analysis. Extraction refers to pulling raw data from the original source. Transform refers to data cleaning, data integration and data enrichment. Load refers to loading the extracted and transformed data tables in data warehouses or other databases which can be used for reporting or analytics (Gonzalez Lopez de Murillas, 2019).

3.2.1 Data extraction

The event logs used in this study are extracted from an accounting software system used by accountants for digital invoice and documents processing. The software automatically recognises the information on the submitted invoices and makes a booking proposal. Afterwards, the employer checks the booking proposal and books the invoice when the proposed information is correct. While doing this, the software system leaves digital traces at every activity which provides detailed information about the individual task. Examples of the detailed information about the individual task (also called the event or activity) might include - the way in which the invoice entered the accounting system (for example by mail, scan or app), the time and date of the entry of the invoice, and the person who checked the booking proposal. With this system, accountants do not have to enter the information of the invoice manually. This saves them time and effort.

The digital traces of the activities are recorded in the database of the accounting software system. To obtain the real-life event data, an official request by the management of the accounting firm is submitted to the database owner. After this request, access was provided to an extensive database. The next part will explain the data transformation process of the extracted data.

3.2.2 Data transformation

This research will apply a quantitative research method because it will analyse event logs to find deviations between the process map and reality. It will use the data (primary data) generated from the accounting system. Section 2.2 describes in detail that event logs must be extracted from a dataset to apply process mining techniques.

The data requirements for process mining are simple. First, a case ID is needed. A case ID is a process instance or case identifier that identifies a specific execution of the process. Second, an activity name needs to be found in the data, explaining the steps that are being performed in the process. And the third requirement is the timestamp. The timestamp brings everything in the correct order.

To build event logs, the dataset must be transformed into the necessary attributes of event logs because the extracted excel file from the database is raw data and is not yet suitable for process mining analyses. Therefore, some data cleaning and integration has been applied to make it usable for processing mining analysis. The dataset is transformed into the desired format by changing data types, formatting, splitting columns, filtering rows and joining tables. To do this the Power Query M formula language is used. Microsoft Power Query presents a powerful data import experience that contains many features, and it works with Excel, Power BI and Analysis Services. Power Query's core capability is to filter and combine mashup data from one or more sources. This capability is useful for this research because the raw data needs some adjustment to make it suitable for analyses.

3.2.3 Loading transformed data

Extracting and transforming data is carried out to create data that can be loaded in data warehouses, databases, or other software. The transformed data of this paper will be loaded in a process mining software system to conduct further analyses.

There are many offering of open-source frameworks for process mining algorithms. Nowadays, there are over 35 process mining software (Leemans et al., 2021). However, this paper will discuss two open-source platforms for mining processes, namely ProM and Appromore. ProM is an open-source tool developed at the Eindhoven University of Technology, the Netherlands. It supports all standard process mining techniques, such as process discovery, conformance checking, decision-mining, organisational mining, social network analysis and many more. ProM is the most used process mining tool and has over 500 plugins that have been developed by several universities for mining operations in business processes ("ProM Tools", 2020). Until now, there are still research groups that help the growth of ProM. ProM provides a free four-week online course that presents insight into different plugins and interfaces.

Apromore is an open-source tool designed for Business Process Modeling and is developed at the Queensland University of Technology in Brisbane, Australia ("Apromore", 2020). Apromore offers many advanced models of business processes and techniques for analysing, displaying and saving the information content of process models. They aim to support those who want to add functionality to the repository. Apromore provides a three-month demo account for academic purposes. Additionally, they provide online tutorials which give insights into tools functionality, a vivid user manual, and online courses about modelling different process mining techniques, such as process discovery, conformance checking, performance mining, and variant analysis.

Both open-source platforms can be applied to this research. However, this paper used the platform of Apromore because of user-friendliness and the number of teaching materials supplied to support the usage of Apromore. In contrast, ProM gave limited information on how to carry out a project.

3.3 Data analysis

The previous parts explained how the predefined invoice booking process and the actual invoice booking process model will be created. First, the predefined invoice booking process method is presented as a model design based on management meetings. Second, the actual invoice booking process method is based on the data extraction from the accounting software system, data transformation to create desired event logs, and data loading for further analyses. This part will introduce the analysis that will compare the actual practice against the predefined model.

3.3.1 Process mining technique

In sections 2.3 the three main types of process mining techniques are discussed. First of all, the process discovery technique. The process discovery techniques produce a process that shows the process flow of a firm's occurring activities. Second, the conformance checking technique. This technique detects and diagnoses the commonalities and differences of the intended behaviour of a process as described in a process model and event logs that have been recorded during the execution of the process. Third, the process enhancement. This technique aims to extend or improve a- priori process model by using the information of the actual process recorded in the event log.

This research will use the conformance checking technique to map the differences between the predefined invoice booking process and the actual invoice booking process using real-life event logs. This technique is in line with this paper's aim because this paper would like to detect the differences between the predefined process model and the actual process. Therefore, the conformance checking technique seems the most appropriate approach to address the papers' problem.

3.3.2 Variant analyses

The variant analysis will assist the evaluation of conformance checking. The variant analysis identifies categories for standard and non-standard variants by analysing the entire population of real-life event log data of the invoice booking process. The variant analysis provides a fuller understanding of the organisation's deviant business processes. The variant analysis aims to present information about what types of deviations occur in the real-life event log data. This aim will be realised by reviewing the standard and non-standard paths in the organisation's business process and further dividing these paths into three categories: "full activity", "missing activity", and "activity not in the correct order".

The discovered standard and non-standard variants enable this paper to gain insights into real-world business processes that conform to or deviate from the predefined invoice booking process. Besides, with the understanding of standard variants and non-standard variants, organisations can detect deviations, most common paths, undesirable performance variation, inefficient processes, and potential risks.

4

Results

This chapter will describe the application of the roadmap presented in chapter 3. Section 4.1 draws the process map of the activities that should happen. Section 4.2 describes the data extraction, data transformation and data preparation to load the transformed data for data analysis. Section 4.3 performs the conformance analysis based on event logs and present the activities that have happened in real life.

4.1 Process map

The predefined invoice processing model is designed based on the description of the management. After the first talk with the management a draft model is drawn. The draft model is several times discussed and modified until the management agreed that the predefined process model, shown below, represent the reality (Figure 8).

