Classifying Process Instances Using Recurrent Neural Networks — **Source link** ⬈

Markku Hinkka, Teemu Lehto, Keijo Heljanko, Keijo Heljanko ...+1 more authors

Institutions: Aalto University, Helsinki Institute for Information Technology

Published on: 09 Sep 2018 - Business Process Management

**Topics:** Process mining, Deep learning and Recurrent neural network

Related papers:

- Classifying Process Instances Using Recurrent Neural Networks

- Predicting process behaviour using deep learning

- Predictive Business Process Monitoring with LSTM Neural Networks

- Deep Learning Process Prediction with Discrete and Continuous Data Features.

- Learning Accurate LSTM Models of Business Processes

Process Mining Based Influence Analysis for Analyzing and Improving Business Processes
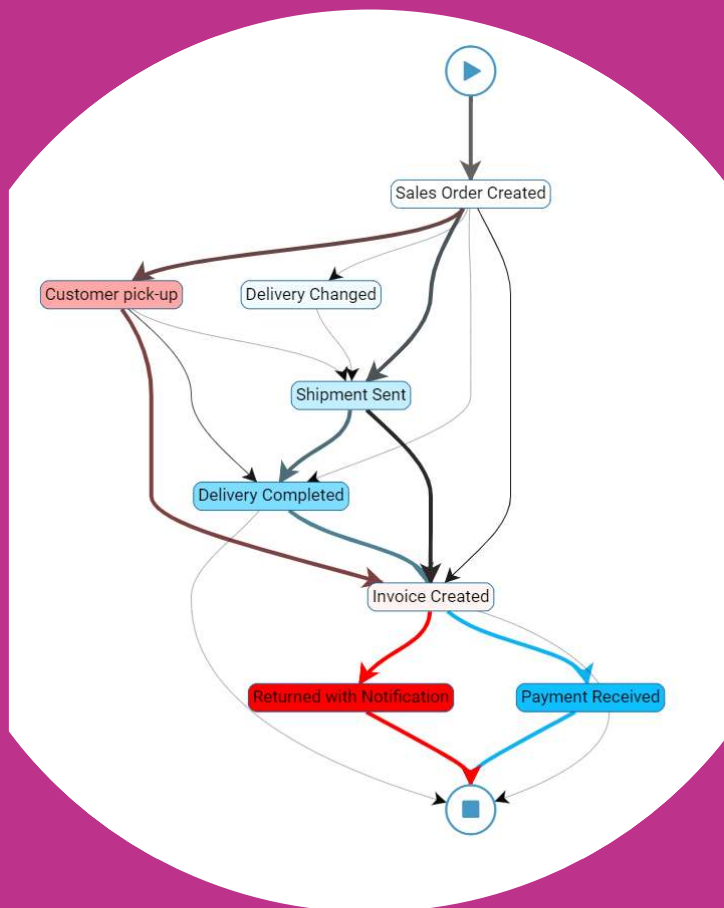
Teemu Lehto

Process Mining Based Influence Analysis for Analyzing and Improving Business Processes

Teemu Lehto

A doctoral dissertation completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Science, Remote connection via Zoom, on 26 November 2020 at 2.00pm.

Aalto University
School of Science
Computer Science

Supervising professor
Assistant Professor Alex Jung, Aalto University, Finland

Thesis advisor
Dr. Jaakko Hollmén, Aalto University, Finland

Preliminary examiners
Prof.dr.ir. Wil van der Aalst, RWTH Aachen University, Germany
Assistant Professor, Felix Mannhardt, Eindhoven University of Technology, The Netherlands

Opponent
Prof.dr.ir. Wil van der Aalst, RWTH Aachen University, Germany

NORDIC SWAN ECOLABEL

Printed matter
4041-0619

Author
Teemu Lehto

Name of the doctoral dissertation
Process Mining Based Influence Analysis for Analyzing and Improving Business Processes

Publisher School of Science

Unit Department of Computer Science

Series Aalto University publication series DOCTORAL DISSERTATIONS 187/2020

Field of research Computer Science

Manuscript submitted 7 April 2020	**Date of the defence** 26 November 2020

Permission for public defence granted (date) 23 October 2020   **Language** English

☐ **Monograph**   ☒ **Article dissertation**   ☐ **Essay dissertation**

Abstract

The ability to improve processes is essential for every organization. Process mining provides a fact-based understanding of actual processes in the form of discovered process diagrams, bottlenecks, compliance issues, and other operational problems. Organizations need to carry out accurate root cause analysis and efficient resource allocation to improve the process and reduce problems.

This work presents a novel influence analysis method to improve the allocation of development resources, detect process changes, and discover business areas that significantly affect process flow. The method combines the usage of process mining analysis with probability-based objective measures and analysis of deviations. The method is specially designed for business analysts, process owners, line managers, and auditors in large organizations, to be used as a set of interactive root cause analyses and benchmark reports. Methods and algorithms are presented for analyzing both binary problems where each case is either successful or non-successful, and continuous variables, including process lead times and costs. A method for using case-specific weights to consider the relative business importance of each case is also presented. This work also includes data preparation methods and best practices for acquiring relevant business operations data in the event log format.

