UNIVERSITY OF TARTU

Institute of Computer Science
Software Engineering Curriculum

Karli Oruste

Process Mining in Industry

Master's Thesis (30 ECTS)

Supervisors:
Frederik Payman Milani
Fabrizio Maria Maggi

Tartu 2017

Process Mining in Industry

Abstract:

Process mining is a set of analysis techniques that provides a data-based overview of how business processes are actually executed. In order to use process mining techniques the data about the business process execution has to be recorded into a chronological sequence of activities called event logs. It is a quite young research area and has not yet been widely adapted by the industry. However,more and more research is being produced in the field. In this paper a systematic literature review was conducted to identify all the different process mining techniques that have been tested on real-life logs. This review was used as an input to a survey among industry representatives to understand how useful each process mining technique is considered from the perspective of the industry. The target group of the survey were industry representatives in Estonia, who were compared with industry representatives from around the world.

Keywords:

Process mining, literature review, industry survey

CERCS: P170 Computer science, numerical analysis, systems, control

Protsessikaeve ettevõtetes

Lühikokkuvõte:

Protsessikaeve on kogum analüüstehnikaid, mis võimaldab saada andmete põhist ülevaadet äriprotsesside tegelikust toimimisest. Protsessikaeve kasutamiseks tuleb äriprotsessi täitmise andmed salvestada spetsiaalselt protsessikaeve jaoks loodud andmetüüpi – sündmuslogisse. Sündmuslogi on äriprotsessi kronoloogiline tegevuste järjekord. Antud uurimisvaldkond on üsnagi noor ning seda pole veel ettevõtetes laialdaselt kasutusele võetud. Siiski, protsessikaeve kohta ilmub üha rohkem teadustöid. Antud töös tehakse süstemaatiline kirjanduse ülevaade, mis selgitab välja erinevad protsessikaeve tehnikad, mida on testitud kasutades päris elulisi sündmuslogisid. Koostatud kirjanduse ülevaadet kasutati küsitluse läbi viimiseks ettevõtete esindajate seas, et aru saada kui kasulikus iga protsessikaeve tehnikat peetakse. Küsitluse sihtgrupp olid Eesti ettevõtete esindajad kelle tulemusi võrreldi ettevõtete esindajate üle maailma.

Võtmesõnad:

Protsessikaeve, kirjanduse ülevaade, ettevõtete küsitlus

CERCS: P170 Arvutiteadus, arvutusmeetodid, süsteemid, juhtimine (automaatjuhtimisteooria)

Table of Contents

1 Introduction

Process mining is a set of analysis techniques that provides a data-based overview on how business processes are executed in real life. Usually the people performing the business processes think they have a pretty good overview on how the processes work. Process mining uses the historical process data to confirm or refute this belief. As the experience of process mining practitioners has shown in many cases, the results of conducting a process mining effort can prove enlightening even for quite experienced process performers.

To start using process mining there are certain preconditions. As mentioned before process mining is a data-based approach. That means the data about the business process execution needs to be captured in an IT system. Furthermore, a business process consists of many small sub steps or activities. Therefore the IT system needs to record events that that can clearly be mapped to these activities. Also, these activities need to have a timestamp representing the time when the event was executed and a case ID referencing the business process instace,the event is part of. The recorded events can have more activities, but the case ID, activity and timestamp are the bare minimum for using process mining.

If all the necessary data is in the system then it has to be extracted, pre-processed and an event log has to be constructed. An event log is a set of events all belonging to the same business process. It is the underlying data type upon which all process mining techniques are built on. Once an event log is constructed process mining techniques can be applied.

As a research area, process mining is relatively young. To categorize it one can say it is bordered by data mining and machine learning on one side and by process modeling and analysis on the other [1]. The discipline has gained some traction by different industries over the last years. One of the reasons for that is that a number of process mining tools have rolled out that are specifically built for industry use. With their simple and intuitive user interfaces, they have made the process of conducting process mining much simpler. Some examples of these tools are Fluxicon Disco and Celonis PI. Compared to more academic processing tools like ProM these tools have limited, but powerful, techniques, are much simpler and scale better for larger datasets. Using these tools does not require a lot of background knowledge. Thanks to these tools and the maturing of the discipline over the last years more and more companies have started using process mining to complement their business intelligence tools. However, process mining is still quite far from being widely adopted. Even though the analytical techniques provided by process mining are developing fast and gaining more traction the knowledge about the benefits process mining provides and how to apply the techniques seems to remain low in the industry.

So far there has been no research done to understand how useful does the industry consider the current process mining techniques. The goal of this paper is to describe how the applicable does the industry find process mining techniques. This was done in two parts.

In the first part, a systematic literature review was conducted that identifies the state of the art techniques of process mining. The aim was to find the techniques that have been tested on real-life logs. This constraint was necessary to understand if the technique is mature

enough to be used by the industry. As this meant looking into what kind of event logs have been used by process mining researchers this presented an opportunity to understand what kind of data is being used to conduct the research. Furthermore, it revealed where the real-life data is coming from and what industries are driving the research of each process mining? The aim of the research is to identify the techniques that the industry might be interested in.

In the second part, the techniques identified in the first part were used to conduct a survey among industry representatives. The aim of the survey was to understand how industry representatives value the techniques that researchers are working on. This survey was held among industry professionals in Estonia and in different business process management groups in LinkedIn. Meaning that there were two main research groups. The Estonian industry representatives and the worldwide industry representatives. This provided the opportunity to compare how Estonian industry representatives value process mining versus industry representatives from around the world. Estonia is the main focus of this research because the Software Engineering Group in Tartu University has been researching and introducing the field of process mining for some years now. This survey gives insight on how successful they have been in introducing the technology in Estonia.

This paper will give insight about the current applicability of process mining techniques in the industry based on industry feedback to process mining. It will identify different process mining techniques and the industries driving the research of the techniques. Furthermore,we will present the results and analysis of a survey on the attitudes of industry representatives towards each identified process mining technique.

The remainder of the paper is organized in the following way. Section 2 describes the methodology of the of the systematic literature review while section 3 gives an overview of its results. The survey methodology is described in section 4. Section 5 discusses the methodology that was used to analyse the results of the survey. Section 6 describes the results of the survey based on each technique. It also summarizes and discusses the insights the survey provided. Section 7 concludes the paper.

2 Literature Review Methodology

In order to understand what uses of process mining have been applied on event logs from the industry, a systematic literature review is conducted. To conduct a systematic literature review a research protocol has to be in place. The protocol was created using the guidelines specified by Kitchenham [2]. This section describes the protocol this paper follows.

Section 2.1 describes the research questions. Section 2.2 describes the derived search string and section 2.3 the data sources that were selected. The inclusion and exclusion criteria are listed in section 2.4. The studies selected are described in section 2.5 and section 2.6 describes what data is extracted from these works.

2.1 Research Question Formulation

The aim of this systematic literature survey is to identify the different process mining techniques that have been applied on real life event logs. This is done through identifying and analysing studies on process mining that have used real life event logs in their work. For this a set of research questions were derived:

1. What are the different process mining techniques that have been applied in the real world?
2. What industry does the data, which is being used to conduct research, come from?

2.2 Search String

To search databases a search string needed to be created. The aim of the survey is to find all process mining techniques. Therefore the search string had to include a general term to capture all the usages of process mining. The search term “process mining” was selected. To capture that this paper is only interested in works have been applied on real-life event logs different synonyms for real-life event logs were included as search terms:

- “real-life”
- “industry log”
- “real-world”
- “case study”

This was all combined. The search string was the following:

"process mining" AND ("real-life" OR "industry log" OR "real-world" OR "case study")

2.3 Data Source Selection

The initial idea was to use Google Scholar, Web of Science and Scopus libraries. These sources contain the most papers on process mining. However, after some initial research Google Scholar was discarded. Google Scholar was discarded because it gives a lot of results and has a limit on results that can be accessed. This means some relevant results can be missed because they are over the limit. When applying the defined search query to Google Scholar it listed over 7800 results making it futile to proceed using it.

Web of Science and Scopus gave relevant works and were selected as the data sources for this survey

2.4 Inclusion and Exclusion Criteria

Inclusion criteria:

1. The study is about a process mining technique.
2. The study is verified on a real-life log.
3. The results of the experiment with the real-log are described in the paper.
4. It is the latest paper on the subject by the same author.

Exclusion criteria:

1. The paper is not in English.
2. The paper uses only synthetic and/or artificial logs.
3. The paper uses real-life logs, where the core characteristics of an event log are artificially changed to fit needs of the research project.
4. Paper is not accessible for free through university library proxy service or found in the top 20 search result on Google when searching the paper by the title.

Exclusion criteria number 2 was included because approaches tested on only synthetic logs might not be applicable in the industry. Exclusion criteria 3 was selected because it can be disputed if the logs are still real-life logs if they have been altered to fit some research goal. As an example of a paper that was excluded based on this criteria is a work about capturing concept drift in process mining by a group of Indian researchers [3]. They came up with a process mining algorithm to detect concept drift. Unfortunately, they did not have real-life event log with concept drift to test the algorithm on. So they artificially induced concept drift in a real-life event log that did not actually contain it. Exclusion criteria 4 is defined, because some papers were not accessible through the university network, but could easily be found by googling the title. The top 20 search result limit is set because if the paper is not in the top results the likeliness of finding it among the other results is low. Even more, it would have been too time-consuming.

If any of the exclusion criteria is met the paper is discarded.

2.5 Study Selection

The search string was applied to Scopus and Web of Science. This gave 729 results. 455 from Scopus and 274 from Web of Science. The initial results were reviewed in 3 iterations and papers that did not match the review criteria were excluded. **Error! Reference source not found.** illustrates the number of papers left after each iteration.

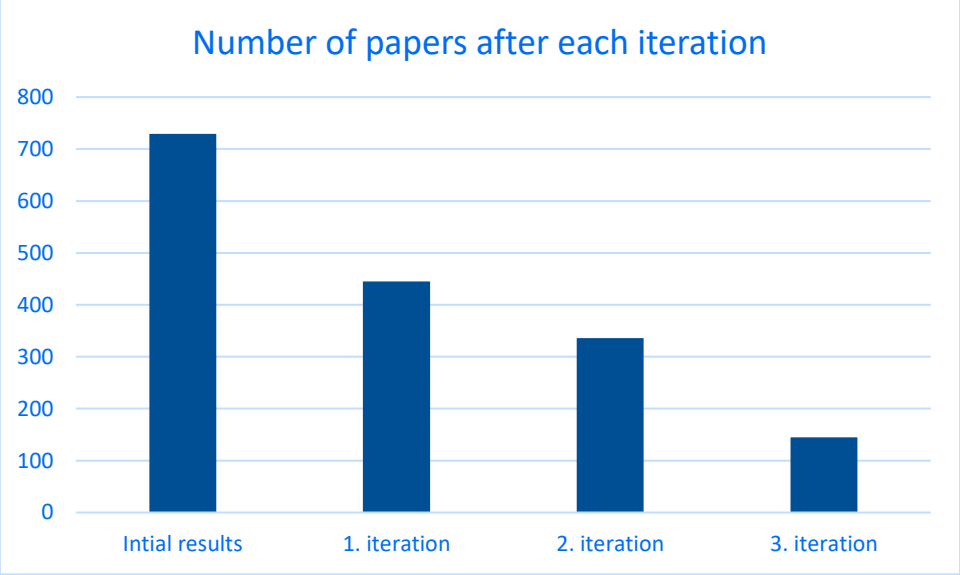

Figure 1 Number of papers after each iteration

In the first iteration, the results of the two databases were compared based on title and author of the papers and duplicates were removed. Scopus contained some conference book proceedings. These were also discarded in this iteration. This iteration left us with 422 papers from Scopus and 23 papers from Web of Science.

In the second iteration, the papers were filtered by title and abstract using the inclusion and exclusion criteria defined in section 2.4. When necessary the introduction section of the paper was read for clarification. Some papers were excluded, because they were related to process mining, but out of scope. For example a paper from Austrian researchers about the way to store event logs in an information system [4]. While this paper is related to process mining, it is about the way to gather and pre-process event logs. Pre-processing is not a process mining technique. Therefore this paper was excluded from this systematic literature review. A large portion of the papers that were excluded in this iteration because it became clear that these papers had not tested their approach on real-life logs. An example to illustrate this is a paper from the Netherlands about electronic data interchange [5]. The abstract of the paper contained phrases "real-world" and "process mining", but the "real world" was referring to real world processes. The paper itself did not contain any studies or experiments using real-life logs. In the end of this iteration 336 papers were left.

During the second iteration the papers were categorized based on the subfield of process mining they belong to. The vast majority of 205 papers were categorized to be on process discovery. This was too be expected as process discovery is the core technology that all other process mining techniques rely on and the bulk of the research done about process mining has been done on process discovery. 113 papers were categorized as other process mining techniques. 18 papers were categorized as process mining, because they contained case studies using many different process mining techniques. These papers were included in the systematic literature review.

A systematic literature review on process discovery has already been conducted in Tartu University by Soo [6]. This review was done to identify different process discovery techniques. In his work he has selected the state of the art papers on discovery that use artificial, synthetic and real-life logs. As Soo's literature survey already describes all the different process discovery techniques out there and is conducted during the same time as this review it makes sense to reuse his work. So in order to understand what had been done in the field of process discovery with real-life logs Soo's literature survey on process discovery techniques was reviewed and the list of papers he included in his work was filtered using the inclusion and exclusion criteria defined in this paper. Through this, only process discovery papers with real-life logs were selected for this systematic literature review. Through this the number of process discovery papers was cut down from 205 to 45. This filtering was included in the third iteration

In the third iteration, the papers were read and the data was extracted based on the data extraction strategy defined in section 2.6. During this iteration, more papers were discarded. In the end of the 3rd iteration, 143 papers were left.

In the final literature review, only the most interesting or illustrative examples out the 145 papers were included to demonstrate the techniques of process mining.

2.6 Data Extraction Strategy

The following information was extracted from each paper:

1. General information. The title of the paper, its authors and the year of publishing.

2. The process mining technique the paper is on. For instace,is on process discovery, performance analysis, conformance analysis or something else.
3. The industry the event logs are from.
4. Process mining tools used.
5. Event log characteristics. The number of events, activities, traces and cases each event log contained.
6. What kind of algorithms were used in the paper? This means the process mining algorithm being used. For instace,alpha algorithm, heuristic miner, fuzzy miner and so on.
7. What was researched? Meaning what was the main contribution of the paper. Was it a discovery algorithm, process mining project methodology, a case study about applying process mining?
8. Where were the real-life logs applied? This question will help to understand if the real-life logs were applied in an evaluation of an algorithm or a case study on applying process mining in a new setting.

3 Process Mining Use Cases

In this section the results of the systematic literature review are discussed.

Figure 2 shows the papers that were included in the survey by the publishing year. It becomes visible that in the last years the number of papers that have used real-life logs has increased year by year only having a slight downfall in 2014. As this literature review was conducted in early 2017 there is only one paper from 2017. However, it is clear that process mining is a growing area of research as there is an upward trend in the number of papers published each year.