The process map describes the flow of the invoice booking process. The process map involves three participants: the customer supplier, the customer, and the accounting firm. To have a better understanding of the three participants a clarification will be given.

- Accounting firm: Is the firm that raised the question about the process model.
- Customer: Is the customer of the accounting firm, for example ABC B.V.
- Customer suppliers: Are the creditors or debtors of the customer, for example creditor KPNAA B.V. sends an invoice to ABC B.V. In this case KPNAA B.V. is the customer supplier.

The invoice booking process starts at the moment an invoice is uploaded in the software system. However, it is also essential to know the person who sends the invoice to the system to have a complete picture of the invoice booking process.

The process map is divided into three columns. At the top of each column the participant is stated. This means that the underlying activities occur within this participant. The small circles without a letter represent the start point and the endpoint of the process map. The circle with a letter is a connector. In the process map, shown below, the circle means that the process flow stops at one participant and connects to another participant. The triangles in the process map represent a decision point, a question is asked and the answer can be a "yes" or "no". After a decision point an activity occurs. A square represents this. *Figure 8* represent the process map of the incoming invoices.

There are several ways to walk through the invoice booking process (see Figure 8). The path of the invoice booking process depends on the customer's degree of outsourcing the set of activities. Some of the customers carry out the invoice booking process within their firm but use the accounting system. The reason for this might be the interest of keeping control of the invoice booking process within their firm or to reduce accounting expenses. When the customer outsources fewer tasks to the accounting firm, more activities the customer has to undertake and the accounting firm less. This can be up to five activities in each process. Other customers might outsource this service, then the customer decides to send the invoice to the accounting firm and does not involve in any other activity. The process map for the customer stops after two activities. This means that employees of the accounting firm have to undertake more activities to complete the invoice booking process.

The process map describes the actions that must be performed until an invoice is booked. The activities of the process map will be compared against the activities occurring in reality. The next part will explain the data generated from the accounting system. And in the final part, the predefined process model will be compared to the accounting system's data.

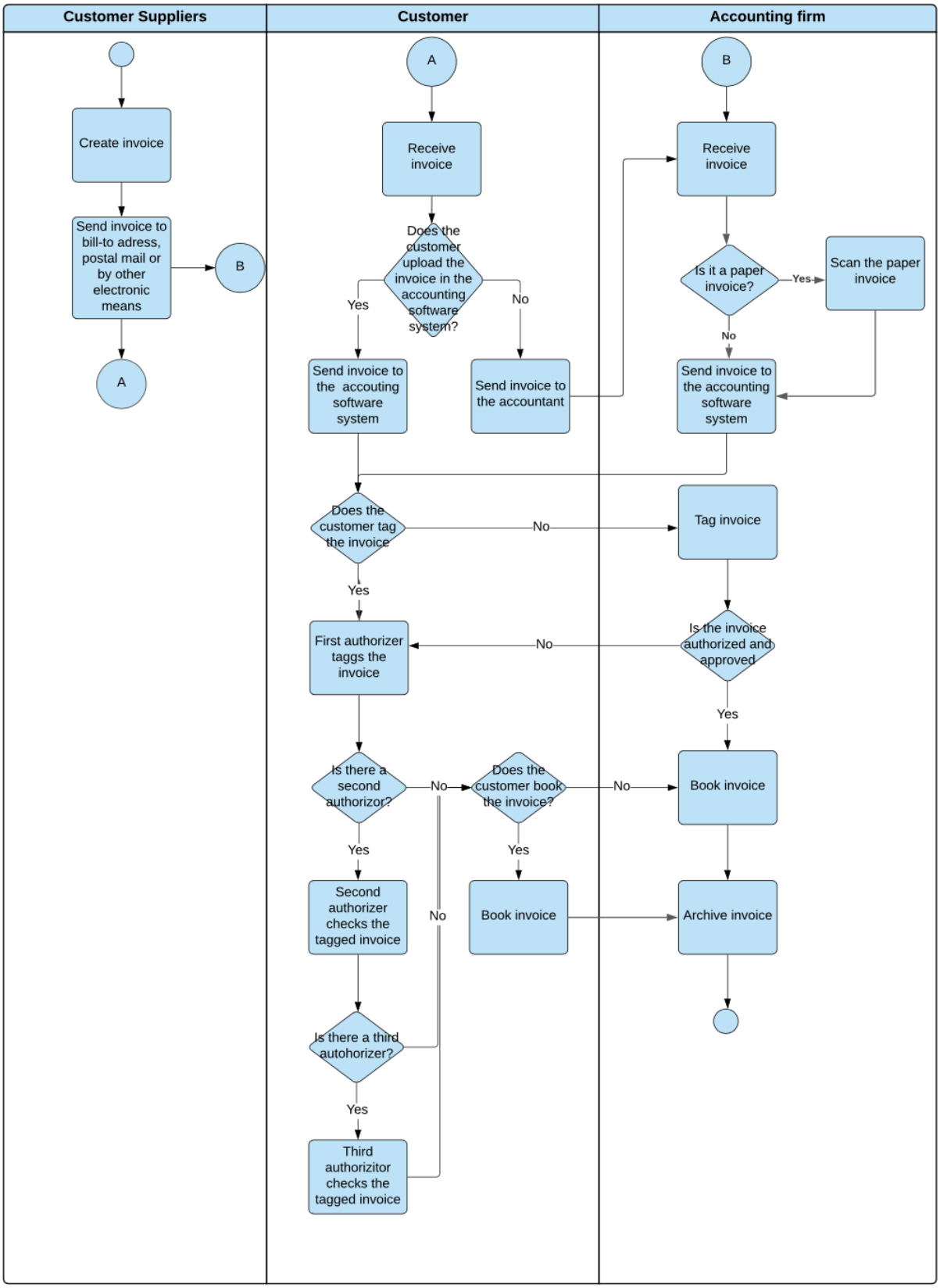


Figure 8 Predefined process map

4.2 The reality

An important part of this research is the data preparation because the event data will be compared to the actual process. This section explains the method of data extraction and data preparation. It explains the data transformation stage, the protection of personal data and the rules and functions applied to the extracted data to prepare it for loading into the process mining software.

4.2.1 Get data

The data regarding the invoice processing represent primary data requested by the database owner of the ERP system. After the request, access was provided to an extensive database. The database gave access to the audit files of 2017, 2018 and 2019. During this research, the 2020 data of the invoice processing was limited and not complete therefore the period from January first, 2019 up to and including December 2019 is chosen.