Concept drift in process mining is a research area that studies business process changes over time. This dissertation shows how process mining can be used to identify changes in business operations by using the influence analysis method to identify business process changes in the business review context. Typical business reviews consist of monitoring key performance indicator (KPI) measures against targets, while the detection of activity level process changes is often based on subjective manual observations alone. Many relevant changes are not detected promptly, making organizations slow to adapt to changes.

Machine learning techniques such as clustering extend the coverage of process mining analyses. A method for clustering cases based on process flow characteristics and using influence analysis to explain the results with business attributes is presented. The method identifies business areas where the process execution differs significantly from the rest of the organization.

Finally, the results of using our methods with publicly available industrial datasets, including service desk data from Rabobank, loan applications process data from a Dutch Financial Institute, and publicly available purchase to pay process data are presented.

Tiivistelmä

 Prosessien jatkuva parantaminen on välttämätöntä jokaiselle organisaatiolle. Prosessilouhinta (process mining) tuottaa tosiasioihin perustuvan tarkan käsityksen operatiivisesta liiketoiminnasta prosessikaavioiden, tunnistettujen pullonkaulojen, vaatimustenmukaisuusongelmien ja muiden prosessihavaintojen muodossa. Prosessien kehittämiseksi ja ongelmien vähentämiseksi organisaatiot tarvitsevat tarkkaa analyysiä havaittujen ongelmien juurisyistä ja menetelmiä kehitysresurssien suuntaamiseen oikein.

 Tämä työ esittelee prosessilouhintaan perustuvan vaikutusanalyysimenetelmän (influence analysis) kehitysresurssien allokoinnin parantamiseksi, prosessimuutosten havaitsemiseksi ja prosessin kulkuun vaikuttavien liiketoiminta-alueiden tunnistamiseksi. Menetelmä yhdistää prosessilouhinnan käyttöä todennäköisyyslaskentaan ja poikkeamien analysointiin. Menetelmä on ensisijaisesti suunnattu suurten organisaatioiden prosessianalyytikoille, prosessien omistajille, operatiiviselle johdolle ja sisäiselle tarkastukselle. Keskeisiä käyttötapoja ovat interaktiiviset juurisyy-analyysit ja vertailuraportit. Esitämme menetelmät ja algoritmit sekä binäärisille ongelmille, joissa jokainen tapaus on joko onnistunut tai epäonnistunut, että jatkuville muuttujille, kuten prosessien läpimenoajat ja kustannukset. Esittelemme myös tapauskohtaisten painotusten käyttöä kunkin tapauksen suhteellisen liiketoimintamerkityksen huomioimiseksi vaikutusanalyyseissä. Lisäksi esittelemme menetelmiä ja kokemuksia tarvittavien lähtötietojen keräämiseen ja esikäsittelyyn.

 Prosessien muutosten seuranta ja analysointi (concept drift) on prosessilouhintaan liittyvä tutkimusalue liiketoimintaprosessien ajallisten muutosten tutkimiseksi. Näytämme miten vaikutusanalyysimenetelmää voi hyödyntää liiketoimintaprosessien muutosten tunnistamiseksi erityisesti liiketoimintakatsausten yhteydessä. Tyypilliset liiketoimintakatsaukset koostuvat pääosin suorituskykymittareiden seurannasta suhteessa tavoitteisiin samalla kun aktiviteettitason prosessimuutosten havaitseminen jää usein pelkästään subjektiivisten havaintojen tasolle. Varsinkin hitaasti etenevät prosessimuutokset havaitaan usein vasta pitkän ajan kuluttua, mikä osaltaan tekee organisaatioista hitaita mukautumaan muutoksiin.

 Koneoppimistekniikat, kuten klusterointi, laajentavat prosessilouhinta-analyysin kattavuutta. Esitämme menetelmän prosessitapausten klusteroimiseksi aktiviteettipolun perusteella ja käytämme vaikutusanalyysiä tulosten selittämiseen liiketoimintakäsitteiden avulla. Menetelmän avulla voidaan helposti havaita sellaiset liiketoiminta-alueet, joissa prosessin kulku poikkeaa muusta organisaatiosta.

 Lopuksi esittelemme tuloksia menetelmien käytöstä julkisesti saatavilla olevien teollisten tietoaineistojen kanssa liittyen IT palvelu-, lainahakemusten käsittely- ja ostoprosesseihin.