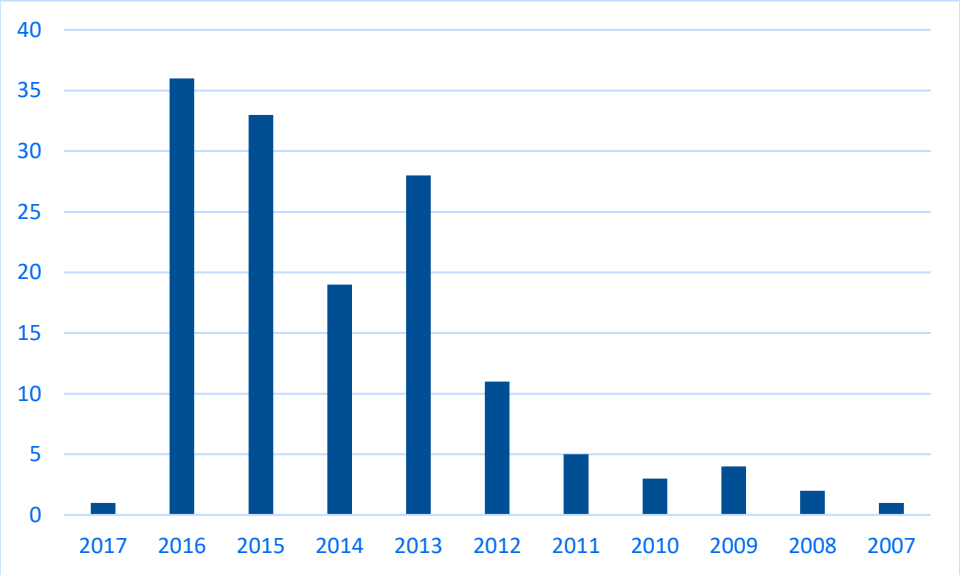

Figure 2 Papers of the 3rd iteration categorized by year

Figure 3 illustrates the results by categorization of techniques. There are 45 papers on discovery, which is adequate considering that most of the research on process mining is about process discovery. The next most popular categories are performance, conformance checking and prediction. This was to be expected. One of the main reasons for using process mining is the ability to analyse the performance of one's processes. Conformance checking is one of the most researched process mining techniques. It is only second to process discovery. Process prediction is gaining more interest from the research community as more and more advances are being made in the field of machine learning and data science. The techniques with least papers are process repair, concept drift and process comparison. These are quite new fields of research so it is not surprising there are only a few papers on these topics.

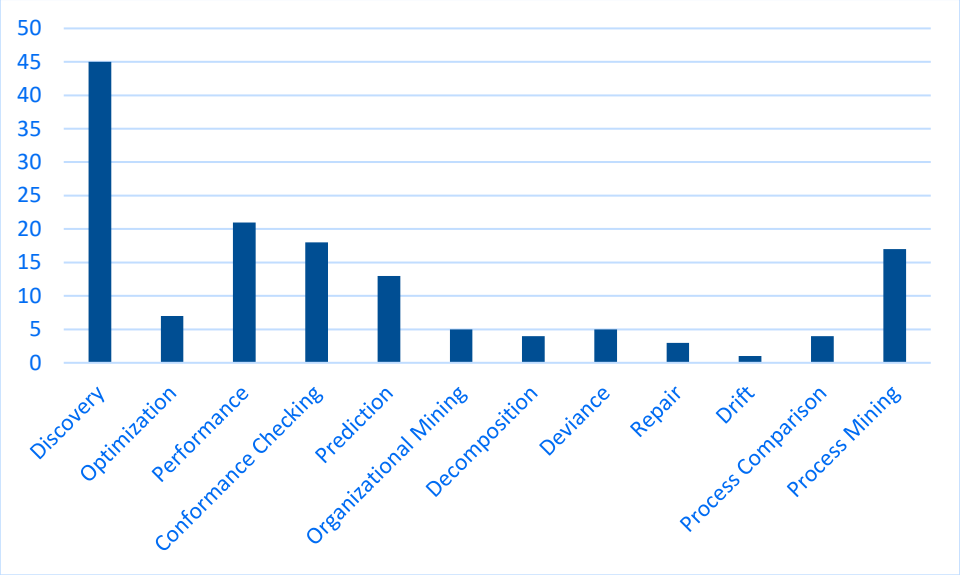

Figure 3 Categorization of the 3rd iteration by techniques

In total, these papers contained 185 real-life logs. Figure 4 demonstrates the industries the event logs are from. It is visible that the most popular logs are about finance, information technology, the public sector and healthcare. One might conclude that these are the key industries that are driving process mining. However, this might not be the case. In the case of the top three industries we are mostly dealing with 3 logs that are used for research over and over again. These three logs from the Business Process Intelligence Challenge of 2011, 2012 and 2013. These logs represent healthcare, finance and information technology processes. They are publicly available and as we can see from the results used widely by researchers. The fourth biggest industry is the public sector. For this sector there is a log from a Dutch public sector project called CoSeLoG. This log makes up almost half of the times a public sector log has been used in research. Figure 5 illustrates how many times BPIC and CoSeLoG logs have been used when compared to other logs in the same industry. We can see that in most cases they make up almost half of the logs in their industries. Even more than half in the case of finance.

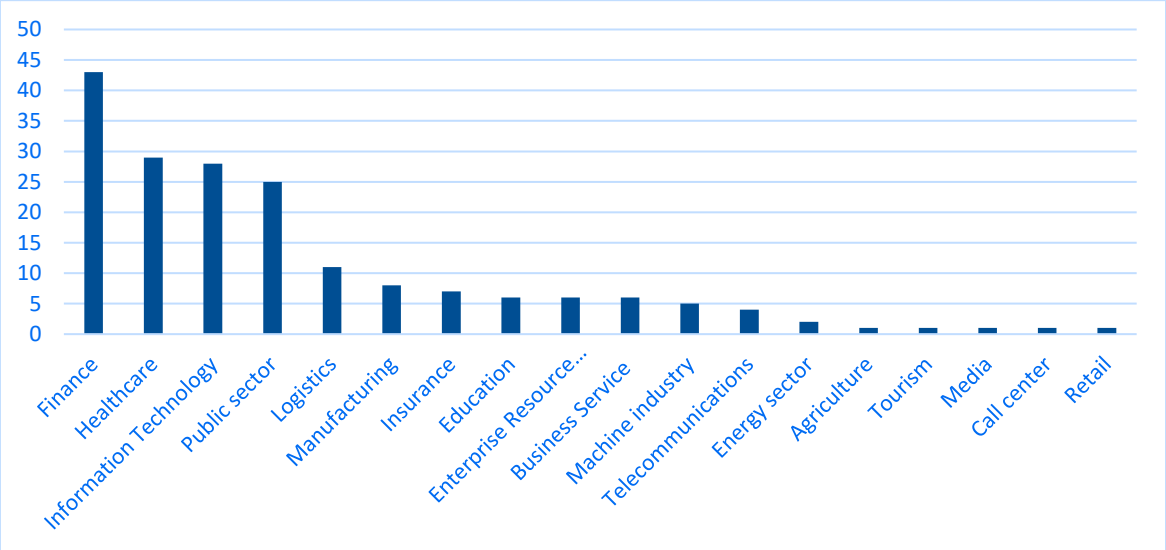

Figure 4 Logs that are used in research papers categorized by industries

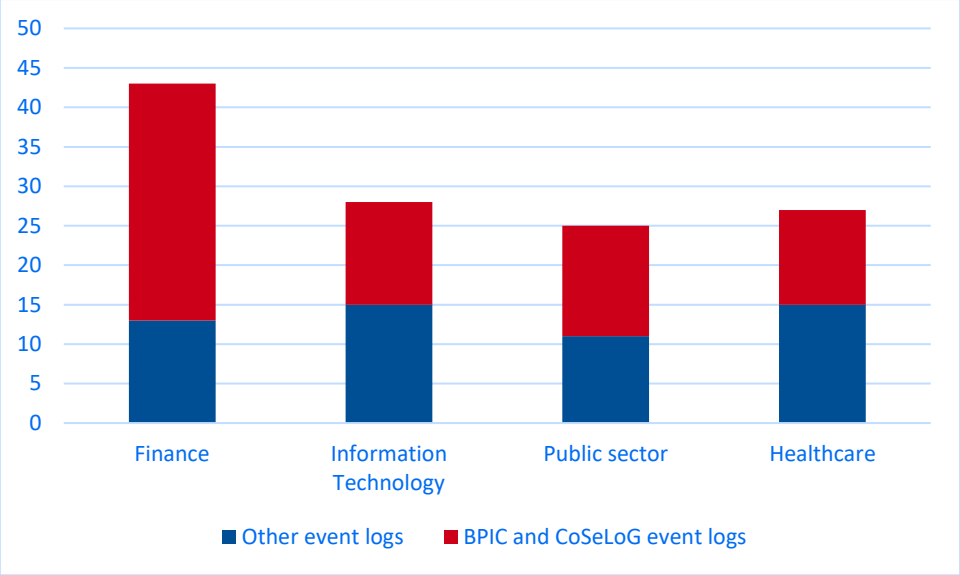

Figure 5 The proportion of BPIC and CoSeLog event logs in their respective categories

Eleven process mining techniques were identified based on the papers read. This section introduces each process mining technique and describes some examples how each process mining technique has been applied in the industry. Appendix I contains list of papers that where included in the survey.

This section is ordered in the following way.

- Section 3.1 describes process discovery. It is used to construct process models from event logs. This is the main technique of process mining that all other techniques rely upon.
- Section 3.2 is about performance analysis. This technique is used to analyse the business processes and find bottlenecks in a process that can be optimized to make the process more efficient.
- Section 3.3 talks about optimization. This technique is used to make processes more efficient.
- Section 3.4 is conformance checking. This process mining technique is used to understand if businesses processes are following the business rules that are set for these processes.
- Section 3.5 describes prediction, which is used to predict the possible outcomes of a business process, for instance, delay time or process cost.
- Section 3.6 is organizational mining. This technique deals with understanding the organizational structure of a company, the handover of work between workers and the resource utilization.
- Section 3.7 shows decomposition. The aim of this technique is to decompose large complicated process models into smaller and easily understandable models.
- Section 3.8 talks about model repair. This technique uses an event log and a process model and creates an updated model of a process. It is used when there is a need to capture some data from the model that does not exist in the event log.
- Section 3.9 is process deviance. The aim of this technique monitor business processes and detect deviances from the predefined business process.

- Section 3.10 shows concept drift. The idea behind this technology is that business some processes change over time without the stakeholders even noticing. Concept drift detects changes in the business process and notifies the stakeholder of the changes that occurred.
- Section 3.11 talks about process comparison. This technique is used to compare two executions of the same business process and find differences between these processes.

3.1 Discovery

Discovery is the first step of any process mining effort. The technique takes the event log of a business process execution as input and produces a model representing the current business process. The model is usually shown using Business Process Model and Notation (BPMN). All the activities that have occurred during the execution of a business process are shown in the model. Process discovery is mostly used to extract the as-is process model, but this is not the only reason. The majority of process mining techniques that will be introduced in the coming chapters like performance analysis, optimization and conformance checking among others, need the discovered model to perform their analysis. This is why discovery is the most researched topic in the field of process mining. Producing an exact model of the as-is process will make further analysis more precise.

Process discovery has been applied to many different industries. Figure 6 shows the different industries discovery has been applied in and how many times logs from each industry have been used in the research of process discovery.

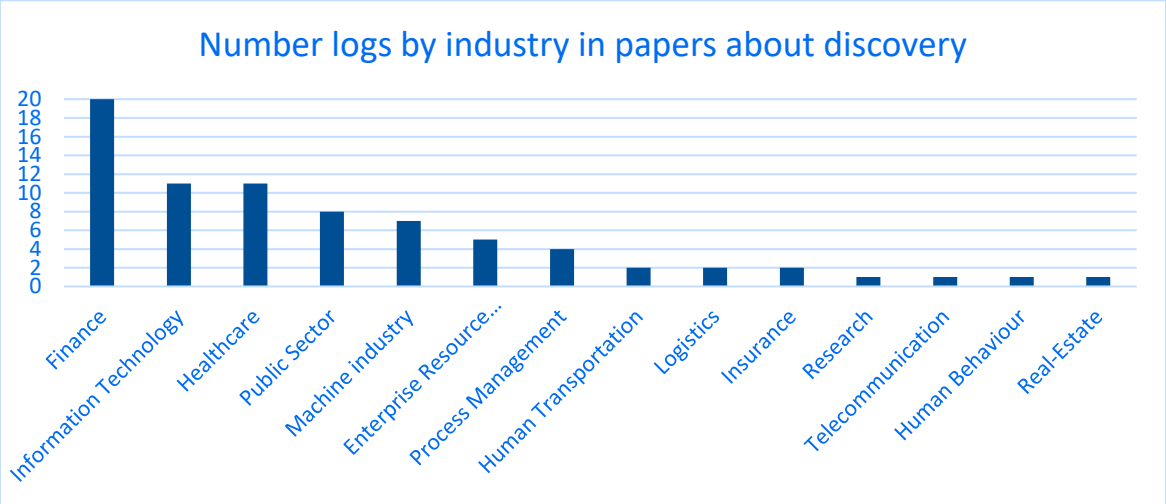

Figure 6 Number of event logs usages by industry

As shown on Figure 6 discovery has been applied mostly on logs from the financial industry. One example of this comes from a group of Austrian-Estonian researchers [7]. They took the event log of a Dutch financial institution and applied their discovery approach on the log. As a result, they produced a set of 18 human readable constraints that describe how the process works.

In another example, discovery was used on a Belgian research funding agency [8]. The process was handling research funding applications. This process consisted of two sub processes. The first was about handling the reviews given to the applications. The second was about giving out funding in two payments. The proposed discovery approach managed to produce a process model with advanced BPMN constructs like sub processes.

3.2 Performance Analysis

Performance analysis is focused on analysing the performance dimensions of a business process. The main dimensions are time, cost, quality and flexibility [9]. Time represents the time spent on the execution of the process. Cost represents the financial expenses of executing the process. Quality relates to how well the process meets the requirements of the customer and flexibility shows the adjustability of the process to a changed environment.

Performance analysis can be conducted in all kinds of processes. For instance, in Chile, there was a case study that compared the use of the learning management system by high and low performing students [10]. The approach used performance analysis to highlight the differences between two groups.

In Korea, process mining was used to conduct human reliability analysis on the control room crews in nuclear power plants [11]. The control room crews receive simulation training of possible failures in the power plant. During one of these trainings, the communication of crew members was recorded. The researchers categorized the most relevant keywords, then used natural language processing and turned it into an event log. Performance analysis techniques were applied to analyse the flow of performing the steps to contain the situation. The researchers concluded that process mining techniques showed great promise in this field.

3.3 Optimization

After conducting performance analysis the bottlenecks and weak points of the process are identified. Knowing these critical places in the process flow is necessary to optimize the process. Optimization means changing the process flow to optimize it for a chosen metric, while not changing the process goal. As mentioned in the previous section the relevant metrics of a process are usually time, cost, quality and flexibility. These metrics are related and while changing the process to optimize one metric other may increase.

An example of an optimization effort is a case study on a cancer institute in the USA [12]. The group of researchers behind the case study set out to answer the question: “How to improve operational performance of the scheduled process?” The researchers came up with a process improvement technique. The algorithm changed the scheduling policy of the cancer clinic based on the identified issues in the process. The experiments resulted in a 20–40% improvement, in process flow and decrease of lateness.

3.4 Conformance Checking

Conformance checking is along with discovery are the two most researched topics in the field of process mining. Conformance checking deals with checking the similarity between a process model and the event log. This approach can show how similar is the actual behaviour in the process to the behaviour perceived by the organization. Having low conformance metrics means that the process models are out of date and the actual processes are conducted in a different way. To update the models process repair is used.

Conformance checking techniques can be split into to two depending on the models they use. Procedural and declarative. The two models have been created because the processes they support have different characteristics. Procedural process models are used mostly for processes where the sequence of events is clear, meaning all the possible next events are specifically stated. Declarative processes take the opposite view. If a next event is not explicitly forbidden it is allowed. Declarative models are used on processes where the sequence of events might not be that apparent. For instance, healthcare is one of the main drivers for using declarative models.