The extracted raw data had a filename extension of .xlsx – Excel workbook. *Figure 9* shows a fragment of the raw data. The excel file consists of 46 columns and 282.548 rows representing the invoice processing of all invoices received from 01-01-2019 till 31-12-2019. In total 6.636.086 cells contain information.

An additional data file is requested from the department in order to filter the dataset to the department customers. A list with department customers is requested to distinguish all customers and department customers. This is important because the focus of this paper is the invoice booking process of a specific department of the accounting firm. After the request, the department delivered an excel file of 300 customer names. This file will be used to remove the data of the non-department customers.

The raw data is not yet suitable for process mining analysis therefore some data cleaning and data integration will be applied to make it usable for processing mining analysis. The next part will explain how the data is formalised into meaningful event logs to conduct further analyses.

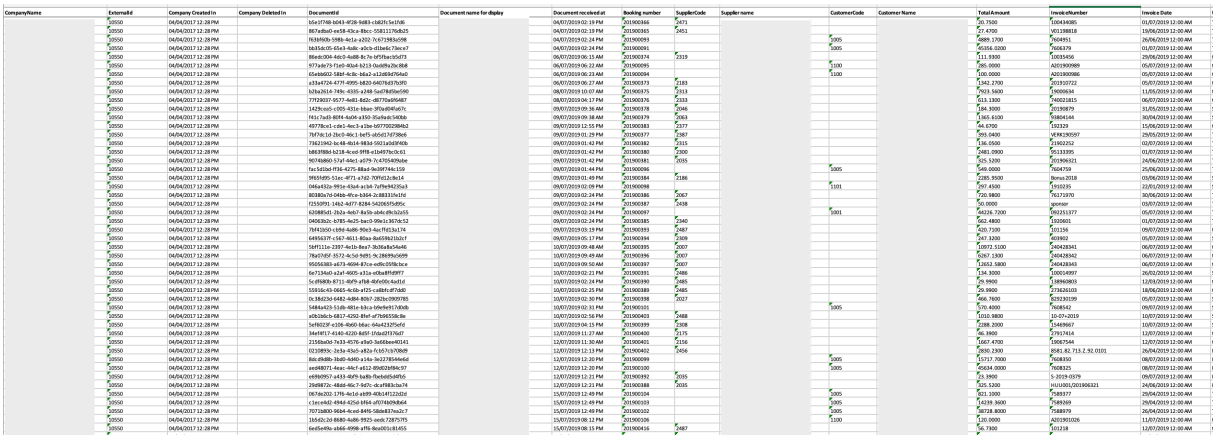


Figure 9 Fragment of raw data

4.2.2 Build event log

The data of the invoice booking process used in this study is extracted from the firm's ERP system. The data represents the invoice booking process of all invoices that led to booking an invoice from January 2019 to December 2019. The dataset is transformed into the desired format to create event logs by removing unnecessary data, changing data types, formatting, splitting columns, filtering rows and joining tables. For example, the raw data includes personal data, and this is removed from the analyses because of the protection of data privacy (GDPR, 2021). The raw data also contains information that is not relevant for this analysis, and this is also removed from the dataset. Appendix 3 list the 24 attributes that are eliminated from the analysis. The remaining 22 attributes are used to build event logs.

The data requirements for process mining are simple. First, a case ID is needed. The case ID identifies a specific execution of the process and refers to the source of the data. This analysis cannot use personal data therefore the personal data is replaced by an ID of random numbers. Second, an activity name needs to be found in the data. This is about the steps that are being performed in the

process. And the third requirement is the timestamp. The timestamp brings everything in the correct order. The first step in the event log building process is to merge the list of department customers with the Audit file 2019. This is done in order to remove the data of non- department customers. Before the merge the dataset consisted of 282.548 rows. After the merge the dataset consists of 87.758 rows.

The dataset is still not in the required format because the activities, resources and timestamps of a process instance, also called a case, is represented in the rows. This means that the invoice booking process is explained in one row. To change this, the column rows are converted into tables. Additionally, a case ID is added to count the number of cases. This led to a total number of five columns (External ID, Case ID, Activity, Resource, and Timestamp) and 269.009 rows. Appendix 4 shows the steps that are undertaken to transform the data into meaningful event logs.

TABLE 3
Description of Event log: Invoice booking process from an accounting software application

Event	269.009
Process instance	87.758
Activity	5
Activity Detail	(1) Received
	(2) Tagged
	(3) Booked
	(4) Inspected by authoriser 1
	(5) Inspected by authoriser 2
Variant	18
Start	01 January 2019, 01:00
End	31 December 2019, 17:17

The log represents 87.758 process instances (a process instance in this event log is a single line of a case in the invoice process). The event log contains 269.009 events and includes five activity types (Received, Tagged, Booked, Inspected by authoriser 1, and Inspected by authoriser 2). All process instances are grouped into 18 variants. Table 3 and Table 4 show the information and statistics about the event log data.

TABLE 4
Activity Frequency

Activity	**Count**	**Percentage**
Received	87.756	31,623%
Tagged	87.755	32,623%
Booked	87.755	32,623%
Inspected by authoriser 1	5.655	2,102%
Inspected by authoriser 2	80	0,03%

4.2.3 Load event logs

The transformed real-life log data is loaded in the process mining software system and creates the process map of the invoice booking process shown in *Figure 10*. The arrows in *Figure 10* represent the process's frequency and direction. The dark and thick arrows stand for more frequent business processes. The number beneath each activity in each box represents such activity's overall occurrence. For example, "Inspected by authoriser 1" happened 5655 times and "Received" happened 87758 times within the dataset. When a box is darker than another box, it means that this particular activity happened more frequently. *Figure 10* presents the process map of the invoice booking process with five activities and eight gateways.