Preface

I have been fascinated by artificial intelligence and decision support systems since 1989 when I started to work in the Knowledge Engineering department of Software Technology Laboratory at the Nokia Research Center. Those early-day symbolic artificial intelligence methods were far from perfect. Still, we managed to develop computer programs to analyze data and support decision making. I received my M.Sc with honours in 1995 from the Department of Information Technology at Aalto University, at that time known as Helsinki University of Technology. My M.Sc thesis was titled *Design and implementation of a functional object-oriented application development system*. It was related to a commercial decision support system software MUST Modeller which was used for strategic planning and scenario analysis in many large organizations during 1990 - 2010. After a series of acquisitions and management buy-outs, we merged our company Planway Oy, where I was the Managing Director, to QPR Software Plc in 1999. Since that, I have been focusing on Business Process Management. First ten years in QPR were full of customer projects around Business Process Analysis, Performance Management, Balanced Scorecard, and Strategy Execution. During a QPR innovation workshop meeting on 16th April 2009, we documented our new innovation of using event log data for drawing the flowcharts, and I started to lead the product development project with a code name *QPR Automated Process Bottleneck Remover*. Six months later, we started our first customer pilot projects, and the official QPR ProcessAnalyzer 2.0 product launch for international markets took place on 15th February 2011. Two days later, on 17th February 2011, we met with Professor Olli Simula in Otaniemi at Aalto University premises, signed my Ph.D. student application, and started my academic journey.

This dissertation is a result of four main activity areas. First, delivering value to the actual end customer organizations by selling them the idea of using process mining, and then providing the value using process mining methods. During these years, QPR has conducted around 400 process mining projects, and I have personally been involved in more than 200 customer projects. Second, connecting with the active process

mining academia by reading articles, following presentations, and having open discussions in meetings, conferences, and other events. I have met and talked with around 100 fantastic researchers from process mining academia, which has given me a lot of ideas and different perspectives to process mining. Third, the studies in Aalto University related to data mining, machine learning, graph theories, information visualization, and creative problem solving have all allowed me to learn many exciting and useful skills. Fourth, the software product development activities in QPR have engaged me in decisions regarding the product vision, features, functionalities and prioritization as well as product briefings with external research companies.

My first words of thanks are to Markku Hinkka, my coauthor, a colleague since 1996, and a great friend for these last 24 years. Your software development skills are incredible, covering all aspects related to being a software product architect, scrum master, user interface designer, BIG Data framework expert, and machine learning guru. You master all possible programming languages such as C, C++, C#, Java, Python, SQL Server, Hadoop, Assembly, and Angular, as well as those numerous programming languages we have designed and implemented together during these years. During our Ph.D. journeys, you always had time to discuss process mining and algorithms, share ideas about our joint research papers, revise papers, formalize definitions, evaluate results and plan the next steps. You are the person I have always been able to count on.

To my Ph.D. supervisor, Professor Alexander Jung: thank you for your always encouraging words, your presence, and your trust. You have excellent skills and motivation to drive clarity, and you fully deserve your selection as the teacher of the year 2018 at Aalto University. It was an honor to conduct a joint presentation together in QPR Conference for process mining. I am very grateful that you accepted to become my Ph.D. supervisor.

To my thesis advisor Doctor Jaakko Hollmén: big thank you for guiding me through these years, pointing out exciting research articles and encouraging me to develop my own ideas. You were always positive and managed to find the time when it was needed.

To my coauthor Professor Keijo Heljanko: it was an honor to work with you. I am deeply impressed by your integrity and dedication to research. I was lucky to join the meetings to discuss, formalize, and evaluate the results of machine learning algorithms with you.

To my initial Ph.D. supervisor, Professor Emeritus Olli Simula: thank you for welcoming me to the academic world when I started my Ph.D. journey in 2011 and needed a professor. I am delighted and grateful for that invitation.

To Professor Wil van der Aalst, Distinguished Humboldt Professor at RWTH Aachen University: thank you for being the Godfather of Process Mining and a great source of inspiration to my studies! As I started to do

process mining in 2009, you already had 10+ years of experience in the field. I met you for the first time in the process mining camp in June 2012. You immediately made a huge positive impression on me; the passion you have for process mining is inspiring and always strives to look for more opportunities and benefits. I am deeply honored for the invitation to the Dagstuhl-Seminar 13481 Unleashing operational Process Mining in November 2013. During those five intense days, I had a chance to learn and share my thoughts directly with the best process mining experts, the leading professors, and researchers in the field. You made everyone feel comfortable and facilitated great discussions. You are the most important driver behind the success of process mining and an excellent figurehead for us all.

To my academic process mining friends Professor Massimiliano de Leoni, Professor Marlon Dumas, Professor Felix Mannhardt, Professor Boudewijn Van Dongen, Professor Josep Carmona Vargas, Professor Paolo Ceravolo, Professor Marcello La Rosa, Professor Hajo Reijers, Professor Paulus Torkki, and Doctor Johan Himberg: thank you for the important guidance and motivation you have all given to me by listening to my presentations about influence analysis and sharing your comments and ideas with me.

To my commercial process mining friends, Doctor Anne Rozinat, Doctor Christian W. Günther, Tobias Rother, Doctor Rafael Accorsi, Aurora Sunna, Stewart Wallace, and Doctor Jan Machač: thank you for your feedback related to influence analysis methodology and implemented process mining tool functionalities. Your open communication, sharing of ideas, and hard work to succeed has given me a lot of energy.