As mentioned there have been a lot of papers about conformance checking. This includes a lot of real life case studies in different industries. Conformance checking has been applied to analyse the process of issuing building permits in a municipality, procurement processes in companies, different healthcare and financial processes among other fields [13,14,15].

One interesting example that illustrates the usefulness of conformance checking comes from the University of Leuven in Belgium. The researchers came up with an advanced rule-based conformance checking approach to check business process compliance to a set of rules [14]. They evaluated their approach by conducting a study on using conformance checking to assess the effectiveness of a purchase-to-pay process in a large non-profit organization. In the purchase-to-pay process the researchers first visually identified a set of deviations from the standard process, which may cause financial loss to the company and therefore can be considered as risks. They focused their research on two key problems. Firstly, making sure no fraudulent bills get registered and paid for. Secondly, ensuring all employees have suitable responsibilities in the information systems based on their role. This was important to reduce the risk of fraud in the process. The researchers then proceeded to formulate a set of logical rules that could be applied to the compliance checking process to avoid such deviations in the process and thereby mitigate the risks the company might face. The researchers concluded that the non-profit organization had managed set proper responsibilities to employees, but had not managed to make sure fraudulent bills would not get registered.

3.5 Prediction

Prediction deals with predicting process instance outcomes based on historical event log data of similar cases. It is a mix of process and data mining techniques. Different metrics can be predicted. For instance, processing time, errors and delays. Prediction techniques have been applied in many different sectors.

Quite a few works have used event logs from healthcare processes to evaluate their approach for business process prediction [16,17,18,19]. One the most interesting ones was conducted in 2013 by researchers at the University of Twente in the Netherlands, who conducted a case study on predicting the cash flow of a hospital [19]. This research was conducted, because of a change in the Dutch healthcare system. If previously the healthcare provider could bill the cost of treatment after the diagnosis was made then after the change the billing could only start after the medical provider had finished giving care and the cost of actual care activities became clear. This meant that the hospitals would have a financial management challenges. When previously they would get paid before starting treatment then after the change they would have to finance the care out of their own pocket until the treatment is finished. This meant that predicting the cost of treatments would help healthcare providers be more efficient at managing their finances. The researchers used several data mining techniques to build a model that predicted the cost of the treatment based on the diagnosis and past care given. The model could predict the next care product in the care activity sequence with the accuracy of a little less than 50%, which is better than picking the care product at random, but not as accurate as needed to really use the model. However, they were much more successful at predicting the next care product in the care sequence reaching 80% accuracy with their best algorithm. The overall result of the research was pretty accurate. Out of the 3 datasets, they used for research for 2 they managed to obtain an error of less than 10% and 17% for the third. These numbers show that using prediction is a viable option for predicting cost to healthcare providers.

3.6 Organizational Mining

Organizational mining focuses on discovering and analysing organizational structures and communication between organizational entities. It is divided into three: organizational model mining, social network analysis and information flows between organizational entities [20]. Organizational mining uses event logs to create the organizational model. Social network analysis deals with analysing communications between performers using specific metrics. The example of one of these is the handover of work metric, which shows how a work case is passed along in the organization. Information flows between organizations shows social networks where the performers are in different organizations. Organizational mining has been applied to many different fields.

A case study was performed in 2008 at a Dutch municipality [20]. The event log is from an invoice handling process from the municipalities' Urban Management Service. Organizational model mining and social network analysis was used to analyse relationships between process performers and departments. The result was an organizational structure model and a social network model that showed the connections between the local municipality and other local public institutions.

Another example is using social network analysis is in the healthcare field. A group of Brazilian researchers used organizational and social network analysis to conduct analysis on a Brazilian hospital [21]. The event log was about a chemotherapy treatment. The analysis managed to map all interactions and identify the departments with the highest number of interactions. The work pointed out that the departments with the highest interactions are more susceptible to errors and need to be monitored more.

3.7 Decomposition

Event logs can be very long and complex. Complex problems can be made simpler by decomposing them into smaller, more manageable problems. This is what decomposition does. It decomposes the complex event log into smaller sets. Decomposition is mostly used for discovery and conformance checking to make the process models more understandable and easier to follow.

A real-life example was made by a group of Spanish-Dutch researchers [22]. They used decomposition to make conformance checking faster. The approach was tested on a real-life log from the Dutch financial institute. The log had roughly 29.6 events per case. The results were astounding. The use of decomposition made conformance checking 99.99% faster than not using it. Such reduction in calculation time shows the benefit of using decomposition.

3.8 Model Repair

Process model repair fits between process discovery and conformance checking. When discovery finds the process model and conformance checking checks it's compliance to the expected model, process repair deals with improving the discovered model based on the event log. As input this technique takes the process model and an event log. The aim is to create a model that allows all the observed behaviour while being as close to the original model as it possibly can [23].

When do use discovery and when to use model repair? Repair is usually used when the model can provide additional information that might not be in the event log. In this case the given model serves as a source of information to discover the updated model.

A case study that illustrates the use of model repair was conducted in a Dutch hospital by a group of researchers [24]. The event log is about treating a certain illness in the urology

department. The case study first starts with designing a so called *de jure* model of how the process should like based on best practices. Conformance checking was used to find discrepancies between the log and the *de jure* model. Next, the model was repaired using the *de jure* model and the event logs. All the discrepancies found during conformance checking were taken into account to produce the *de facto* model. The result is an updated process model that has the correct constraints and is in tune with reality.

3.9 Deviance

In companies there usually are predefined process models that prescribe how to execute a business process. However, the process execution does not always follow the process model. Deviations from the model occur. Process deviations can be split into three categories "explicit exceptions", "implicit exceptions" and "anomalies" [25]. Explicit exceptions are approved guides when and how to deviate from the regular processes. Implicit exceptions are deviations that are accepted by the company, but there are no guidelines how to deal with them. Anomalies are unexpected deviations that can refer to errors, mistakes or even fraud.

There have been some works that use real-life logs to show the value from using process deviance. In a 2012 paper, a group of Belgian researchers tried out their approach for detecting deviance on a large financial institutions procurement event log [25]. They found anomalies in the procurement process that the implied the weakness of the internal controls in the company. Another group of researchers in Romania applied deviance mining on a software automotive company [26]. They discovered major inefficiencies in the software engineering process. For instance, in some cases the client of the software team skipped the step of analysing the request sent to the engineering team. This caused the engineering team to build and ship a solution that did not fit the needs of the client. The result of this was that the engineers had to do a lot of rework.

3.10 Concept Drift

Process mining techniques tend to assume that processes are always in a steady state [27]. In real life processes evolve and change over time. A business process captured now might not be same as a year from now. This is where process drift comes to play. Concept drift refers to the situation in which the process is changing while being analysed. The three main problems when dealing with process mining are change point detection, change localization and characterization and change process discovery [27]. Change point detections means realizing that a change has actually taken place in a process. Change localization and characterization means the region of the change is found the type of the change is identified. Change process discovery means putting the change into context. Showing what caused the change and how it evolved.

A real-life case study applying concept drift techniques was done by researcher in the Netherlands. The process they took under observation was process of applying for advertising permits in a Dutch municipality. If a person wants to post an advertisement on a building they have to apply for a permit in the local municipality. A number of changes in the process model was detected. Two examples are made. The first change was with making the completeness check of the registered documents from optional to mandatory. The second change was that a new activity had been added to the process. They conclude that the framework they proposed shows significant promise in detecting concept drift from event logs [27].

3.11 Comparison

Consider an organization that has many branch offices that are geographically apart. The execution of the same business process can be quite different in these branches. Process comparison is used to compare variants of the same process using event logs and find statistically significant changes.

Process comparison has been applied by Dutch researchers on an event log from an Italian Municipality's IT system that deals with "road fine management" [28]. They split the event log into two one contained fines below 50 euros and the other over 50 euros. They found that low fines are paid much faster. Also, high fines are sent to the offenders earlier. Another interesting find was that it is significantly more frequent that high fine payments occur when a financial penalty is added for missing the payment or transferring an incomplete sum. Other examples of applying process comparison on event logs include a case study dealing with the process of stroke management and a case study comparing students performances based on watching video lectures [29,30].

4 Survey Methodology

In the literature review, 11 process mining techniques were identified. These techniques serve as the basis of the survey performed among representatives of the industry.

The aim of the survey was to understand how industry representatives value the techniques that researchers are working on.

The survey was constructed in the following way. Respondents were asked some general information about their backgrounds like the industry they work in, their role in the organization and their experience of working with processes. After that, each process mining was introduced and their usage explained. Each description was followed by 3 questions:

1. How familiar are you with the presented technique?
2. How useful is the presented technique from your company's perspective?
3. What could be the possible application of the presented technique in your company? Please describe.

The questions were constructed so we could understand how well to the respondents feel they know the technology, would this technique be useful to their company and how do they think it could be applied in their company.

The first two questions were mandatory. The respondents were asked to assign a rank from 1-7. One being the lowest and seven the highest value. The third question was open ended and optional.

The target respondents were industry representatives who have experience in managing or improving processes. Initially, two groups of interest were defined.

1. Quality control group - Industry representatives, who have no deeper knowledge of the capabilities of process mining.
2. Industry representatives, who have been introduced to the capabilities of process mining.

The idea behind questioning these two groups was to understand if there is any difference in rating the usefulness of the techniques based on their familiarity.

The initial scope of the research was Estonia. The reason for choosing Estonia is that the Software Engineering Group in Tartu University has been researching and introducing the field of process mining for some years now. This survey gives insight on how successful they have been in introducing the technology in Estonia.

To extend on this. One more focus group was added to the survey. Industry representatives worldwide. This gives us the opportunity to compare the Estonian results to how process mining is viewed by industry specialists around the world.

So at the end, the respondents were divided into three subgroups:

1. Quality control group. Industry representatives, who have no deeper knowledge of the capabilities of process mining.
2. Estonian industry representatives, who have been introduced to the capabilities of process mining.
3. Worldwide industry representatives, who have been introduced to the capabilities of process mining.

The survey was done online using Google Forms. The total number of respondents was 92 who were divided into subgroups. Subgroup sizes are indicated in Figure 7.

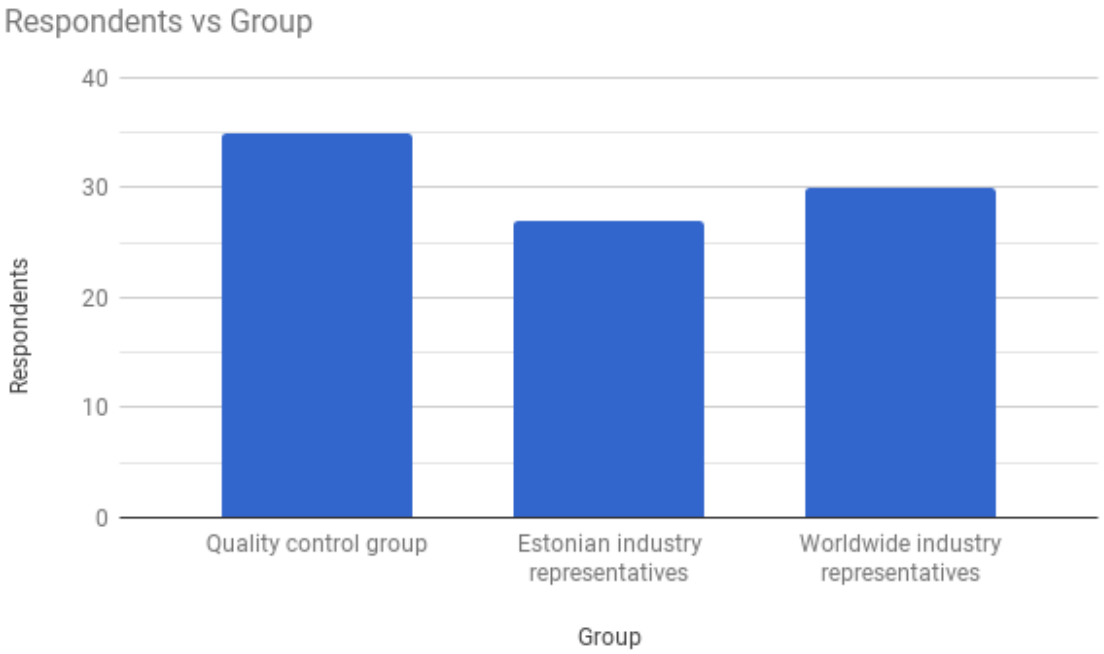

Figure 7 Respondents divided into subgroups

The following sections describe how each group was targeted and approached.

4.1 Quality Control Subgroup

The quality control group is made up of industry representatives who are familiar with the concept of process mining, but have no deeper knowledge how to apply it. The aim of surveying this group is to understand the general sentiment towards process mining as a technology by a group who has never used it.

For the quality control group, Tartu University's Open University program students taking the courses on Business Process Management and Business Analysis were chosen. The majority of these students are industry professionals in their respective fields. During their studies, they have been introduced to the concept of process mining, but they have not seen it in action.

There were 35 respondents in this group. The group was approached during their lecture and asked to fill out the questionnaire. This group is made up of Estonians so the questionnaire was in Estonian. Figures Figure 8 and Figure 9 give some insight about the industries that the respondents are from and the how long they have worked with processes