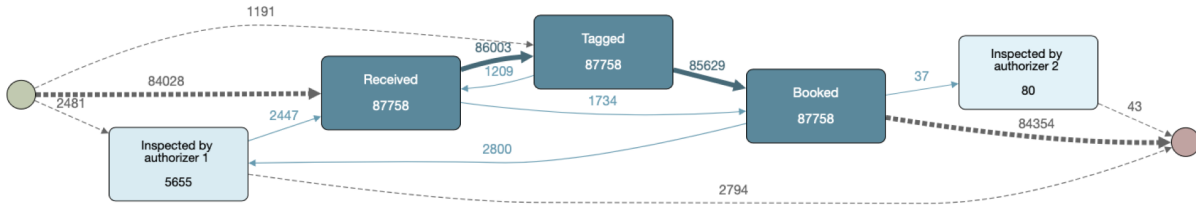


Figure 10 Actual process

4.3 Conformance analysis

The conformance checking technique detects and diagnoses the commonalities and differences of a process's intended behaviour as described in a process model and event logs that have been recorded during the execution of the process. The variant analysis will assist the evaluation of conformance checking by discovering standard and non-standard variants. The analysis enables this paper to gain insights into real-world business processes that conform to or deviate from the predefined invoice booking process.

4.3.1 Variant analysis

The variant analysis examines the 18 variants from the event log's entire population and lists these variants into standard and non-standard variants. One category for the standard variant and three categories for non-standard variants are discovered by analysing the event log's entire population. The standard variant consists of one variant, which is the "Standard invoice booking process" because it contains the required activities in the invoice booking process. The three categories for non-standard variants consist of 17 variants, as shown in Table 5. The three categories of non-standard variants are: "Full activity", "Missing activity" and "Activity not in the correct order". The category "Full activity" describes the process paths that contain five, the maximum number of, activities. The category "Missing activity" describes the process paths that miss one of the activities specified in the predefined invoice booking process. The category "Activity not in the correct order" describes the process paths with unexpected or unusual activity orders.

Table 6 represent the classification results of the standard and non-standard variants. This table shows that 91.54 percent of the variants are categorised into standard variants, and 8.56 percent of the variants are classified as non-standard variants. The standard variant could describe the efficient and most common paths. The non-standard variants could describe inefficient processes and possible errors in the business processes. This depends on whether the organisation's business rules acknowledge certain deviations from the standard variants. The identified non-standard variant could be further compared with the firm's policies to separate approved variant paths from unapproved variants.

The goal of recognising the standard and non-standard variants is to generalise possible standard and non-standard business processes in real-life companies. This study's classification results illustrate the total population of the event log into standard or non-standard categories. The identified non-standard variants can be understood as deviations that could be prioritised in the risk assessment method to improve efficiency.

TABLE 5
Standard and non-standard variants

Standard variant

Category	Activity order	Cases	Frequency
Standard invoice booking process	Received- Tagged- Booked	80333	91.54%

Non-standard variant

Category	Activity order	Cases	Frequency
Full activity	Inspected by authoriser 1- Received- Tagged- Booked- Inspected by authoriser 2	36	0.04%
Full activity	Received- Tagged- Booked- Inspected by authoriser 1- Inspected by authoriser 2	6	0.01%
Full activity	Inspected by authoriser 1- Inspected by authoriser 2- Received- Tagged- Booked	3	0%
Full activity total		**45**	**0.05%**
Missing activity	Received- Tagged- Booked- Inspected by authoriser 1	2787	3.18%
Missing activity	Inspected by authoriser 1- Received- Tagged- Booked	2411	2.75%
Missing activity	Received- Tagged – Inspected by authoriser 1- Booked	370	0.42%
Missing activity	Inspected by authoriser 2- Received- Tagged- Booked	34	0.04%
Missing activity	Received- Inspected by authoriser 1- Tagged- Booked	3	0%
Missing activity	Received- Tagged- Booked- Inspected by authoriser 2	1	0%
Missing activity total		**5606**	**6.39%**
Activity not in the correct order	Tagged- Received- Booked	1168	1.33%
Activity not in the correct order	Received- Booked- Tagged	528	0.6%
Activity not in the correct order	Inspected by authoriser 1- Tagged- Received- Booked	31	0.04%
Activity not in the correct order	Booked- Received- Tagged	22	0.03%
Activity not in the correct order	Tagged- Booked- Received	15	0.02%
Activity not in the correct order	Tagged- Received- Booked- Inspected by authoriser 1	7	0.01%
Activity not in the correct order	Booked- Tagged- Received	2	0%
Activity not in the correct order	Tagged- Received- Inspected by authoriser 1- Booked	1	0%
Activity not in the correct order total		**1774**	**2.03%**

The number of cases and the percentage of each classification for the one category of standard variant and three categories of non-standard variant are presented in the third and fourth columns of Table 5. The standard variant, "Standard invoice booking process", occurs 91.54% times in the entire population of real-life event log data of the invoice booking process. The sum of the standard variant is exceeding the total non-standard variant in the event log data, as presented in Table 5 and 6. The three activities, "Received- Tagged- Booked", are the required activities in the invoice booking process. This routing is the most common path and without one of these activities, an invoice cannot be booked. "Full activity" is the category that represents 0.05% and is the routing that is the least often performed. The categories "Missing activity" and "Activity not in the correct order" require more attention because their presence could be an inefficient business process or a possible error in the business processes.

The variant analysis's categorisation results can help identify the occurrence, completeness, and accuracy. The category "Missing activity" can be linked to "occurrence" and "completeness". For example, the management can examine whether all process instances have the activity "Tagged". This

activity can help detect unapproved documents (occurrence). Also, the management can determine whether a process instance has both "Received" and "Booked" activities to ensure that invoices that are received are booked (completeness). The category "Activity not in the correct order" can be related to "accuracy". For example, testing if all process instances have "Tagged" happening after "Received" is essential to identify because this gives insights about whether the process has been completed correctly (accuracy).

TABLE 6
Results variant analysis

	Count	Percentage
Standard variant	80333	91.54%
Non- standard variant	7425	8.56%
Total	87758	100%

4.3.1.1 Standard invoice booking process

The standard variant consists of one variant which is the "Standard invoice booking process" because it contains the required activities in the invoice booking process. "Standard invoice booking process" occurs 91.54% times in the entire population of real-life event log data of the invoice booking process. The three activities, "Received- Tagged- Booked", are the required activities in the invoice booking process. This routing is the most common path and without one of these activities, an invoice cannot be booked (Figure 11). The processing time of this category is 1.93 days (Table 7), which has the second fastest running time.

Figure 11 Standard invoice booking process map

4.3.1.2 Full activity

The category "Full activity" describes the process paths that contain five, the maximum number of, activities (Figure 12). "Full activity" is the category that represents 0.05% and is the routing that is the least often performed. This category has a processing time of 7.03 months (Table 7), which is the category with the slowest running time.