To my colleagues in QPR Software Plc, Jari Jaakkola, Matti Erkheikki, Olli Komulainen, Miika Nurminen, Jaakko Riihinen, Olli Vihervuori, Jaakko Niemi, Tuomas Aalto, Jari Luomala, Vesa Kivistö, Polina Hietanen and Antti Manninen: You all have had an essential role in motivating and supporting me to carry on with my work. It has been awesome to see the success we have made together by implementing the functionalities into our QPR ProcessAnalyzer tool and helping customers to use those methods for improving their business processes.

To Peter Selberg, the CEO of Agilon Analytics, Sweden: thank you for the great discussion we had regarding the influence analysis. I had been developing the idea of a continuous version of the method, and you immediately saw the benefits. It was awesome to have you as our first pilot customer for the functionality implemented into our product.

To Rob Van Agteren, Managing Partner at Ackinas, Belgium: thank you for highlighting the importance of working capital management. Your insight supported us in productizing a process mining solution for working capital maangement based on influence analysis.

To Jari Vuori, Matti Ketonen and Outi Aho, Vice Presidents and Directors from Metsä Board, Finland: it was an honor to work with you very actively

during 2014-2015 for deploying process mining and influence analysis, and to take continuous analysis and monitoring into use for supporting your LeanSCM project. I am grateful for your positive attitude, eagerness to reach business results, and willingness to share visionary thoughts about the process mining possibilities.

To Marc Kerremans, Senior Research Director at Gartner Research: thank you for being the open, honest, and energetic person always ready to share thoughts and ideas related to process mining and digital twin of an organization (DTO) technologies. You have been a great companion in my quest for mastering process mining - both from the top-down perspective, including other DTO components, as well as to the practical strategies for boosting commercial process mining business. I am also super proud to have you as my friend with the same hairstyle - as documented in the picture taken during the ICPM 2019 conference after party!

To Henriikka Maikku and Doctor Yrjänä Hynninen: thank you for being the Happy Writers! There are always moments in my life when I question my decisions, objectives, and goals. Regarding the goal of completing my dissertation, I have always had one extra motivator, and that has been you two! From the very moment when we found each other in the excellent course of Carol Kiriakos related to academic writing, I knew that I want to complete this project. It felt so good to be able to send you those draft versions of my articles - big thanks!

To Professor Esa Saarinen: thank you, Esa, for the applied philosophy and creative problem-solving guidance you have given me during my studies at Aalto University. You are the best philosopher I know. Your ideas facilitate my thinking and somehow result in so many positive outcomes that I can only wonder how it is possible. I was so happy to join all the lectures this spring 2020 in the famous Aalto lecture hall A. Good job Esa.

I want to thank Professor Wil van der Aalst, Distinguished Humboldt Professor at RWTH Aachen University and Professor Felix Mannhardt, Norwegian University of Science and Technology (NTNU) for agreeing to act as pre-examiners for this dissertation and Professor Wil van der Aalst also for agreeing to act as an opponent in defense of this dissertation.

To my family: Kirsi, Joonas, Juhani, Jaakko, Mia, Liisa, Jukka, Liina, and Jenni: thank you for your love and support, and thank you for continuously bringing joy, love, smile, positive energy, interesting ideas, exciting discussions and happiness to my life.

Espoo, November 2, 2020,

Teemu Lehto

Contents

List of Publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

I Teemu Lehto, Markku Hinkka, Jaakko Hollmén. Focusing Business Improvements Using Process Mining Based Influence Analysis. In *Business Process Management Forum. BPM 2016.*, Rio de Janeiro, Brazil, pages 177-192, 9 2016.

II Teemu Lehto, Markku Hinkka, Jaakko Hollmén. Focusing Business Process Lead Time Improvements Using Influence Analysis. In *7th International Symposium on Data-Driven Process Discovery and Analysis (SIMPDA 2017)*, Neuchatel, Switzerland, pages 54-67, 12 2017.

III Teemu Lehto, Markku Hinkka, Jaakko Hollmén. Analyzing Business Process Changes Using Influence Analysis. In *8th International Symposium on Data-Driven Process Discovery and Analysis (SIMPDA 2018)*, Seville, Spain, 12 2018.

IV Teemu Lehto, Markku Hinkka. Discovering Business Area Effects To Process Mining Analysis Using Clustering and Influence Analysis. In *23rd International Conference on Business Information Systems (BIS 2020)*, Colorado Springs, USA, 6 2020.

V Markku Hinkka, Teemu Lehto, Keijo Heljanko. Assessing Big Data SQL Frameworks for Analyzing Event Logs. In *24th Euromicro International Conference on Parallel, Distributed, and Network-Based Processing*, Heraklion, Crete, Greece, 101-108, 2 2016.