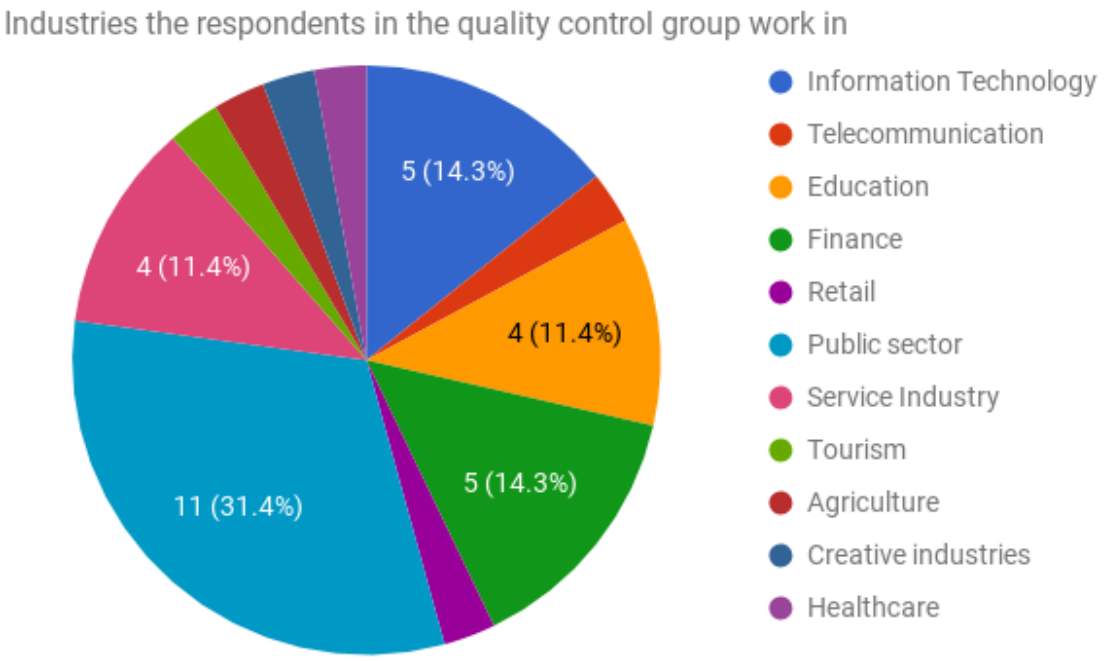

Figure 8 Industries the quality control subgroup respondents work in

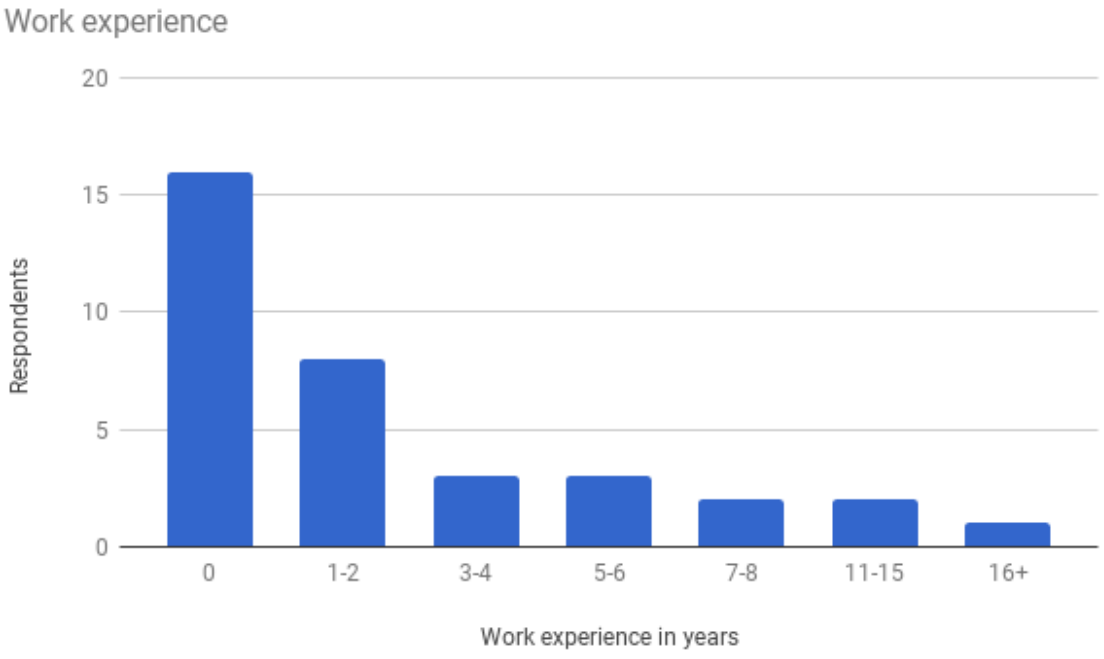

Figure 9 Respondents grouped by work experience with processes in years

Almost a third of the respondents in this group work in the public sector. The next biggest categories are information technology, finance, education and the service industry. 16 people out of 36 say that they have no experience of working with processes.

4.2 Estonian Industry Representatives Subgroup

The Estonian industry representatives group is made of industry representatives who have had a deeper introduction to process mining and its capabilities. This group is made up of members of the Estonian Business Process Management Roundtable (BPM Roundtable). This is a group of industry representatives who are interested business process management. They have had a couple of lectures where they have been demonstrated the capabilities of process mining by academic experts in the field like Wil van der Aalst and Anne Rozinat. Some of the members of this group have conducted process mining pilot projects in their own companies. Therefore they are more familiar with the field of process mining than the quality control group and the ratings from this group should differ from the quality control group.

The Estonian BPM roundtable has their own mailing list the Estonian version of this questionnaire sent to this mailing list.

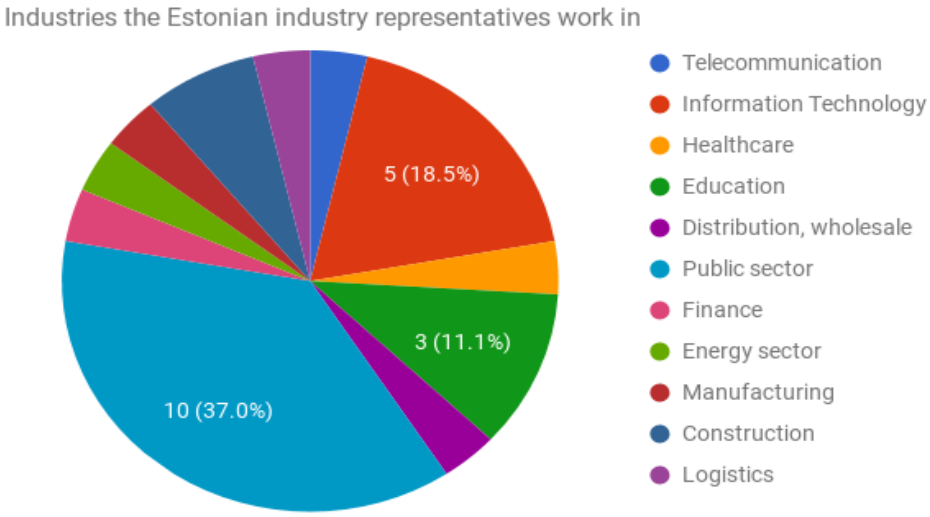

Figure 10 Industries the respondents in the Estonian subgroup work in

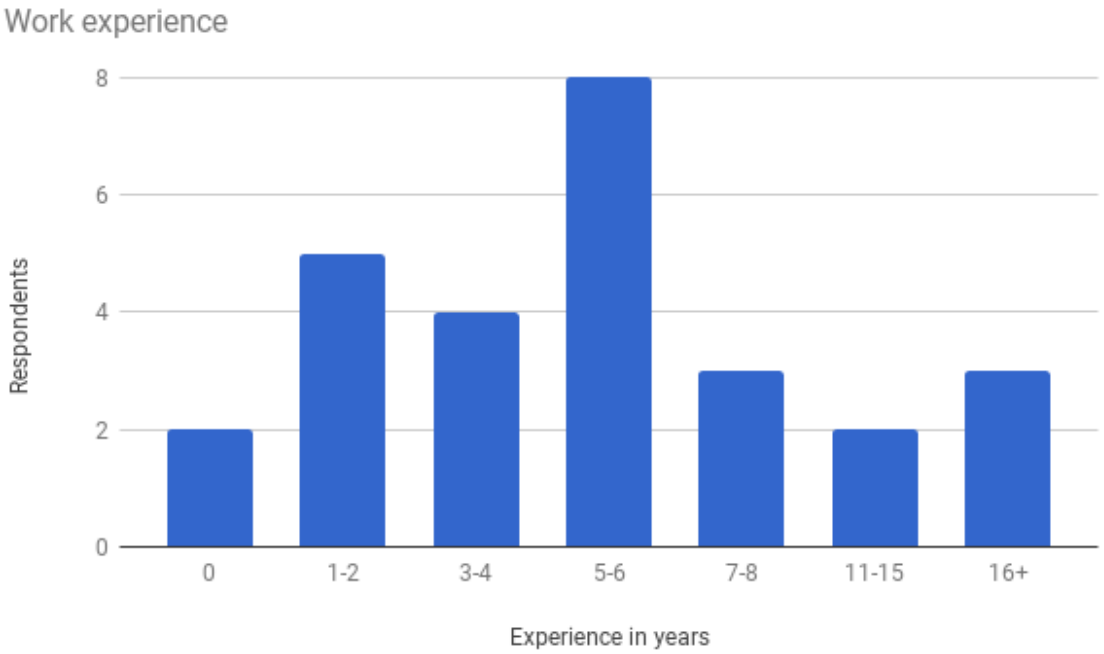

Figure 11 Respondents of the Estonian subgroup grouped by work experience with processes in years.

This group also contains a lot of representatives from the public sector, over one-third of the group, followed by information technology, education and construction. The work experience of these groups is much more evenly distributed than in the previous group with most people having 5 to 6 years of experience in working with processes.

4.3 Worldwide Industry Representatives Subgroup

The worldwide industry representative group is made up of members of 5 different Business Process Management groups on LinkedIn. This target group was defined in order to understand if there is a difference between the Estonian industry representatives and industry representatives globally. An English version of the questionnaire was posted to these groups. 30 responded worldwide. Figure 12 shows the countries where the responses came. The most responses came from the United States followed by Canada, India and Australia.

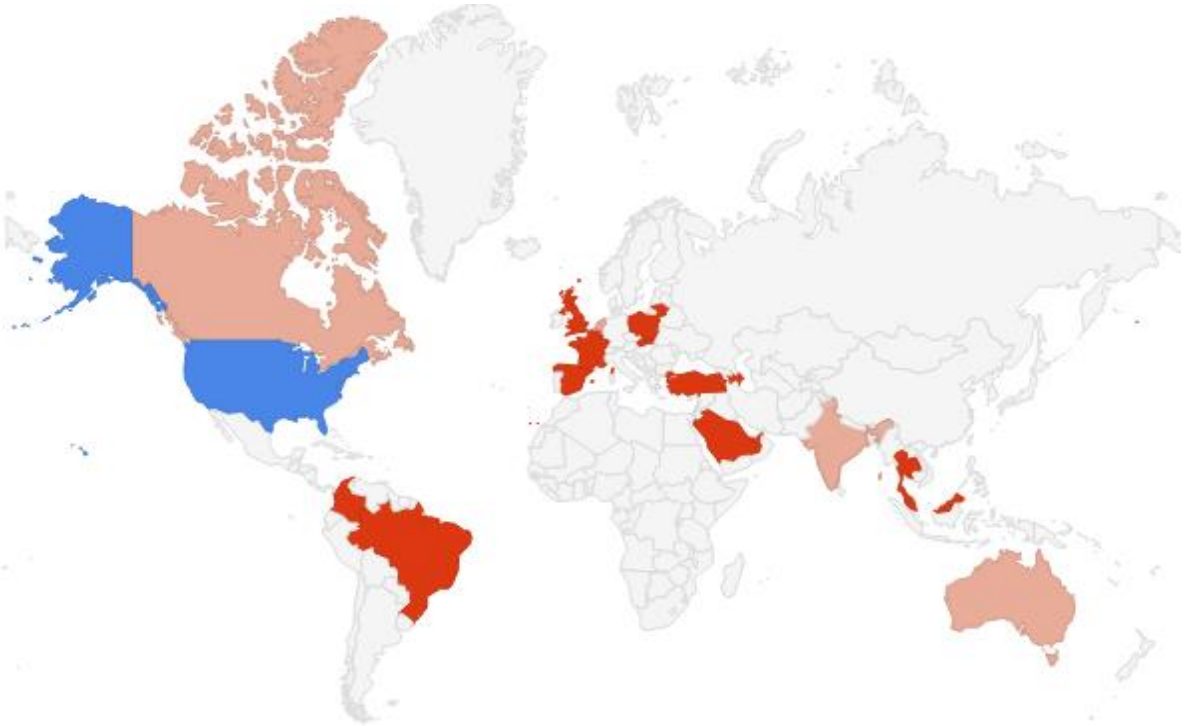

Figure 12 Countries where the worldwide the respondents are from

As Figure 13 shows the most represented industries are information technology, logistics, manufacturing, retail and education. Almost half of the respondents have more than 10 years of experience with working with processes.

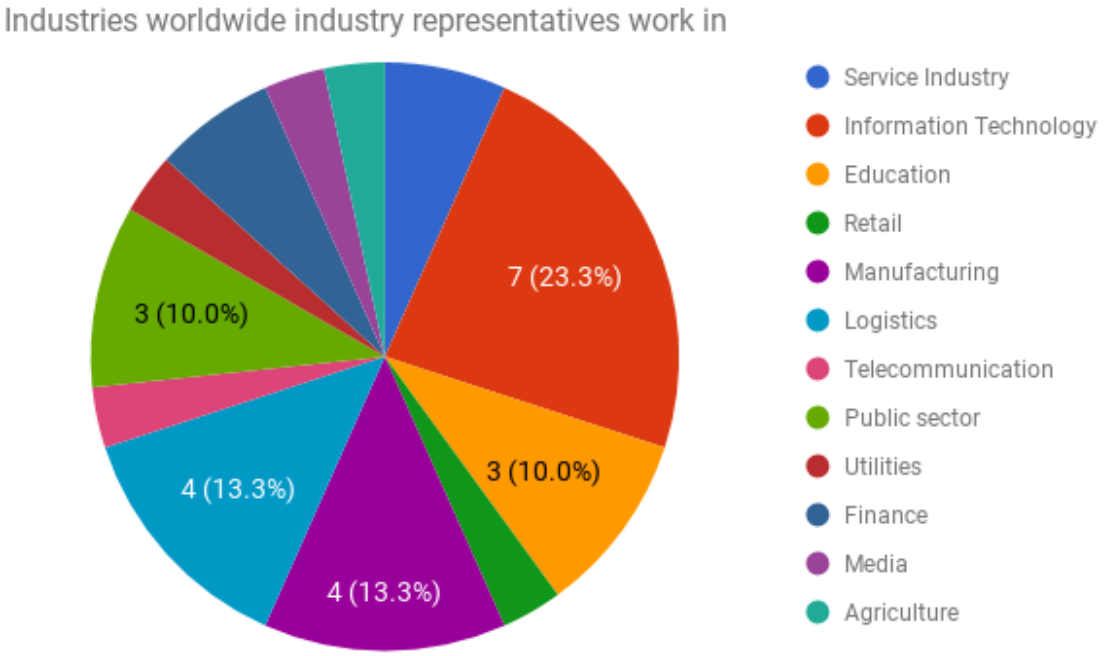

Figure 13 Industries the respondents in the worldwide subgroup work in

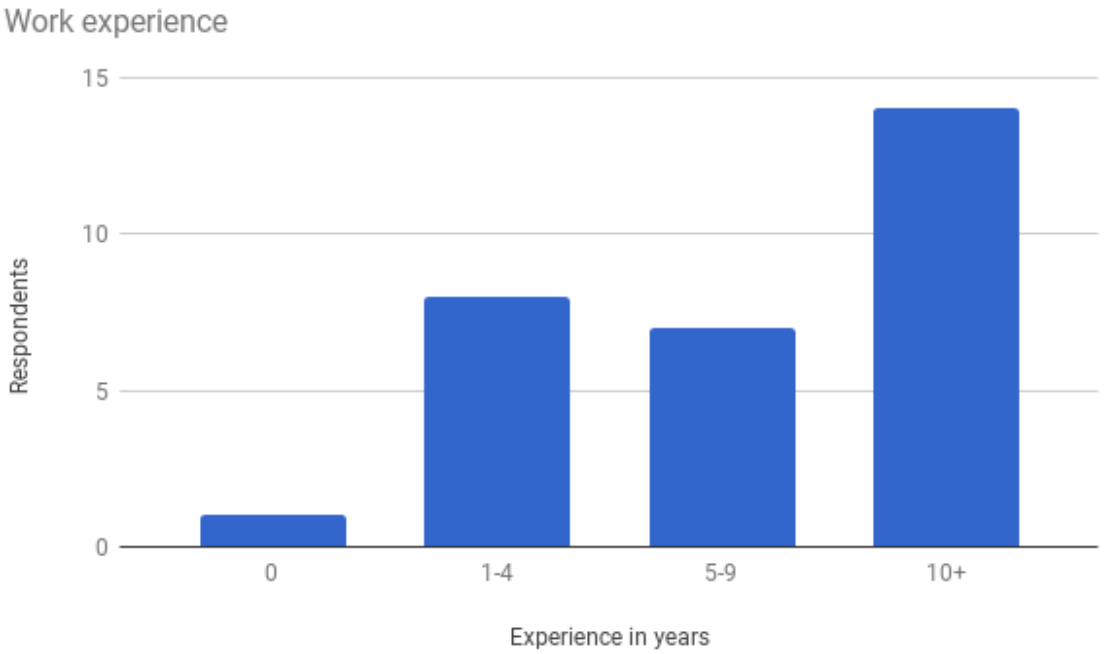

Figure 14 Respondents of the worldwide subgroup grouped by work experience with processes in years.