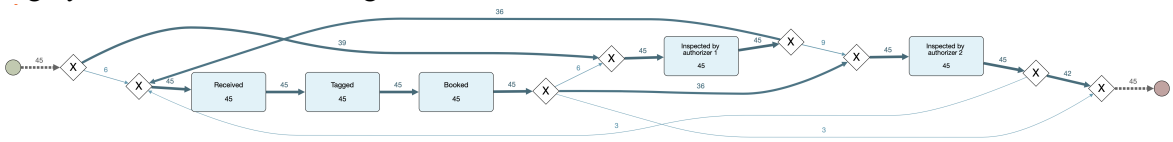


Figure 12 Full activity process map

4.3.1.3 Missing activity

The category "Missing activity" describes the process paths that miss one of the activities specified in the predefined invoice booking process. The categories "Missing activity" occurs 6.39% times and is the second most performed routing (Figure 13). Processing time of this category is 4.31 months (Table 7). The presence could be an inefficient business process or a possible error in the business processes.

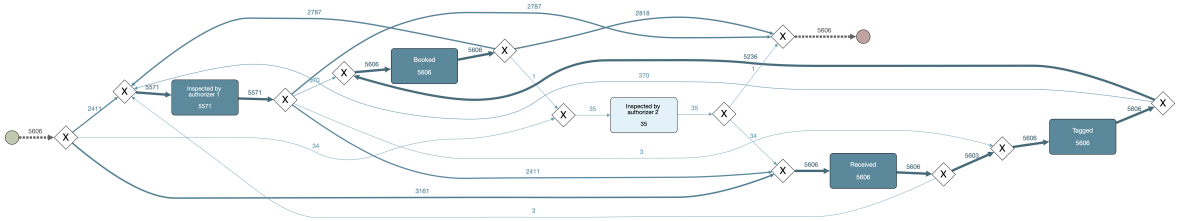


Figure 13 Missing activity process map

4.3.1.4 Activity not in the right order

The category "Activity not in the correct order" describes the process paths with unexpected or unusual activity orders. For example, an invoice booking process that has "booked" occurs before "received". The categories "Activity not in the correct order" occurs 2.03% times and is the third most performed routing (Figure 14). This category has a processing time of 37.50 minutes (Table 7), which is the category with the fastest running time.

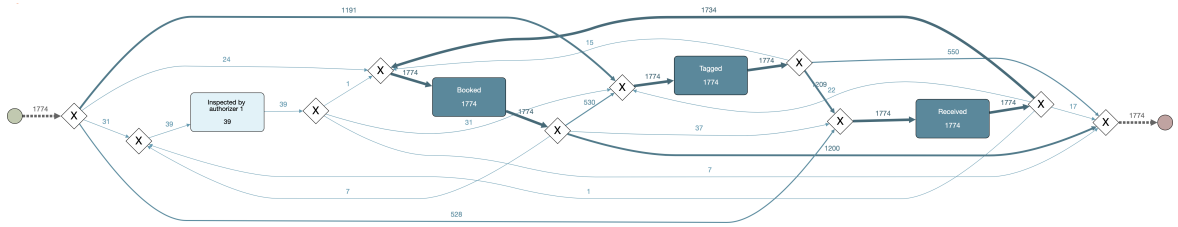


Figure 14 Activity not in the right order process map

TABLE 7
Case duration of each category

	Min	Median	Average	Max
Standard invoice booking process	Instant	1.93 days	6.45 days	1.43 years
Full activity	1.99 months	7.03 months	6 months	9 months
Missing activity	1.15 hours	3.09 months	4.31 months	19.55 years
Activity not in the correct order	Instant	37.50 min	1.05 weeks	9.89 months

5

Conclusion

The present study aimed to examine the alignment and deviation between the activities of the expected behaviour and the actual observed behaviour of the invoice processing. Based on this aim, this study seeks to answer the following research question: "*What is the difference between the predefined process and the actual process model?".*

The accounting firm mentioned a need to uncover the difference between the handmade process model of the invoice booking process and the event log generated from the accounting software application. They would like to have fact-based insights about the predefined invoice booking process and the actual observed behaviour of the invoice booking process using a process mining technique. A process mining technique will be used because the company mentioned that they would like to explore the potential of process mining.

The focus of this paper is on the alignments and deviations between the reality and modelled invoice process. The predefined invoice booking process will be compared to the observed invoice booking process retrieved from the logs. Understanding the deviated paths of the invoice booking processes is interesting because the accounting firm might know which customers do not comply with the predefined business model. And understanding the invoice processing path might help the accounting firm to check the quality of the predefined model. Process mining shows the alignment and deviation between the expected behaviour and the actual observed behaviour of the invoice processing. The accounting firm might use the information about the alignment and variations for inspection and control purpose. Furthermore, the company will acquire knowledge of process mining to apply for other internal processes within the same context.

This study starts with a literature review to understand the essential concepts for process mining and the process mining principle by examining various scholars' insights within the process mining domain. Then, based on the literature and this paper's aim, a process mining technique is chosen. Finally, the conformance checking technique is considered the most applicable method to answer the central research question. Because with conformance checking, a process model is compared with the logs of the same process model.

This study has identified that the predefined invoice processing model clearly states each activity's participants; however, in reality, this is not very clear. It is not possible to define who has uploaded the invoice in the accounting system. The invoice can be uploaded via the app by the customer but also by the employers. The data only mentions that the invoice is uploaded by the app but not by who. Also, the predefined invoice processing model discusses several activities before the invoice entered the accounting system, but this information is not available in the logs because no notation is made of these activities.

This study visualizes the actual invoice booking process in a model. Figure 10 presents the process map of the invoice booking process with five activities and eight gateways. The most common activity order is "received-tagged-booked", which appeared in 91,54 % of the cases. The predefined process model also notates these activities, however the visualization provides fewer insights about the frequency.

The result of the process mining analysis shows that in reality the process model consists of 18 variants. These variants are divided into one "standard variant" and three "non-standard variants" (table 5). The predefined process model displays seven decisions points that can lead to different flows. However, this model does not clearly describe the number of variants. Additionally, this study has found that generally 8,56% of the variants (7425 process instances) are classified as non-standard variants.

They are indicating that these variants do not conform to the standard invoice booking process. The research has also shown that the process flow "received-tagged-booked" has a case duration that has on average 6.45 days, which is the category with the slowest running time on average.