VI Markku Hinkka, Teemu Lehto, Keijo Heljanko, Alex Jung. Structural Feature Selection for Event Logs. In *Business Process Management Workshops - BPM 2017 International Workshops*, Barcelona, Spain, Revised Papers, volume 308 of Lecture Notes in Business Information Processing, pages 20-35, 9 2017.

VII Markku Hinkka, Teemu Lehto, Keijo Heljanko, Alex Jung. Classifying Process Instances Using Recurrent Neural Networks. In *Business Process Management Workshops - BPM 2018 International Workshops*, Sydney, NSW, Australia, September 9-14, 2018, Revised Papers, volume 342 of Lecture Notes in Business Information Processing, pages 313-324, 9 2018.

VIII Markku Hinkka, Teemu Lehto, Keijo Heljanko. Exploiting Event Log Data-Attributes in RNN Based Prediction. *Lecture Notes in Business Information Processing*, Volume 379, Data-Driven Process Discovery and Analysis 8th and 9th IFIP WG 2.6 International Symposium, SIMPDA 2018 - 2019, Revised Selected Papers, 2020.

Author's Contribution

Publication I: "Focusing Business Improvements Using Process Mining Based Influence Analysis"

The author of this dissertation is the main contributor and responsible for the contents of this publication. The original ideas related to influence analysis methodology, including usage of interestingness measures and different change types, came from the author, who also designed the methods. The experiments and the case study were designed, executed, and reported by the author. Research topic selection and evaluation of results were done together with Markku Hinkka, who also contributed to the design of the change types. Jaakko Hollmén provided guidance, supervision, and comments to the manuscript. The paper was written by the author of this dissertation.

Publication II: "Focusing Business Process Lead Time Improvements Using Influence Analysis"

The author of this dissertation is the main contributor and responsible for the contents of this publication. The original ideas related to using the influence analysis for analyzing lead times to including the usage of continuous variables and case-specific weights, came from the author, who also designed the methods. The experiments and the case study were designed, executed and reported by the author. Research topic selection and evaluation of results were done together with Markku Hinkka, who also contributed to the design of the weighted contribution analysis. Jaakko Hollmén provided guidance, supervision, and comments to the manuscript. The paper was written by the author of this dissertation.

Publication III: "Analyzing Business Process Changes Using Influence Analysis"

The author of this dissertation is the main contributor and responsible for the contents of this publication. The original ideas related to using the influence analysis for analyzing business process changes, including the usage of event-level data instead of case-level data, came from the author, who also designed the methods. The experiments and the case study were designed, executed, and reported by the author. Research topic selection and evaluation of results were done together with Markku Hinkka, who also contributed to the design of the event level influence analysis. Jaakko Hollmén provided guidance, supervision, and comments to the manuscript. The paper was written by the author of this dissertation.

Publication IV: "Discovering Business Area Effects To Process Mining Analysis Using Clustering and Influence Analysis"

The author of this dissertation is the main contributor and responsible for the contents of this publication. The original ideas related to using clustering and influence analysis techniques for analyzing differences in business areas and consolidation of results from multiple clusterings as well as the consolidation of individual business areas to case attribute level, came from the author, who also designed the methods. The experiments and the case study were designed, executed, and reported by the author. Research topic selection and evaluation of results were done together with Markku Hinkka, who also contributed to the design and implementations of the clustering and influence analysis algorithms used in this paper. The paper was written by the author of this dissertation.

Publication V: "Assessing Big Data SQL Frameworks for Analyzing Event Logs"

Markku Hinkka is the main author of this publication. Markku Hinkka, the author of this dissertation and Keijo Heljanko contributed in research topic selection and evaluation of results. The author of this dissertation also contributed to the design of the process mining based Flow and Trace analysis methods.

Publication VI: "Structural Feature Selection for Event Logs"

Markku Hinkka is the main author of this publication. Markku Hinkka,

the author of this dissertation, Keijo Heljanko and Alex Jung contributed in research topic selection and evaluation of results. The author of this dissertation also contributed to the design of the process mining based structural features.

Publication VII: "Classifying Process Instances Using Recurrent Neural Networks"

Markku Hinkka is the main author of this publication. Markku Hinkka, the author of this dissertation, Keijo Heljanko and Alex Jung contributed in research topic selection and evaluation of results. The author of this dissertation also contributed to additional optimizations to shorten input vector lengths and further speed up RNN training times.

Publication VIII: "Exploiting Event Log Data-Attributes in RNN Based Prediction"

Markku Hinkka is the main author of this publication. Markku Hinkka, the author of this dissertation and Keijo Heljanko contributed in research topic selection and evaluation of results. The author of this dissertation also contributed in the design of process mining based activity type specific clustering for event attributes.