5 Survey Analysis Methodology

The results of the survey can be split into two parts: the quantitative and the qualitative results. The quantitative results consists of the ratings that were given by respondents about the familiarity and usefulness of each technique. The qualitative part consists of the comments they left on how they would use the solution.

On the quantitative part, statistical analysis was used to gain insights on respondents' attitudes towards process mining. To conduct analysis on the data gathered using Google Forms the data was downloaded in .csv format. The data headers were changed and then the statistical analysis tool R was used to melt this data into one big dataset and conduct the analysis.

In the first part, each technique was analysed based on the familiarity and usefulness rating of each respondent group. For each technique, a null hypothesis and an alternative hypothesis was formulated.

As an example, the following hypothesis was formed for discovery techniques.

For familiarity:

- H_0: Discovery techniques are not known in the subgroup
- $H_{1:}$ Discovery techniques are known in the subgroup

For usefulness

- H_0: Discovery techniques are not useful for the subgroup
- $H_{1:}$ Discovery techniques are useful for the subgroup

To test the hypothesis the One-Sample Wilcoxon Signed Rank Test was used. This test was used because the response data is not normally distributed and the best way to get the measure of central tendency for this data is too use the median. Wilcoxon Signed Rank Test is nonparametric, meaning it does not assume that the data is following a specific distribution.

On the qualitative part, only the most interesting proposals and insights are discussed.

6 Survey results

The survey results are discussed by techniques. Under each description, the ratings from the three subgroups are discussed based on the aspects of familiarity and usefulness. Each section also includes the qualitative part, the summary of respondents answer to the question how each technique could be used in the company the respondents work in.

This section is categorized in the following way:

- Section 6.1 is about discovery.
- Section 6.2 is about performance analysis.
- Section 6.3 is about optimization.
- Section 6.4 is about conformance checking.
- Section 6.5 is about prediction.
- Section 6.6 is about organizational mining.
- Section 6.7 is about decomposition.
- Section 6.8 is about model repair.
- Section 6.9 is about process deviance.
- Section 6.10 is about concept drift.
- Section 6.11 is about comparison.
- Section 6.12 sums up the paragraph and discusses the insights.

6.1 Discovery

The familiarity with discovery technique was rated low by Estonian and quality control subgroups with the median being 3 and 2 respectively. The worldwide subgroup was much more familiar with discovery the median value of the group is 5. After checking the result with the One-Sample Wilcoxon Signed Rank Test it can be said that worldwide group is familiar with process discovery while the quality control and Estonian subgroups are not.

In the case of usefulness, the median ranking of the quality control group is 4. This means the quality control group does not think discovery techniques are useful. The Estonian and worldwide groups differ in their opinions. In both subgroups, the median rating is 5. After checking the results of the Wilcoxon test we can conclude that the Estonian and worldwide subgroup find discovery techniques useful.

When looking at the figure it is clear that in the worldwide subgroup more people have given discovery the maximum rating in both aspects. While in the Estonian group there are not that many high ratings. It seems that the worldwide subgroup is much more familiar and sees more value in discovery than the Estonian subgroup does.

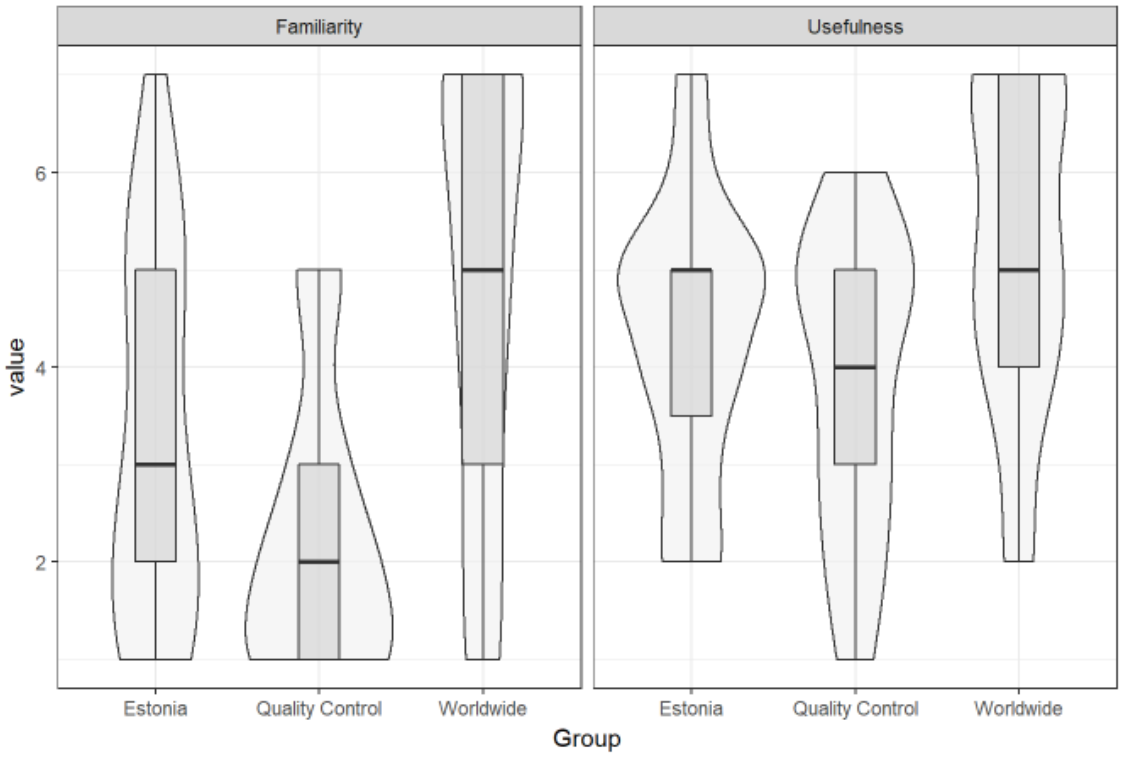

Figure 15 The ratings of familiarity and usefulness of each subgroup for discovery

When asked about how process discovery could be used in their company 34 out of the 92 respondents replied to this question. The respondents had quite a few ideas. The use they mentioned the most in different contexts was process mapping and performance analysis. Some people went even further and mentioned that the discovered model could be used to discover bottlenecks in the processes, optimize the process, use it for conformance checking with business rules and find deviances. There were some mentions about using process discovery to understand clients' movement through and information system and using discovery to understand and improve the usability of the IT system. Quite a few people who said that they work in small companies said that they have so few processes and IT systems that they see no use for process discovery. Overall the answers to this questions were to the point and it seems the people who answered it have a clear understanding of how this technique works.

6.2 Performance Analysis

Again as with discovery, the median familiarity with performance analysis is lower in Estonian and quality control subgroups. In both subgroups, it is 2. In the worldwide subgroup, the median value for familiarity is 4. For all three subgroups, it can be said that they are not familiar with process analysis technique.

The median rating of all three subgroups for the usefulness is 5, but the results of the Wilcoxon Signed Ranked Test showed that we do not have statistically significant evidence to conclude that performance analysis is useful for the Estonian and Quality Control subgroups. However, the Wilcoxon Signed Ranked Test showed that there is enough statistically significant evidence to say that performance analysis techniques are useful for the worldwide subgroup.

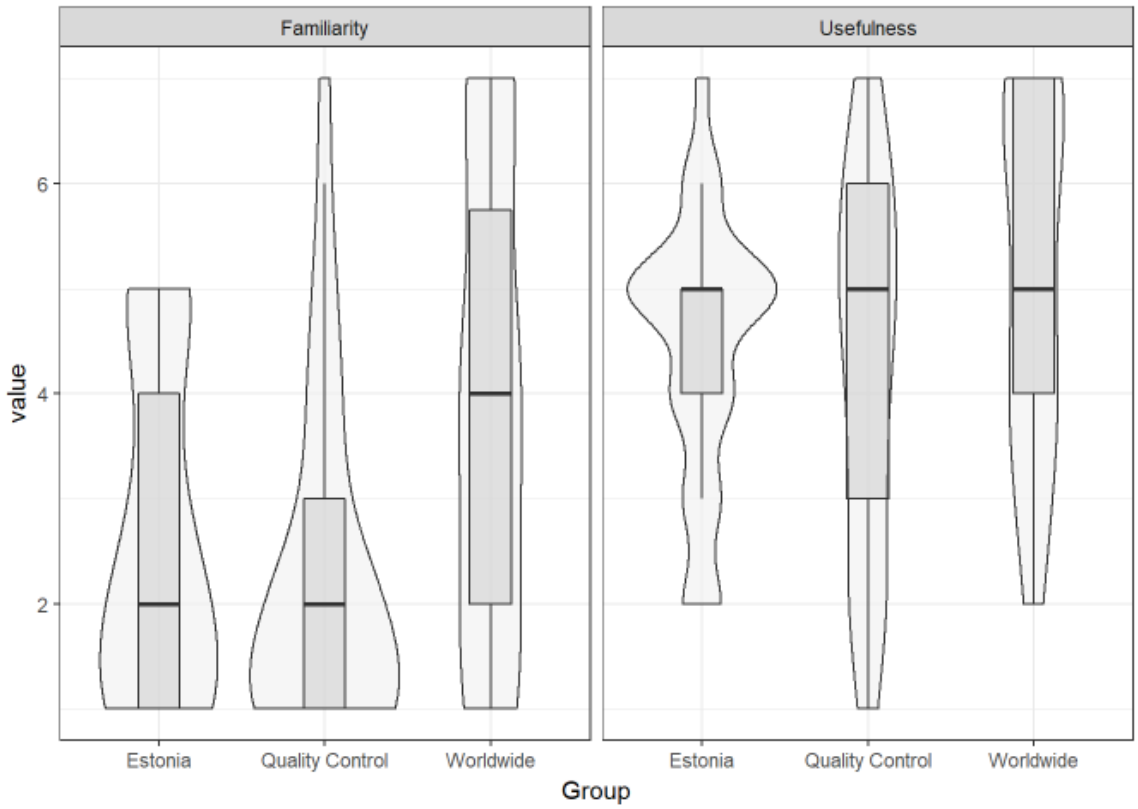

Figure 16 The ratings of familiarity and usefulness of each subgroup for performance analysis

37 people answered the question about how to use performance analysis in their company. The vast majority named that they would be interested in gaining insight how the process works or in finding the bottlenecks in the process so they could be optimized. Some people mentioned that this could be used to gain insight into resource usage and to see where the more resources are needed. There were a few answers in the worldwide subgroup that expressed doubt that this technique can be used for their processes.

6.3 Optimization

The median familiarity of quality control and Estonian subgroups with optimization is rated at 3 and 4 respectively. Meaning that neither subgroup is familiar with process optimization. However, the worldwide subgroup median familiarity with optimization is at 6. Meaning that this is the second technique the worldwide group is familiar with.

All three groups agree that optimization is useful. All of the groups have the median rating at or near 5. The One Sample Wilcoxon Rank Signed Test also supports this median. This means all three subgroups find optimization useful.

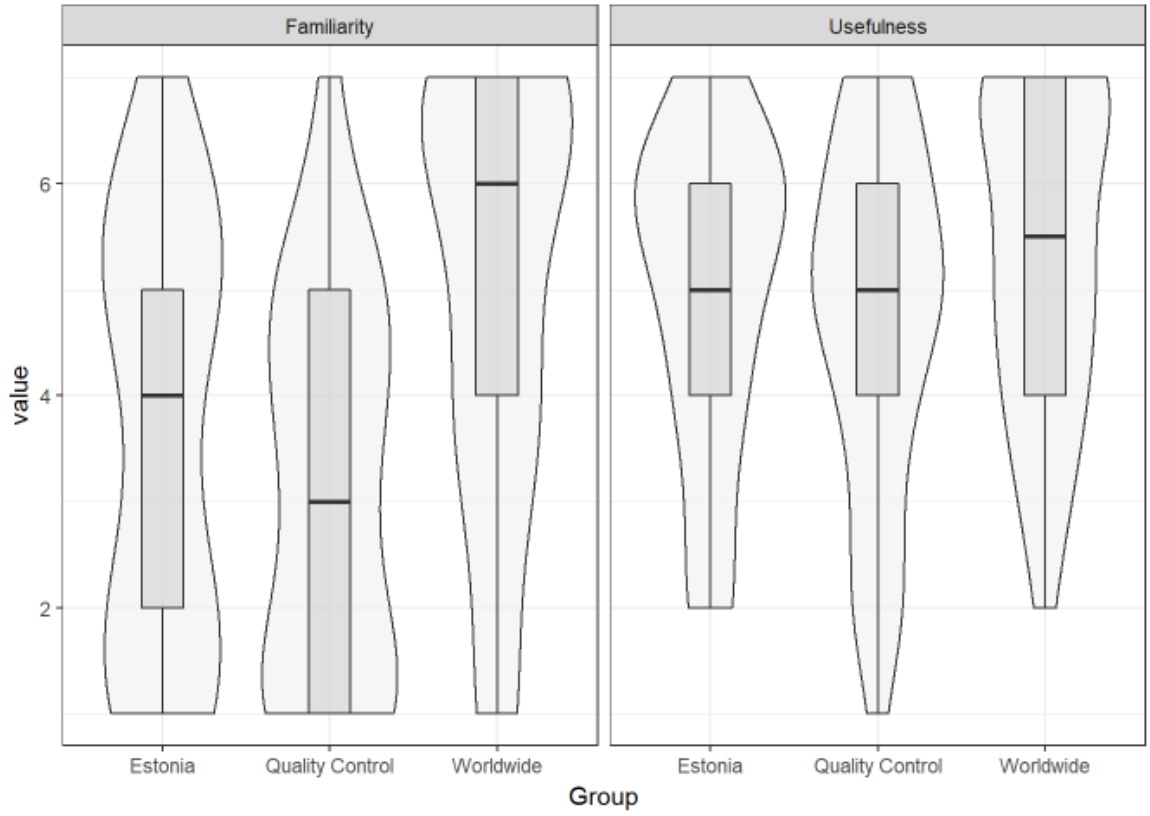

Figure 17 The ratings of familiarity and usefulness of each subgroup for optimization

29 people answered the optional question about how optimization could be used in their company. Based on the answers the respondents generally understood the value of optimizing processes and quite a few people mentioned that they do it all the time in their company. However, there were some respondents in the Estonian subgroup who said they do not see the value of using process mining to conduct process optimization as processes optimization efforts can be conducted by using only interviews and going through the documentation. This gives the impression that among the Estonian subgroup there are people who do not understand that process mining can new insights into processes that other techniques might not.

6.4 Conformance Checking

The familiarity of conformance checking ranges highly between groups. The median familiarity rating in the quality control group is 1. In the Estonian group it is 2. Again the worldwide group sticks out with a higher median rating of 4.5. However, there is not enough statistical evidence to reject the null hypothesis. In conclusion all the subgroups are not familiar with conformance checking.

The median rating of usefulness is 4 for both the quality control and Estonian groups meaning conformance checking is not useful Estonian and quality control subgroups. On the other hand the worldwide group median rating is 5. Meaning that the worldwide subgroup sees conformance checking as useful.