5.1 Implications for the management

The management had the problem that different customers use different billing processes. The accounting firm has included the different billing processes in their invoicing process, making it unclear which process flow variants exist and which one is the most efficient. This had the consequence that the management felt that they miss the opportunity to be more efficient. With this research, insight is gained into process variants and the duration of each variant. For example, the standard invoice booking process consists of the activities "received – tagged – booked" and has an average duration of 6,45 which has the fastest running time. Also, the company can check the quality of the predefined model and inform about the alignment and variations for inspection and control purpose. In addition, the company acquires knowledge of process mining to apply for other internal processes within the same context. With the insights of this research, the company can further investigate which employees or clients do not perform the standard invoice booking process. This investigation will result in saving process time and better service provision.

6

Discussion

This chapter discusses the limitations of the study, the practical and the theoretical implications and areas for future research.

6.1 Limitations

The first limitation was that this research might have overlooked some relevant publications. When conducting a systematic literature review, there is a risk that relevant publications might be omitted. To conduct a comprehensive literature review, this research kept the topical focus on the process model, business process management, business process, event logs, event log analysis, workflow management, process mining and process models alone. This might be seen as a limitation because this literature review will not meet the need of readers looking for a review on conformance checking in general. However, the concern about potentially overlooking relevant material is addressed by cross-checking the search topic in additional databases.

The second limitation was that the researcher could not extract the data from the accounting system by herself. Although the system automatically recorded the event log, the top management might have access to alter or delete the existing data. Therefore, it is crucial to ensure the integrity of data before conducting process mining analysis. Additionally, the researcher could not obtain additional data, however, incorporating other variables might have provided more insight into the findings.

The third limitation is that the data did not distinguish between the app file upload by the customers or employers. It is not possible to define who has uploaded the invoice in the accounting system. The invoice can be uploaded via the app by the customer but also by the employers. The data only mentions that the invoice is uploaded by the app but not by who.

The fourth limitation is that a flow analysis does not consider whether a process behaves differently depending on the load. For example, the cycle time for handling invoices would be much slower if the company handles thousands of invoices at once. This might happen in the period when

customers must file the VAT declaration. However, when the load goes up, and the number of resources (for example, invoice checkers) remains nearly constant, in this case the processing times will be longer. This is known as resource contention. Resource contention happens when there is more work to be done than resources available to perform the job. Some tasks will be in a waiting mode in such scenarios until one of the necessary resources are finished. The flow analysis does not warn about the consequences of increased resource contention.

The fifth limitation is that the categories of standard and non-standard variants are based on the standard invoice booking process. But these categories need to be adjusted when considering event logs from other business processes.

The sixth limitation is that it is difficult to analyse where the faults are in the process mining technique.

6.2 Theoretical and practical implications

Process mining is a relatively new domain that combines data mining and process modelling. The process mining phenomenon increased in recent year. However, limited study has been done approaching the improvement of data quality in the process mining domain. Process mining research mainly concentrates on the development of new methods or the application of existing techniques. Still, the quality of all analyses eventually depends on the quality of the event log.

This paper expands previous studies in a similar area by: (1) considering the entire population of event logs, (2) presenting categories of standard and non-standard variants depending on a real-life business process from an accounting firm. This research captured existing approaches and tools in transforming the data of invoice processing to event logs. The process mining technique demonstrated to be a helpful method of recognizing differences in a process. The proposed process mining analyses can eventually become an automatic analytical tool that allows the firm to identify unusual behaviors based on the entire population of event logs. Process mining shows a broad view of the process flows but does not show the causes and consequences. Therefore, process mining alone is not suitable for an in-depth analysis of an invoice booking process as the sole research method.

With the findings of this study, the company can check the quality of the predefined model and inform about the alignment and variations for inspection and control purpose. In addition, the company acquires knowledge of process mining to apply for other internal processes within the same context.

6.3 Future research

This paper has several future research suggestions that will be explained in this part. Future research could compare the categories of standard and non-standard variants with the organization's business rules. This could classify whether the non-standard variants correspond to business rules and define the riskiness of each category in the non-standard variants. It is also interesting to research by who the invoices are entered in the accounting system, currently this could not be analysed because of lack of data. Additionally, in the introduction of this research it is mentioned that this paper will not focus on the cause of deviant behaviour and the initiatives that increase efficiency. But future research should analyse the reasons for deviant behaviour and initiatives that increase efficiency. This will provide a thorough process understanding and improvement. Based on the in-depth insights, actions might be described, deployed, and tracked.

Reference

Apromore. (2020). *About Us - Apromore*. [online] Retrieved 12 December 2020, from https://apromore.org/about/

Augusto, A., Dumas, M., Maggi, F. M., Soo, A., Conforti, R., La Rosa, Marrella, A. Mecella, M. (2019). Automated discovery of process models from event logs: review and benchmark. *Ieee Transactions on Knowledge and Data Engineering*, *31*(4), 686–705. https://doi.org/10.1109/TKDE.2018.2841877

Bhimani, A., & Willcocks, L. (2014). Digitisation, 'Big Data' and the transformation of accounting information. *Accounting And Business Research*, *44*(4), 469-490. https://doi.org/10.1080/00014788.2014.910051

Bpmb.de. (2011). Retrieved 9 September 2020 from http://www.bpmb.de/images/BPMN2_0_Poster_EN.pdf

Burattin, A. (2015). *Process mining techniques in business environments : theoretical aspects, algorithms, techniques and open challenges in process mining*(Ser. Lecture notes in business information processing, 207). Springer. https://doi.org/10.1007/978-3-319-17482-2

Carmona, J., Dongen, B. van, Solti, A., & Weidlich, M. (2018). *Conformance checking : relating processes and models*. Springer. https://doi.org/10.1007/978-3-319-99414-7_3

Caserta, J. and Kimball, R., 2013. *The Data Warehouse ETL Toolkit*. Hoboken, N.J.: Wiley, p.20.

Chiu, T., & Jans, M. (2019). Process mining of event logs: a case study evaluating internal control effectiveness. *Accounting Horizons*, *33*(3), 141–156. https://doi.org/10.2308/acch-52458

Cook JE, He C, Ma C. Measuring behavioral correspondence to a timed concurrent model. In: Proceedings be of the 2001 International Conference on Software Maintenance Florence, Italy: IEEE; 2001, 332–341.