List of Figures

List of Tables

Abbreviations

ABPD Automated Business Process Discovery

BPA Business Process Analysis

BPM Business Process Management

BPMN Business Process Model and Notation

DBMS Database Management System

DDBMS Distributed Database Management System

ERP Enterprise Resource Planning

GBM Gradient Boosting Machine

GRU Gated Recurrent Unit

KPI Key Performance Indicator

LSTM Long Short-Term Memory

ML Machine Learning

RDBMS Relational Database Management System

RNN Recurrent Neural Network

ROI Return On Investment

SQL Structured Query Language

XES eXtensible Event Stream

1. Introduction

Organizations face a variety of problems related to their business processes, including long lead times, many process variants, delayed customer deliveries, bad product quality, and unnecessary rework. The problems cause reduced customer satisfaction, loss of business to competitors, high operational costs, and failures to comply with regulations.

The inability to identify root causes for business problems means that business improvements do not target the right issues. This leads to:

1. Increased costs when inefficient operations are not improved and resources are spent on improving things that provide minimal benefit

2. Decreased sales when the constraints for making more sales are not removed

3. Continuing regulatory problems when issues keep repeating.

This dissertation addresses the problem of supporting operational development initiatives in large organizations using process mining. Process mining is a method for discovering and analyzing business processes based on event data. This event data is typically extracted directly from the database tables in enterprise resource planning (ERP) systems or logs in workflow management systems. Based on this data, the as-is version of the process is discovered and presented using flowcharts and other analysis diagrams. Process mining is a fact-based, easy-to-repeat, and accurate method compared to the traditional subjective method of documenting processes based on interviews, discussions, and human interpretation.

One problem when using process mining to support operational development is that business analysts and managers often consider the event type (activity name) information extracted from ERP systems as very technical, too detailed, and unrelated to the everyday business problems. In this dissertation, we show methods for overcoming this problem by using the case and event attribute values that are often better understood by the business people. The usage of attribute values requires advanced methods

for finding the most significant values since, many organizations have thousands of customers, employees, and products.

Another problem is the lack of understanding of root causes for identified problems. Many potential problems can be identified just by looking at the discovered as-is flowchart. However, a detailed root cause analysis is needed to confirm the real reason for the problem in order for the company to succeed in its operative development initiatives. Finding these root causes is the other major topic of this dissertation.

1.1 Objectives and scope

This dissertation provides answers and methods for the following three research questions:

1.1.1 Research Question 1 (RQ1) - How can process mining be used for resource allocation to maximize business improvement?

Business improvements can be achieved by improving the process design or fixing operative issues. The business process re-engineering approach provides methods and tools for developing a better process design and deploying it to all businesses. Alternatively, the problem-solving approach focuses on discovering the current problematic areas where the actual operations deviate from the intended design, finding root causes for the problems, and fixing those operative issues, for example, by delivering training to the right employees or providing better instructions for customers. All identified ideas for improvement should be prioritized based on the potential benefits and required implementation efforts, as shown in Table 1.1.

Table 1.1. Benefit vs. effort matrix for a business improvement

		Potential business benefit	
		small	large
Effort: Resources and time needed to implement the business improvement	small	Good if small improvements are enough	BEST CASE: small investment and large benefits
	large	WORST CASE: large investment and small benefits	Good if large improvements are needed

In order to maximize business improvements, the available development resources must be allocated to those development projects that give the largest benefits with the available resources, as shown in Table 1.1. This dissertation addresses the following aspects:

1. How to use process mining data preparation methods to acquire relevant data of business operations in case and event log format,

2. How to define a binary business problem by categorizing each individual case as problematic or successful. Alternatively, how to define a business problem using a continuous variable like lead-time such that the degree of problems within each individual case is presented as a value in the continuous variable.

3. How to define a variety of potential development projects by setting a scope for each project based on case attribute values.

4. How to combine potential benefits and resource needs into one measure such that development projects can be ordered based on this measure, which is then used to allocate development resources to those projects that maximize business improvement.

Figure 1.1 illustrates the main problem of RQ1 and the key solution for resource allocation. Black circles represent problematic cases, and white circles represent successful cases. The total area represents the whole organization, and the division of the are into smaller different sized rectangles represents the business areas based on distinct values of one case attribute. The figure contains three layers, which represent three different case attributes such as *region*, *product group*, and *customer group*. Our objective now is to find those business areas, in any of the layers, that contain a large amount of problematic cases. More specifically, we search for areas that have a high enough density and high enough absolute amount of problematic cases by ordering all groupings of cases based on all case attributes and their distinct values. Development resources need to be allocated to those identified areas to maximize business improvement benefits. In Figure 1.1, where business areas are represented as different sized rectangles, the upper right corner business area is a good candidate for resource allocation since it meets the two main criteria: 1. The density of problematic cases is high (75%); and 2. The number of problematic cases is sufficiently high, as six out of a total of 60 problems are in this area. The problem setup for RQ1 is presented in Section 2.2.

1.1.2 Research Question 2 (RQ2) - How can process mining be used to identify changes in business operations?

Here our objective is to show what has changed in the process. Changes need to be shown in the order of business significance. An important use

Each layer corresponds to one grouping of cases, for example based on one case attribute.

Each rectangle in a layer represents cases that have the same value for the particular case attribute.