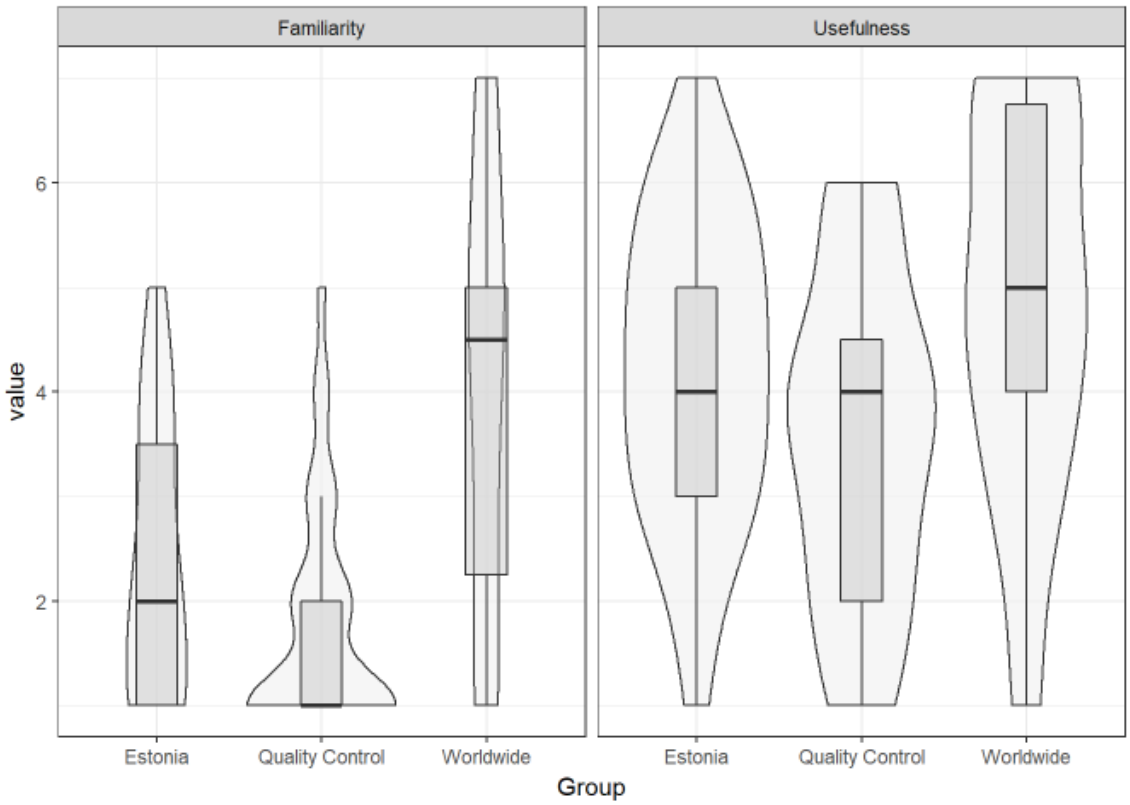

Figure 18 The ratings of familiarity and usefulness of each subgroup for conformance checking

28 people expressed their thoughts about using conformance checking in their company. Most people felt it could be used to check if the execute activities are following the guidelines set out for the process. Some people thought this technique could best be used to check if processes are compliant with business and data protection rules. One person mentioned this technique could be used in their company for inner audits on processes. Another respondent was interested in finding deviances using conformance checking. What seemed to be a problem for some respondents is that they do not have their processes documented or they are not logging the data that is necessary to use this technique. However, most of the respondents felt that if this technological side was set up in their company this technique would provide value for them.

6.5 Prediction

The median familiarity with prediction is low in all the groups it is below 4 meaning none of the subgroups are familiar with prediction.

In case of usefulness things are not much better. Although the median usefulness rating in all the groups is 4 it still means that prediction is thought of as a useless technique.

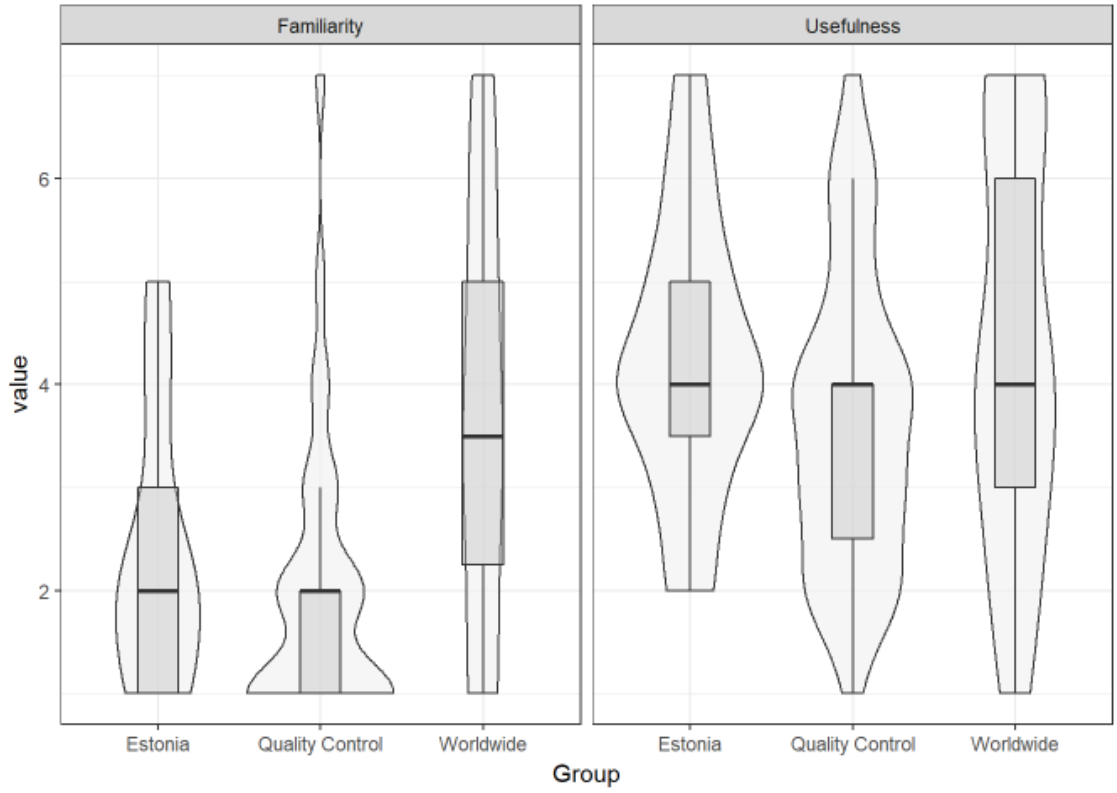

Figure 19 The ratings of familiarity and usefulness of each subgroup for prediction

23 people answered the question about the usage of this technique in their company. There were quite many respondents who were very positive about using prediction techniques. For instance, one respondent said that in the hospital they have a lot of pilot projects, where they test out changed processes. Using prediction techniques to try out the optimized scenario, the hospital would not have to have so many pilot projects. The other respondents also mentioned that if prediction could be used to understand how the process would work after the optimization this would be extremely useful for them. There were some respondents whose attitudes were negative towards this technique. One respondent said he don't see this technique being used in their company anytime soon as they would first need to use the basic process mining techniques like process discovery or performance analysis. Another respondent said that their processes change very fast and there are a lot of factors that need to be taken into account. Due to these reasons he did not think prediction techniques could be used in their company.

6.6 Organizational Mining

The median familiarity with organizational mining is low. Estonian and quality control group median is 2, while the worldwide median is 3. Based on this none of the subgroups are familiar with organizational mining.

The median usefulness of organizational mining is rated higher than the median familiarity. The median is 4 for both the quality control and the Estonian subgroup. The worldwide subgroup's median is a bit higher, being 4.5. However, the Signed Rank Wilcoxon Test still showed that there is not enough statistically significant evidence to conclude that worldwide subgroup finds organizational mining useful. Therefore we have to conclude that all subgroups find organizational mining not useful.

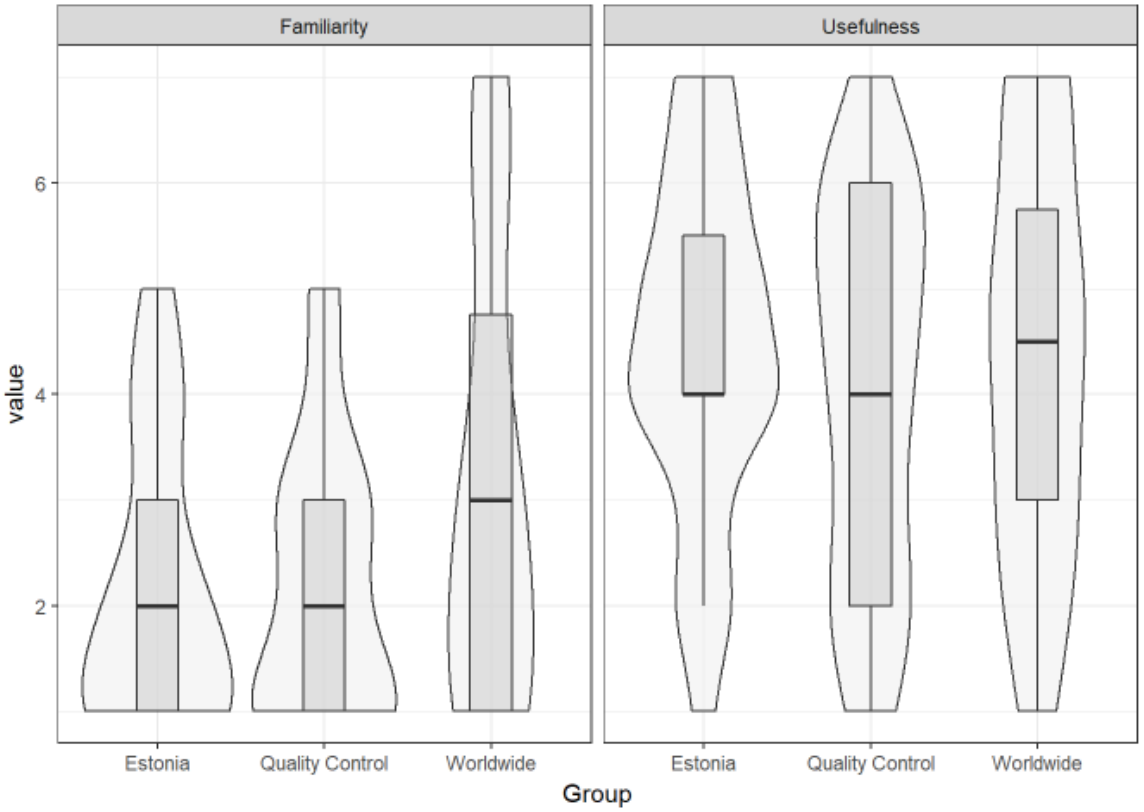

Figure 20 The ratings of familiarity and usefulness of each subgroup for organizational mining

28 respondents described how they see this technology being used in their company. 19 people were saying that they see value in this technique. The most mentioned value point was understanding the work load of the workers and having a better resource allocation to optimize the workforce use. Quite a few of these respondents also mentioned this is because they have problems in their company with some workers being overflowed by work. Some respondent said mapping the duplication of tasks by workers would be useful.

9 people were negative about this technique. They said that they do not see the value in this technique for their company, they have other means of gathering this data or they just do not think it could be used in their company.

6.7 Decomposition

Decompositions median familiarity is different in each subgroup. In quality control it is 2, In Estonia, it is 3 and worldwide it is 4. With none of the values being above for. We can say that none of the subgroups are familiar with decomposition.

The median ratings of usefulness are different in each subgroup. For quality control, it is 3, 4 for Estonia and 5 worldwide. Although for quality control and Estonian subgroups decomposition is not useful, it is found useful by the worldwide subgroup.

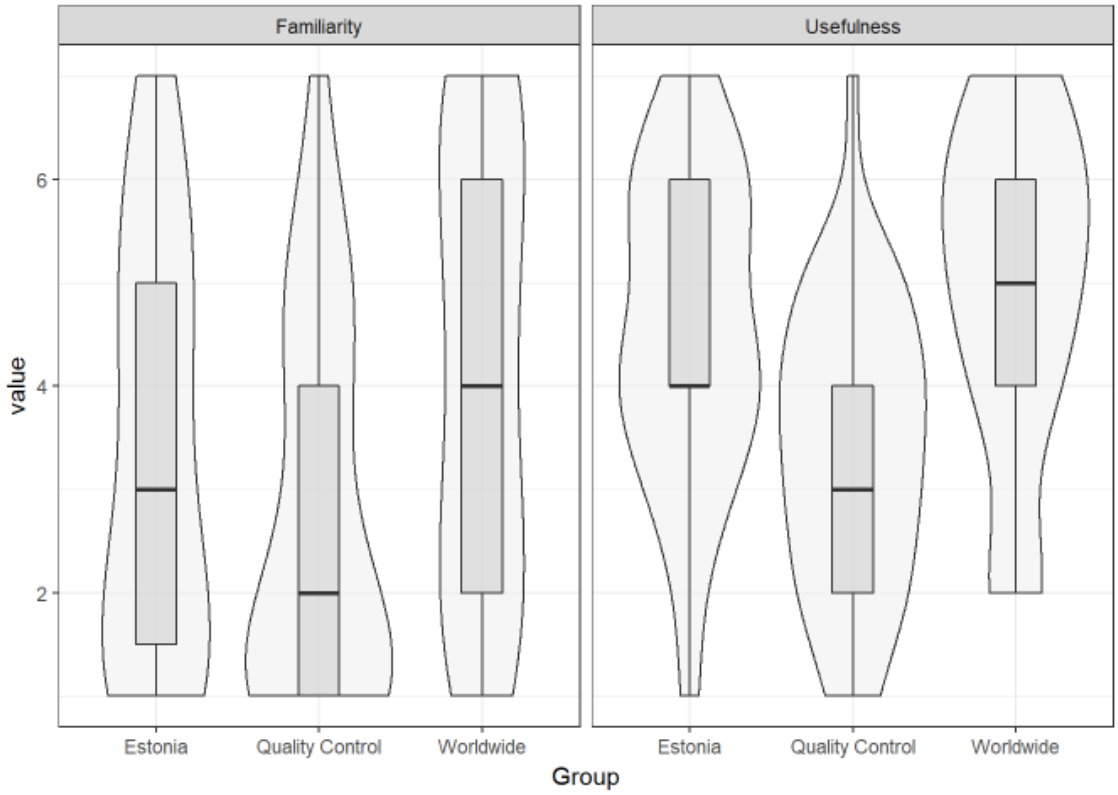

Figure 21 The ratings of familiarity and usefulness of each subgroup for decomposition

15 people described how they see decomposition could be applied in their companies. Quite a few of the respondents correctly pointed out that this is a preparatory technique that in itself might not provide a lot of value, but when used in conjunction with other techniques it would be valuable, because it gives insight into parts of the process. Some people even mentioned processes they would be interested in decomposing like the processes bought from other vendors or the bill of material and assembly process. 1 person said that they see no use for it in their company. The reason for this being that they do not have complicated processes that need to be decomposed.

6.8 Model Repair

The median familiarity of model repair is quite low among all subgroups. It is the lowest among the quality control subgroup with median being only 1. Among the Estonian subgroup it is 2 and in the worldwide subgroup the median is 3. With such a low median we can conclude that model repair is not known to any subgroup.

The median rating for usefulness are higher. Quality control subgroups median rating is 3, Estonian subgroup has rated it as 4 and the worldwide subgroup as 5. Although the median rating of the worldwide subgroup is over 5 the One Sample Wilcoxon Signed Rank Test shows that there is not enough statistical evidence to conclude that model repair is useful for the worldwide group. Therefore model repair techniques is not considered useful by any of the subgroups.