Cook, JE., & Wolf, AL. (1999). Software process validation. *ACM Transactions On Software Engineering And Methodology*, *8*(2), 147-176. https://doi.org/10.1145/304399.304401

Dijkman, R. M., Dumas, M., & Ouyang, C. (2008). Semantics and analysis of business process models in bpmn. *Information and Software Technology*, *50*(12), 1281–1294. https://doi.org/10.1016/j.infsof.2008.02.006

Dumas, M., La Rosa, M., Mendling, J., & Reijers, H. A. (2018). Fundamentals of business process management (Second). Springer. https://doi.org/10.1007/978-3-662-56509-4

General Data Protection Regulation (GDPR) – Official Legal Text. General Data Protection Regulation (GDPR). (2021). Retrieved 12 January 2021, from https://gdpr-info.eu.

Gonzalez Lopez de Murillas, E. (2019). *Process mining on databases: extracting event data from real-life data sources.* Technische Universiteit Eindhoven.

Hassani, M., Zelst, S. J., & Aalst, W. M. P. (2019). On the application of sequential pattern mining primitives to process discovery: overview, outlook and opportunity identification. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(6). https://doi.org/10.1002/widm.131

Jans, M., Alles, M., & Vasarhelyi, M. (2013). The case for process mining in auditing: Sources of value added and areas of application. *International Journal Of Accounting Information Systems*, 14(1), 1-20. https://doi.org/10.1016/j.accinf.2012.06.015

Knudsen, D. (2020). Elusive boundaries, power relations, and knowledge production: A systematic review of the literature on digitalisation in accounting. *International Journal Of Accounting Information Systems*, *36*, 100441. https://doi.org/10.1016/j.accinf.2019.100441

Kindler, E. (2009). Model-Based Software Engineering and Process-Aware Information Systems. *Transactions On Petri Nets And Other Models Of Concurrency II*, 27-45. https://doi.org/10.1007/978-3-642-00899-3_2

Leemans, S., Fahland, D., & van der Aalst, W. (2013). Discovering Block-Structured Process Models from Event Logs - A Constructive Approach. *Application And Theory Of Petri Nets And Concurrency*, 311-329. https://doi.org/10.1007/978-3- 642-38697-8_17

Leemans, S. J. J., van der Aalst, W. M. P., Brockhoff, T., & Polyvyanyy, A. (2021). Stochastic process mining: earth movers' stochastic conformance. *Information Systems*. https://doi.org/10.1016/j.is.2021.101724

Mans, R. S., Schonenberg, M. H., Song, M., van der Aalst, W. M. P., & Bakker, P. J. M. (2008). Application of process mining in healthcare–a case study in a dutch hospital. *International Joint Conference on Biomedical Engineering Systems and Technologies*, 425–438.

Munoz-Gama, J. (2016). Conformance Checking and Diagnosis in Process Mining. *Lecture Notes In Business Information Processing*. https://doi.org/10.1007/978-3-319-49451-7

Murata, T. (1989). Petri nets: properties, analysis and applications. *Proceedings of the Ieee*, *77*(4), 541–580.

OMG. Business Process Model and Notation (BPMN). Object Management Group, dtc/10-05-04, 2014.

ProM Tools. Promtools.org. (2020). Retrieved 12 December 2020, from https://www.promtools.org/doku.php.

Apromore. (2020). *About Us - Apromore*. [online] Retrieved 12 December 2020, from https://apromore.org/about/

Rozinat, A., & van der Aalst, W. M. P. (2008). Conformance checking of processes based on monitoring real behavior. *Information Systems, 33*(1), 64–95. https://doi.org/10.1016/j.is.2007.07.001

van der Aalst, W., van Dongen, B., Herbst, J., Maruster, L., Schimm, G., & Weijters, A. (2003). Workflow mining: A survey of issues and approaches. *Data & Knowledge Engineering*, *47*(2), 237-267. https://doi.org/10.1016/s0169-023x(03)00066-1

van der Aalst, W. M. P., Weijters, A. J. M. M., & Maruster, L. (2004). Workflow mining: discovering process models from event logs. *Ieee Transactions on Knowledge and Data Engineering*, *16*(9), 1128–1142.

van der Aalst, W. (2011). *Process mining: discovery, conformance and enhancement of business processes*. Springer. https://doi.org/10.1007/978-3-642-19345-3

van der Aalst, W., Adriansyah, A., & van Dongen, B. (2012). Replaying history on process models for conformance checking and performance analysis. *Wires Data Mining And Knowledge Discovery*, *2*(2), 182-192. https://doi.org/10.1002/widm.1045

van der Aalst, W. M. P. (2013). Business process management: A comprehensive survey. ISRN Software Engineering, 1–37. https://doi.org/10.1155/2013/507984

van der Aalst, W. (2016). Process mining: data science in action (Second). Springer. https://doi.org/10.1007/978-3-662-49851-4

van der Aalst, W. (2018). Process discovery from event data: Relating models and logs through abstractions. *Wiley Interdisciplinary Reviews: Data Mining And Knowledge Discovery*, *8*(3), e1244. https://doi.org/10.1002/widm.1244

Weijters, A., & van der Aalst, W. (2003). Rediscovering workflow models from event-based data using little thumb. *Integrated Computer-Aided Engineering*, *10*(2), 151-162. https://doi.org/10.3233/ica-2003-10205

Appendix 1. Method of literature research

Literature reviews present, in a reasonable way, insight from various scholars within a domain. Eisenhardt (1989) specifies that comparing existing literature is crucial for theory building. Webster and Watson (2002) mention that prior research knowledge is necessary for every academic work because it sets the groundwork for new knowledge creation and theory testing. Additionally, they mention that systematic literature reviews are appropriate when the goal is to obtain an overview of an emerging issue or concept, such as process mining (Webster and Watson, 2002). This research will use Webster and Watson's framework to structure and analyse the emerging literature on process mining. The main objective of the literature review is:

1. Describe process models theory and introduce modelling languages
2. Describe event log and log requirements
3. Identify a process mining technique that is useful for the research aim

A broad search strategy is used to identify relevant articles on online databases. The literature search acquired relevant studies at databases Scopus, Web of Science, ScienceDirect, and Google Scholar with the following search terms:

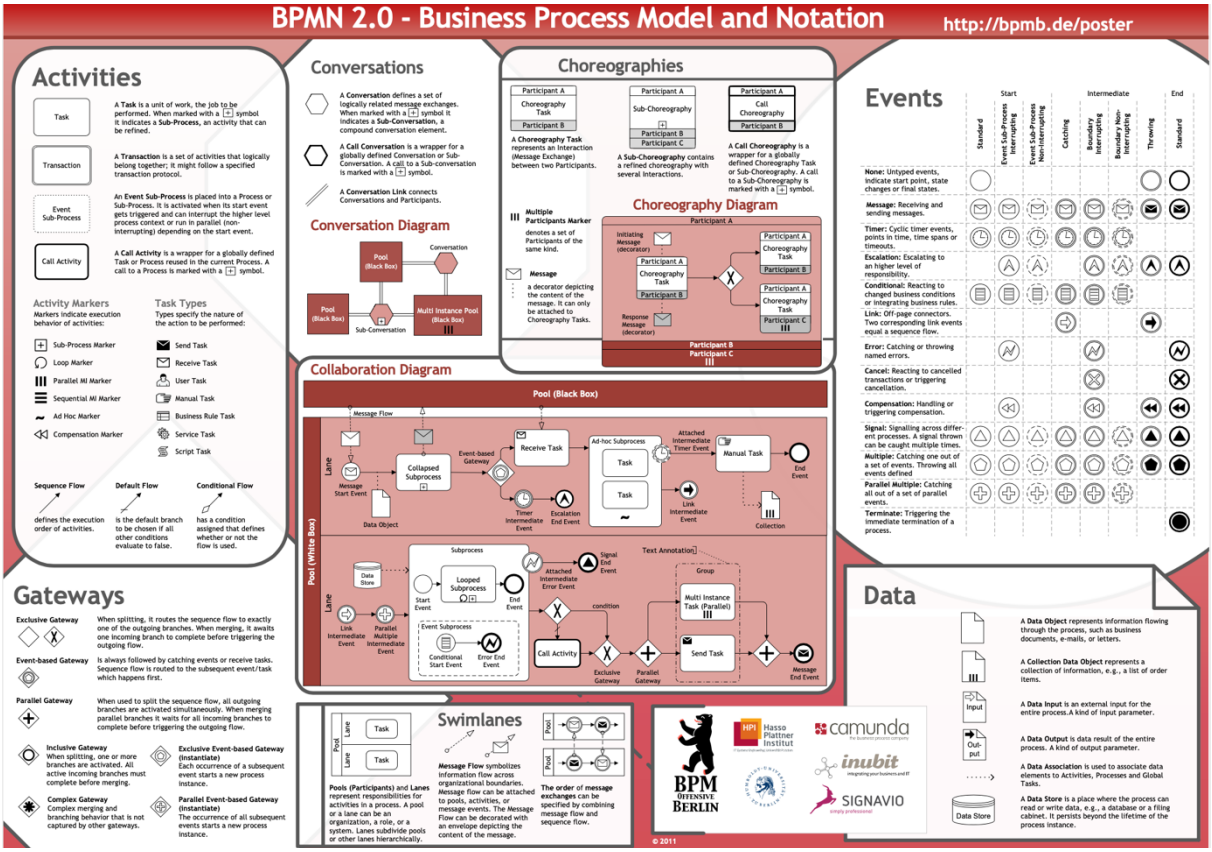
- Process model, business process management, business process
- Event logs, event log analysis, workflow management
- Process mining, process models

The articles are reviewed and assessed based on the articles title, abstract, and conclusion. Additional three review criteria are formulated namely.

1. Limit the search to peer-reviewed academic journals and exclude unpublished materials
2. To ensure the article's quality, respect highly cited papers and exclude papers that are not cited, although an exception is made for new articles (from 2018 to 2020).
3. Articles published in the English language are examined.

Van der Aalst and Dumas et al. add a significant contribution to the research on event logs, process models, process mining and conformance checking because they provide an excellent theory of various variables. Wil van der Aalst' is a bestseller on process mining and therefore his papers are most often used in this research.

Appendix 2. BPMN 2.0 poster



Appendix 3. Removed attributes

1. CompanyName,
2. Company created in
3. Company deleted in
4. Document ID
5. Document name of display
6. Booking number
7. Supplier code
8. Supplier name
9. Customer code
10. Customer name
11. Total amount
12. Invoice number
13. Invoice date
14. Email of uploader
15. Before email of authoriser 1
16. Before email of authoriser 2
17. Before email of authoriser 3
18. After email of authoriser 1
19. After email of authoriser 2
20. After email of authoriser 3
21. After email of tagger
22. After email of booker
23. Document type
24. Email of the original sender

Appendix 4. Steps for event log building

- Collect a dataset with the invoice booking process
- Dataset is collected and the file is called "Audit file 2019" with 282.584 rows
- Import Audit file 2019 file into the Query editor
- Delete colums:
    1. CompanyName,
    2. Company created in,
    3. Company deleted in
    4. document ID,
    5. document name of display,
    6. booking number,
    7. supplier code,
    8. supplier name,
    9. customer code,
    10. customer name,
    11. total amount,
    12. Invoice number,
    13. invoice date,
    14. email of uploader,
    15. Before email of authoriser 1,
    16. Before email of authoriser 2,
    17. Before email of authoriser 3,
    18. After email of authoriser 1,
    19. After email of authoriser 2,
    20. After email of authoriser 3,
    21. After email of tagger,
    22. After email of booker,
    23. document type
    24. email of the original sender.
- Collect a dataset with only department customers
- Dataset is collected and the file is called "BTW voorganglijst" with 331 rows
- Import the BTW voortganglijst file into the Query editor
- Make sure that both data files have a unique table that match both files
- A match is detected between "External ID" in Audit file 2019 and "Application code" in BTW voorganglijst
- Rename "Application code" of BTW voorganglijst code into "External ID"
- Remove the duplicate "External ID" in the BTW voorganglijst
- After removing duplicates 300 rows remained, meaning we have 300 customers with unique ID's
- Merge the two excel files and filter on "External ID" (Inner "Alleen overeenkomende rijen")
- After merge 87.758 rows, 25 colums and 213 unique External IDs are left
    - Audit file 2019 (after merge 87758) (before merge 282548)
- Use the UNPIVOT statement to convert columns to rows
- Convert the column Case ID, External ID, Activity and Resource to type "text" and Timestamp into "date"
- The log consists of 263.274 events