Development resources should be first allocated to this business area because:

1. Density of problems is high (75%) and

2. Number of problematic cases is high enough (6 problems out of total 60 are in this area)

Figure 1.1. Illustrative groupings of cases for detecting problematic business areas.

for this analysis is a periodical business review, for example, a monthly business review, where our analysis is needed to show the most significant changes by comparing the review period against previous historical periods. Another use is called concept drift analysis, where the properties of the process change over time.

Considering RQ2, the analyst person is not initially aware of a business problem in the process. The objective of RQ2 is to detect changes in business operations comparing the *review* period with the *history* period. Changes are ordered by their significance. *Example:* We compare the customer order data from January 2020 against the customer order data from the full year 2019. We find out that in January 2020, a larger amount of *Route Changed* events are taking place compared to historical data from 2019. This change can be considered as a business problem, and the analysis is then continued using the approach described in RQ1 to identify a development project for improving the process.

Figure 1.2 shows a timeline containing process cases, each consisting of individual events related to those cases. Position on the y-axis is used for presenting events belonging to the same case in a horizontal sequence. Each event has a label specifying the activity that took place and the events and connected with arrows to show the transition from one event to the next within the same case. Timeline is divided into a history period and a review period. Detecting changes reveals that during the history period the process typically contained events a, b and c, compared to the

situation in review period when the process more often contains the events a, x and c. Other changes that should be detected include events occurring in a different order, skipping of events, extra events, repeating events, change in the event attribute value of any activity, and change of case attribute value. The problem setup for RQ2 is presented in Section 2.3.
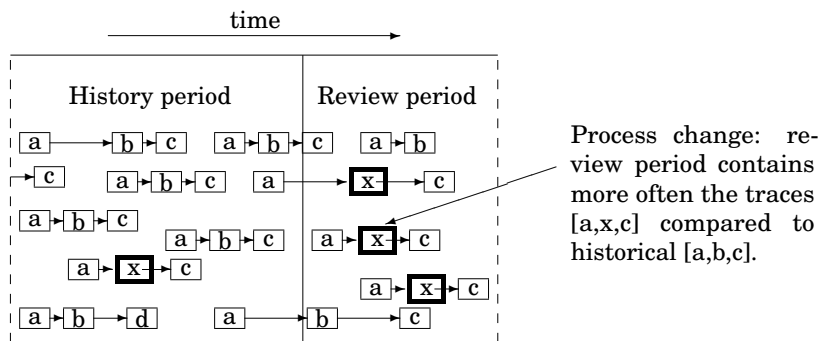
Figure 1.2. Illustrative traces for identifying changes in business operations.

1.1.3 Research Question 3 (RQ3) - How can business areas that have a significant effect on process flow behavior be discovered using clustering?

Our objective is to use clustering on historical data in order to group similar kinds of process instances into the same clusters and then finding the business areas that correlate most with these identified clusters.

Considering RQ3, the business analyst wants to understand what is causing differences in process flow executions. Our approach finds out those business areas that are causing the process flow to be different from other business areas. *Example:* We want to understand better the *order-to-cash* process flow starting from end customer placing an order up to the point of the cash being collected, including production, logistics, and other activities. Using our clustering-based method, we find out that *Route Type* and *Delivery Country* have a big effect on the process flow. Especially we find out that the process flow for cases having *Route Type = Train* is significantly different from cases belonging to the other *Route Types*. With this information, we can now use our Influence Analysis method to discover actual differences in the process flow of *Route Type = Train* cases compared to the other *Route Types* cases in order to find potential business problems. If the *Route Type = Train* cases are entirely different from the other cases, then it may even make sense to analyze those cases separately from each other.

Figure 1.3 shows the results of clustering 129 cases using the process flow information into five clusters. Each case is represented with a marker corresponding to one of the five clusters. The total area represents the

whole organization, and the division of the area into smaller different sized rectangles represents the business areas based on distinct values of one case attribute. The main focus of RQ3 is to discover those business areas that correlate with the clustering results. Figure 1.3 highlights one business area containing six cases, out of which five share the same distinct process flow behavior characterized by a particular cluster. The problem setup for RQ3 is presented in Section 2.4.
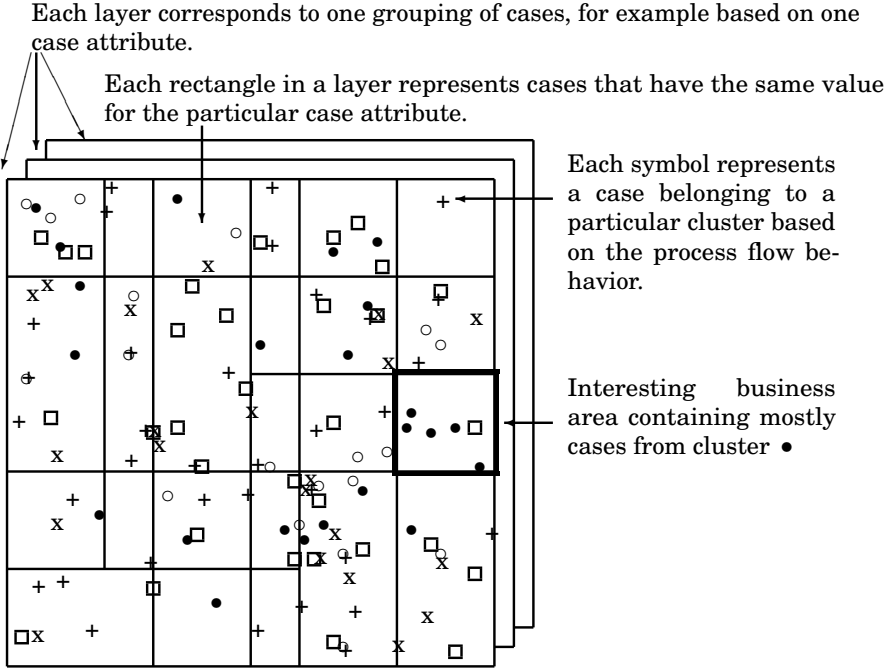
Figure 1.3. Illustrative groupings of cases for discovering business areas with a significant effect on process flow.