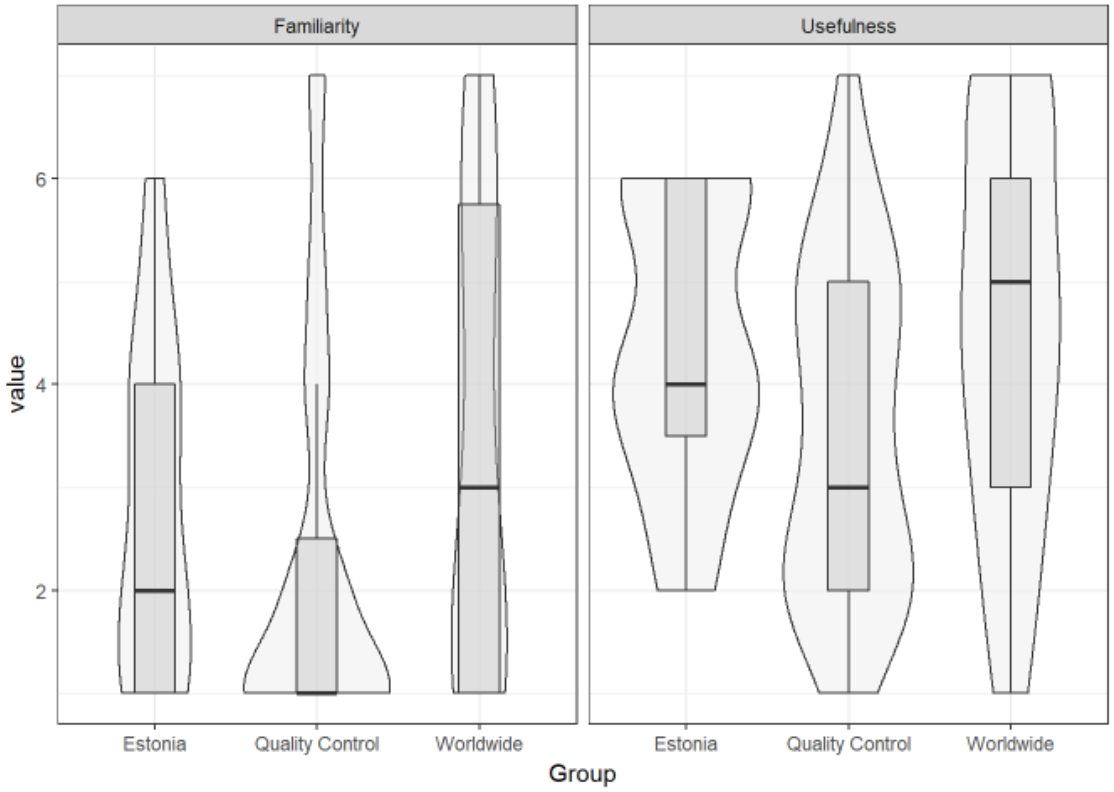

Figure 22 The ratings of familiarity and usefulness of each subgroup for model repair

11 people answered the question about the usage of model repair in their company. It seems this technique remained quite vague to the respondents. Most of the answers were one-liners where is it is hard to understand what the respondents actually meant. 3 people said that they do not really understand the technique and how the value it gives. In one case it seems the respondent misunderstood that the technique somehow shows inefficiencies in process. Only one respondent said it would be useful to identify the gaps in process models.

6.9 Process Deviance

The median familiarity with process deviance ranges from 2 to 3.5 meaning that process deviance is not known for any of the subgroups.

The median usefulness is 4 for Estonian and quality control subgroups. For the worldwide group it is 5. After checking the result for the worldwide subgroup with Wilcoxon signed ranked test, it shows that process deviance is considered useful by the worldwide subgroup but it is considered not useful by Estonian and quality control subgroups.

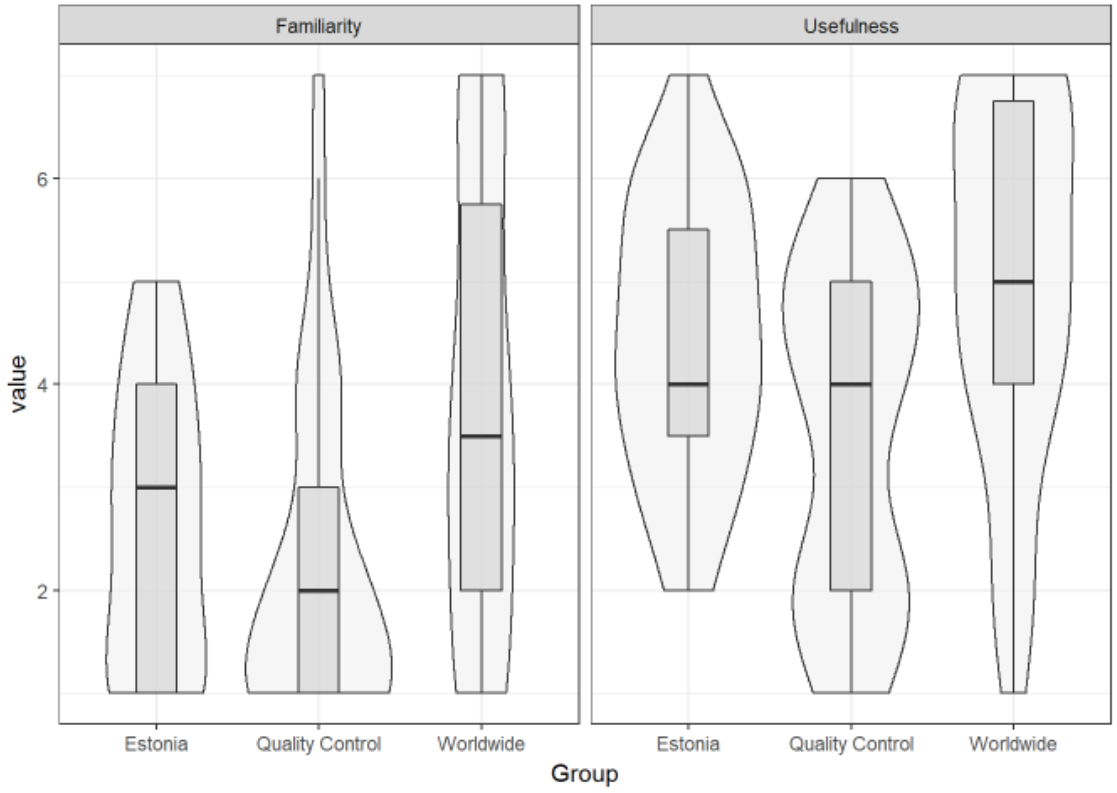

Figure 23 The ratings of familiarity and usefulness of each subgroup for process deviance

17 people gave a response on how this technique could be used in their company. All the respondents saw value in this technique. Most of the respondents felt process deviance is best used to gain an understanding where process deviate and then use this knowledge to drive process optimization. Another main idea that came out of the responses was that this technique would be good to use in order to check if the process has been conducted following the requirements of the process. In general the respondents did not discuss in detail how they see this technique being applied. One the few respondents who proposed an actual use to this technique thought it could be used to analyse the work of customer service representatives.

6.10 Concept Drift

As the literature review showed concept drift is a fairly new and quite under researched topic. Therefore seeing that the median familiarity rating of concept drift is under 4 for all subgroups is too be expected. The quality control group rated the familiarity lowest with the median being only 1, the Estonian group rated it 2 and the worldwide group rated it at 3. Therefore we can conclude that concept drift is not known to any of the subgroups.

The usefulness rating is higher compared to the familiarity ratings. In the quality control subgroup the median is 3, in the Estonian subgroup it is 4 and in the worldwide group it is 4.5. Checking the median of the worldwide group with the One Sample Wilcoxon Signed Rank Test it shows that there is not enough statistically significant evidence to reject the null hypothesis. This means that all subgroups find concept drift technique not useful.

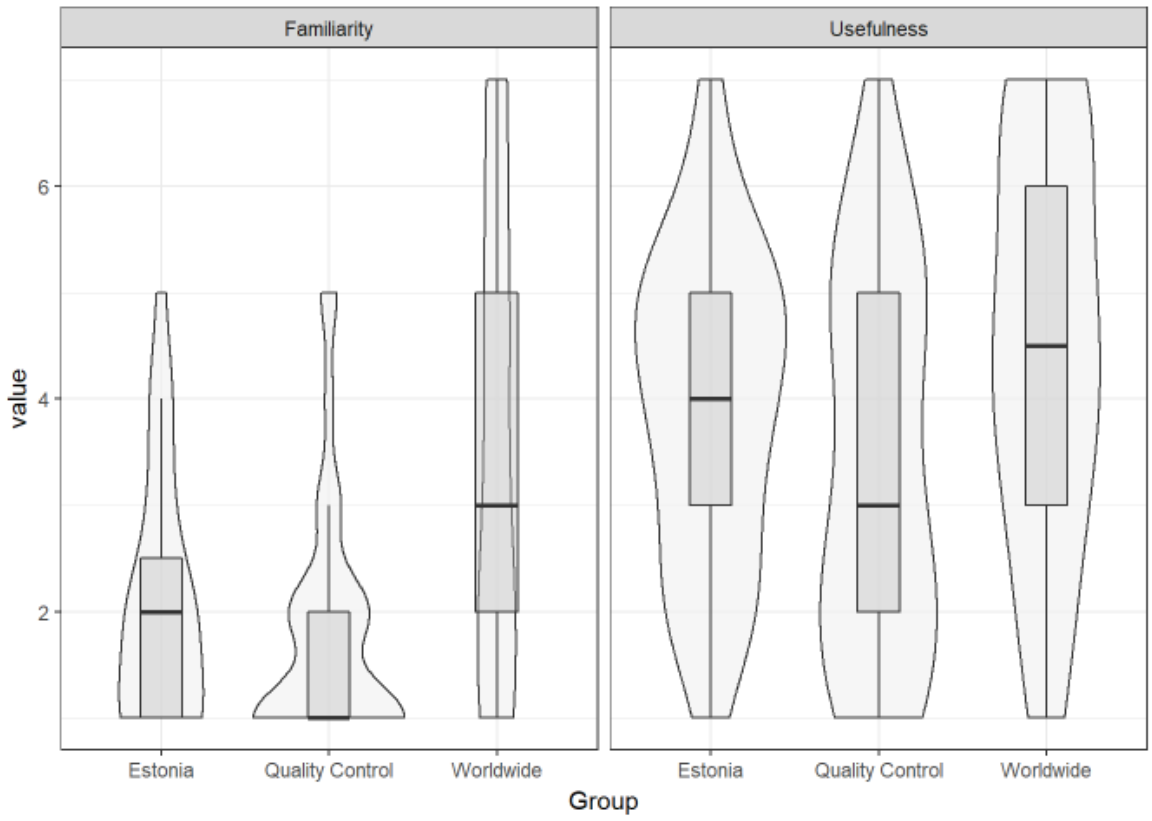

Figure 24 The ratings of familiarity and usefulness of each subgroup for concept drift

14 people wrote an additional comment how they see this technique being used in their company. The sentiments ranged from positive to negative. Half the people understood that some processes change over time and saw value in using this technique to signal changes in the processes. Half the respondents did not really understand the concept or felt this would not be useful for them.

6.11 Process Comparison

Process Comparison is also quite a new field and there are only a few research papers on this topic. So it is to be expected that the median familiarity rating in the quality control group is just 1 and 2 in the Estonian group. What is a bit surprising is that the median in the worldwide group is 4.5 which is quite high. The One Sample Wilcoxon Signed Rank Test supports this median. On the familiarity aspect, we can conclude that for the quality control and Estonian subgroup process comparison technique is not known, while for the worldwide group process comparison technique is known.

When looking at the median usefulness rating the subgroups are closer in their rating quality control and Estonian subgroups median rating is 4. While it is 5 for the worldwide subgroup. This means we can say that there is not enough statistically significant evidence to conclude that the quality control and Estonian subgroups find process comparison techniques useful. However, the median being over 4 for the worldwide group we needed to run a One Sample Wilcoxon Signed Rank Test to validate the median. The result validated the median meaning that we can say that there is enough statistically significant evidence to say that process comparison technique is useful to the worldwide subgroup.

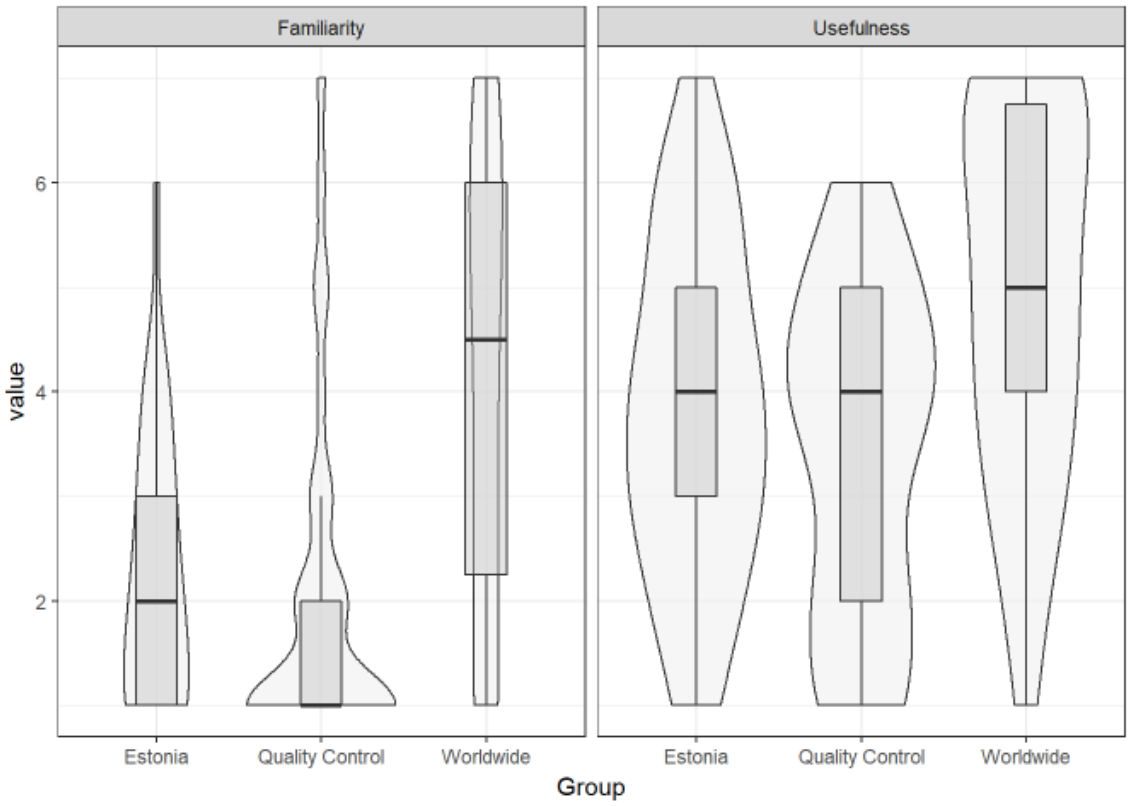

Figure 25 The ratings of familiarity and usefulness of each subgroup for process comparison

The responses to how process comparison could be used in companies were rather plain. The general sentiment was positive that this technique would be useful, but only a few people actually described with more details. One respondent said it could be useful to identify incremental changes to processes or differences between multiple process runs. Another brought out that this technique could be used to identify similarities in processes.

6.12 Summary & Discussion

This section will sum up the results of the survey and discuss the results. This will be done on the subgroup and technique basis.

The quality control and the Estonian subgroups were quite similar in their ratings. As can be seen from Figure 26Figure 1. Both groups were not familiar with any of the process mining techniques. Also, both groups found optimization useful. The only difference between the groups was that the Estonian group also found process discovery useful. However, one still needs to note that the median rating of familiarity on the majority of the techniques was 1 point higher in the Estonian subgroup. Therefore we can say that the Estonian group is a bit more informed with process mining techniques than the quality control group was. The same applies for the usefulness ratings although there the groups have given more techniques the same ratings and the difference is smaller. Still, we need to conclude that the both quality control group and the Estonian group are not familiar with any of the process mining techniques and in general do not find process mining techniques useful. When considering that the results from the worldwide group were much more positive we can say that the Estonian industry is lagging behind. It seems that there is still a lot of work to do in Estonia on introducing process mining to the industry.

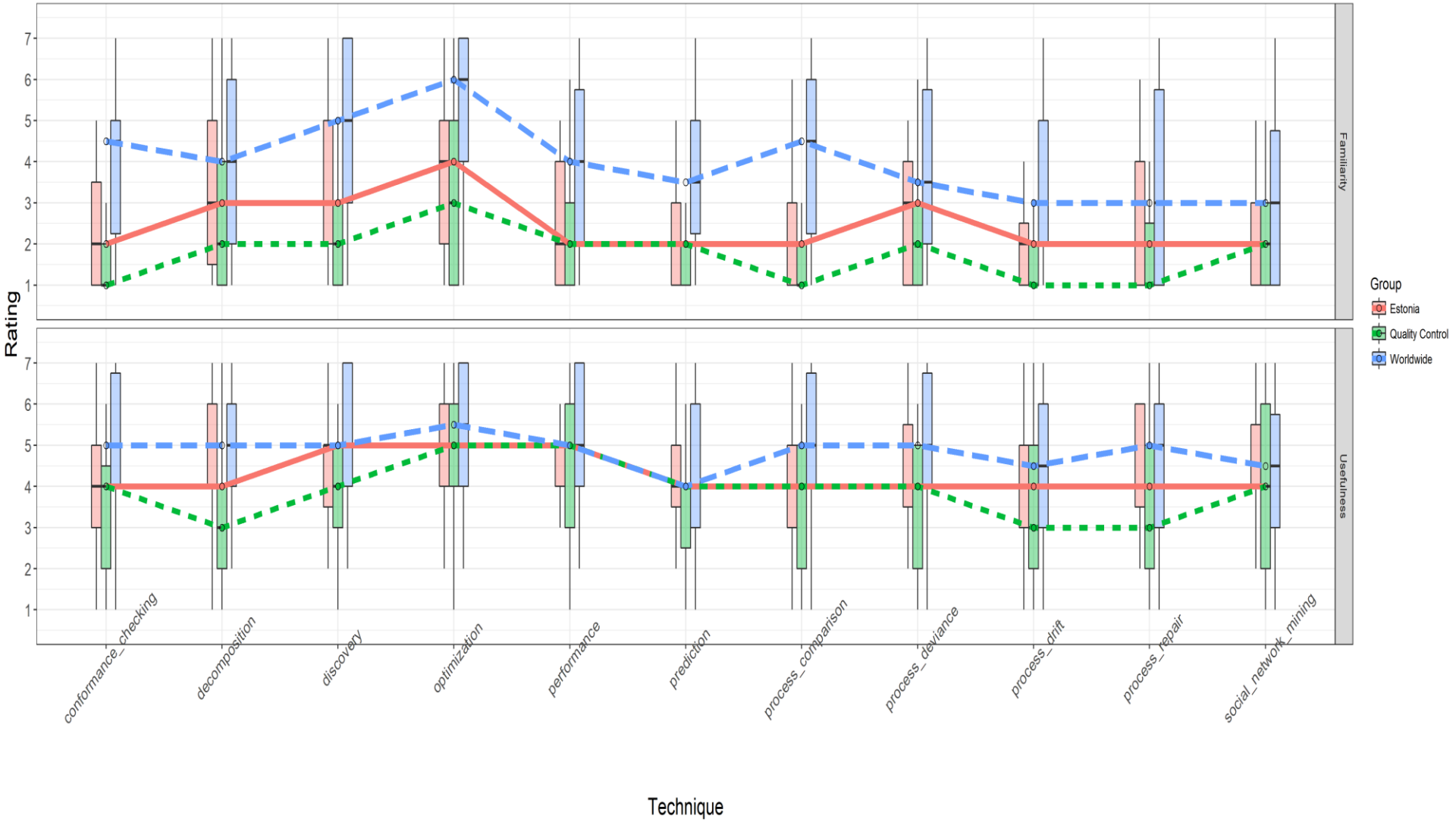

Figure 26 Median results of the survey (familiarity & usefulness)

When looking at the worldwide subgroup we can notice that it is different from the other subgroups. To start with they are familiar with two process mining techniques: discovery and optimization. Furthermore, out of the 11 process mining techniques this group found 6 of them useful. The six being discovery, performance analysis, conformance checking, optimization, decomposition, process deviance and process comparison. Considering the low amount of papers that were found during the literature review on decomposition and process comparison it was surprising to see that the worldwide group finds these techniques so valuable. Process deviance also had a low number of papers, but considering how much deviations from the standard processes can cost to companies this was more or less expected. However, from this results we can conclude that worldwide the industry sees a need to have techniques that enable comparing different business processes and they also need better techniques to decompose complicated process models. Overall we can say that the worldwide subgroup sees much more value process mining than Estonian subgroup does.

When considering the techniques definitely the most interesting finding is that prediction techniques were not found familiar or useful by any subgroup. This is quite surprising as more and more companies are looking towards data mining and machine learning to grow their businesses. When considering the answers given to the question about how this technique could be used in their companies it was clear that the respondents were divided into two. One group saw the future in predictive technologies and the other felt these could not be used. What seemed to characterize people, who were negative about using prediction was that they came from companies that were very small or had very variable processes. Indeed it might not make sense to build predictive models for companies with only a few employees. The same goes for processes that have too many variables. However, in large companies with standardized processes predictive technologies could create a significant advantage. As this survey did not inquire about the size or characteristics of the company it is very hard to speculate what were the drivers of rating prediction so low. This is something that could be researched more in the future.

Another interesting technique to look at is optimization. When looking at how much research has been done on this field it will fall in almost the same place as organizational mining, between prediction and less researched techniques like decomposition and model repair. Even though there is not that much research in the field of process mining about this technique the industry is very familiar with this technique and sees it as very useful. This was to be expected as optimization of processes is something companies have done for a long time before process mining. So they have the knowledge on how to optimize and why is it useful. However, it remains unclear whether they understand how process mining could be utilized to optimize their processes better.

Discovery was found useful by both the Estonian and the worldwide group. Although it was not rated as familiar by the Estonian subgroup this shows that the industry considers discovery as one of the most important techniques of process mining. When we take into consideration the fact that optimization is known to companies even without the context of process mining one could even say that it is the flagship technique of process mining.

Conformance checking, performance, decomposition and process comparison were found useful by the worldwide subgroup, but not by the Estonian subgroup. This indicates either that the worldwide group understands the value these techniques provide better or they are struggling with these issues more than the Estonian subgroup.

Organizational mining, process repair and concept drift were rated low on both aspects by all the groups. Still the usefulness ratings of these techniques were near 4 in all subgroups meaning these techniques might prove to be useful to the industry in the future when they evolve more.

7 Conclusion

In this paper, a systematic literature review was conducted to understand what the different process mining techniques are. A literature review protocol and search string were set into place. This was followed by searching papers from online databases using the predefined search string. The papers went through 3 iterations of filtering were irrelevant papers were discarded. From 729 papers in the first iteration, only 145 papers were left. The papers were then read to determine the most interesting ones that were included in the systematic literature results to illustrate the process mining techniques.

The literature review served as an input to a survey on the usefulness of process mining techniques held among industry representatives from different fields. The aim was to understand which process mining techniques industry representatives are most interested in. The survey was held among a quality control group made up of industry representatives with no knowledge about process mining, Estonian industry representatives with knowledge of process mining and a worldwide group of industry representatives. The results were analyzed based on each subgroup and technique.

The literature review identifies different process mining techniques and shows that process mining offers many different analysis techniques to better analyze business processes. Even more, it shows that each year more and more papers are being published in the field of process mining using real-life data. This shows that the discipline is maturing and it the time, when it will be widely adopted by the industry is now not a question of decades.

The main findings of the survey among industry representatives were that compared to the worldwide subgroup the Estonian subgroup rates their familiarity lower for all process mining techniques. Furthermore, Estonian subgroup finds only process discovery and optimization techniques useful while the worldwide subgroup sees process discovery, performance analysis, conformance checking, optimization, decomposition and process comparison as useful techniques. Based on this it was concluded that the worldwide industry representatives value process mining techniques more than Estonian industry representatives.

When looking at the findings based on each technique it was surprising to see that prediction was rated so low that it had to be concluded that none of the subgroups find process prediction useful. Considering how much there is talk in the industry about leveraging data science and machine learning it was interesting to see that the industry representatives rated process prediction so low. It shows that the process mining community needs to do a lot more to persuade the industry in the usefulness of process prediction.

This work can be extended in the future. Currently, the main drawback of this research is the limited sample of industry representatives who took the survey. This survey had 92 respondents. All 3 subgroups contained about 30 people. While it is enough to get insight into how the industry feels about process mining techniques it is not enough to really understand the circumstances driving the opinions of the industry. To do this the sample of industry representatives needs to be larger. The questionnaire should also include questions about the background of the company the respondents work in and capture the respondents' experiences with process mining.

8 References

1 van der Aalst, W.M.P. *Process Mining - Discovery, Conformance and Enhancement of Business Processes*. Springer, 2011.

2 Kitchenham, B. *Procedures for Performing Systematic Reviews*. Software Engineering Group, Department of Computer Science, Keele University, 2004.

3 Manoj Kumar, M.V., Thomas, L., Annappa, B. Capturing the sudden concept drift in process mining. In *CEUR Workshop Proceedings* (2015), 132-143.

4 Paster, F., Helm, E. From IHE Audit Trails to XES Event Logs Facilitating Process Mining. In *Studies in Health Technology and Informatics* (2015), 40-44.

5 Engel, R., Krathu, W., Zapletal, M., Pichler, C., Van Der Aalst, W.M.P., Werthner, H. Process mining for electronic data interchange. In *Lecture Notes in Business Information Processing* (2011), 77-88.

6 Soo, A. *Benchmarking process discovery*. Tartu University, Tartu, 2017.

7 Di Ciccio, C., Maggi, F. M., Mendling, J. Efficient discovery of Target-Branched Declare constraints. In *Information Systems* (2016).

8 Conforti, R., Dumas, M., García-Bañuelos, L., La Rosa, M. BPMN Miner: Automated discovery of BPMN process models with hierarchical structure. In *Information Systems* (2016).

9 Nguyen, H., Dumas, M., Ter Hofstede, A.H.M., Rosa, M.L., Maggi, F.M. Business process performance mining with staged process flows. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (2016), 167-185.

10 Kelly, N., Montenegro, M., Gonzalez, C., Clasing, P., Sandoval, A., Jara, M., Saurina, E., Alarcón, R. Combining event- and variable-centred approaches to institution-facing learning analytics at the unit of study level. In *International Journal of Information and Learning Technology* (2017), 63-78.

11 Park, J., Jung, J.-Y., Jung, W. The use of a process mining technique to characterize the work process of main control room crews: A feasibility study. *Reliability Engineering and System Safety*, 154 (2016), 31-41.

12 Senderovich, A., Weidlich, M., Yedidsion, L., Gal, A., Mandelbaum, A., Kadish, S., Bunnell, C.A. Conformance checking and performance improvement in scheduled processes: A queueing-network perspective. *Information Systems*, 62 (2016), 185-206.

13 Rozinat, A., van der Aalst, W.M.P. Conformance checking of processes based on monitoring real behavior. *Information Systems*, 33, 1 (2008), 64-95.

14 Caron, F., Vanthienen, J., Baesens, B. Business rule patterns and their application to process analytics. In *Proceedings - IEEE International Enterprise Distributed Object Computing Workshop, EDOC* (2013), 13-20.

15 Ramezani, E., Fahland, D., Van Der Aalst, W.M.P. Where did I misbehave? Diagnostic information in compliance checking. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (2012), 262-278.

16 Verenich, I., Dumas, M., La Rosa, M., Maggi, F.M., Francescomarino, C.D. Complex symbolic sequence clustering and multiple classifiers for predictive process monitoring. In *Lecture Notes in Business Information Processing* (2016), 218-229.

17 Leontjeva, A., Conforti, R., Di Francescomarino, C., Dumas, M., Maggi, F.M. Complex symbolic sequence encodings for predictive monitoring of business processes. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (2015), 297-313.

18 Maggi, F.M., Di Francescomarino, C., Dumas, M., Ghidini, C. Predictive monitoring of business processes. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (2014), 457-472.

19 van der Spoel, S., van Keulen, M., Amrit, C. Process Prediction in Noisy Data Sets: A Case Study in a Dutch Hospital. In *Lecture Notes in Business Information Processing* (2013), 60-83.

20 Song, M., van der Aalst, W.M.P. Towards comprehensive support for organizational mining. *Decision Support Systems*, 46, 1 (2008), 300-317.

21 Riz, G., Santos, E.A.P., Loures, E.D.F.R. Process mining to knowledge discovery in healthcare processes. In *Advances in Transdisciplinary Engineering* (2016), 1019-1028.

22 de Leoni, M., Munoz-Gama, J., Carmona, J., van der Aalst, W.M. Decomposing alignment-based conformance checking of data-aware process models. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (2014), 3-20.

23 Polyvyanyy, A., Van Der Aalst, W.M.P., Ter Hofstede, A.H.M., Wynn, M.T. Impact-driven process model repair. *ACM Transactions on Software Engineering and Methodology*, 25, 4 (2016).

24 Rovani, M., Maggi, F.M., De Leoni, M., Van Der Aalst, W.M.P. Declarative process mining in healthcare. *Expert Systems with Applications*, 42, 23 (2015), 9236-9251.

25 Swinnen, J., Depaire, B., Jans, M.J., Vanhoof, K. A process deviation analysis - A case study. In *Lecture Notes in Business Information Processing* (2012), 87-98.

26 Sebu, M.L., Ciocarlie, H. Business activity monitoring solution to detect deviations in business process execution. In *SACI 2015 - 10th Jubilee IEEE International Symposium on Applied Computational Intelligence and Informatics, Proceedings* (2015), 437-442.

27 Bose, R.P.J.C., Van Der Aalst, W.M.P., Zliobaite, I., Pechenizkiy, M. Dealing with concept drifts in process mining. *Journal of Software: Evolution and Process*, 26, 7 (2014), 714-728.

28 Bolt, A., De Leoni, M., Van Der Aalst, W.M.P. A visual approach to spot statistically-significant differences in event logs based on process metrics. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (2016), 151-166.

29 Montani, S., Leonardi, G., Quaglini, S., Cavallini, A., Micieli, G. A knowledge-intensive approach to process similarity calculation. *Expert Systems with Applications*, 42, 9 (2015), 4207-4215.

30 van der Aalst, W.M.P., Guo, S., Gorissen, P. Comparative process mining in education: An approach based on process cubes. In *Lecture Notes in Business Information Processing* (2015), 110-134.

Appendix

I. List of literature

The list of literature is available on the following address:

https://docs.google.com/spreadsheets/d/1eWDlDPQa7Ddy1C2QJKXNl0GJp7jM8TeWq-gGRcF5umE/edit?usp=sharing

II. License

Non-exclusive licence to reproduce thesis and make thesis public

I, **Karli Oruste**,

(*author's name*)

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to:

   1.1. reproduce, for the purpose of preservation and making available to the public, including for addition to the DSpace digital archives until expiry of the term of validity of the copyright, and

   1.2. make available to the public via the web environment of the University of Tartu, including via the DSpace digital archives until expiry of the term of validity of the copyright,

of my thesis

Process Mining in Industry,

(title of thesis)

supervised by Frederik Payman Milani, Fabrizio Maria Maggi,

(supervisor's name)

2. I am aware of the fact that the author retains these rights.

3. I certify that granting the non-exclusive licence does not infringe the intellectual property rights or rights arising from the Personal Data Protection Act.

Tartu, **15.08.2017**