Figure 1.4 depicts the dependencies between research questions and corresponding solutions. RQ1 is the main research question introducing the Influence Analysis methodology. RQ2 and RQ3 can be used as standalone questions providing new analysis information, or the results from these research questions can further be investigated using RQ1 to find the best way to focus improvement resources for fixing the identified challenges. Table 1.2 summarizes the Publications related to each research question.
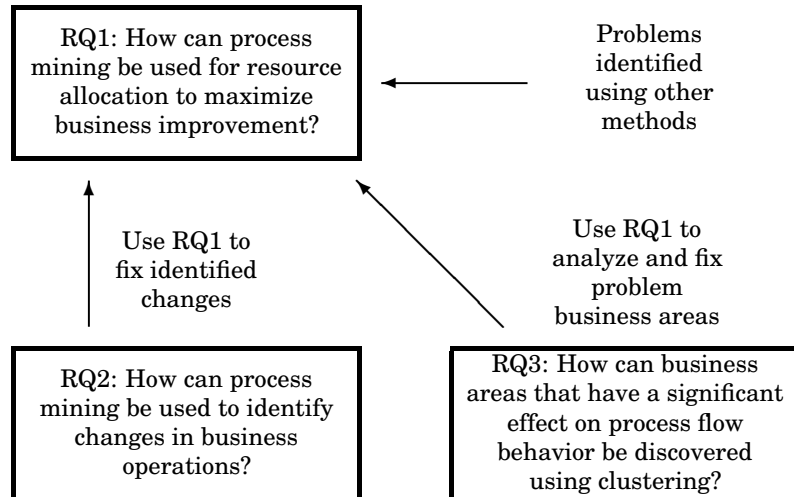
Figure 1.4. Hierarchy of Research Questions

Table 1.2. Relationship between research questions and publications

	Pub. I	Pub. II	Pub. III	Pub. IV	Pub. V	Pub. VI	Pub. VII	Pub. VIII
RQ1: How can process mining be used for resource allocation to maximize business improvement?	X	X			X			
RQ2: How can process mining be used to identify changes in business operations?	X	X	X		X			
RQ3: How can business areas that have a significant effect on process flow behavior be discovered using clustering?	X	X		X	X	X	X	X

1.2 Related work

RQ1: How can process mining be used for resource allocation to maximize business improvement?

With current methodologies, it is difficult, expensive, and time-consuming for organizations to identify the causes of their operational business prob-

lems. One reason for the difficulty is that causality itself is a difficult concept in dynamic business systems [48]. In addition, the theory of constraints highlights the importance of finding the most relevant constraints that limit any system in achieving its goals [29].

The idea of root cause analysis has been widely studied. It includes steps such as problem understanding, problem cause brainstorming, data collection, analysis, cause identification, cause elimination, and solution implementation [6].

Practically all large organizations use business data warehouse and business intelligence systems to store the operational data created during business operations [20], [35]. In 2012 the amount of available data had grown to such an extent that the term Big Data was introduced to highlight new possibilities of data analysis [44]. There are many data mining and statistical analysis techniques that can be used to turn this data into knowledge [49] [51]. There has also been more work carried out in the detection of differences between groups [61] and finding contrast sets [8].

Studies in the field of process mining have highlighted the usage of process mining for business process analysis [2]. Decision tree learning has been used to explain why a certain activity path is chosen within the process, discovering decisions made during the process flow [52]. Decision trees generated from different process variants have also been used to find relevant differences between those process variants [11]. A context-aware framework for analyzing performance characteristics from multiple perspectives has been presented in [33]. Causal nets have been further studied as a tool and notation for process discovery [1]. The approach for detecting cause-effect relations between process characteristics and process performance indicators based on Granger causality is presented in [34]. Our work is partly based on enriching and transforming process-based logs for the purpose of root cause analysis [56]. We also adopt ideas from the framework for correlating business process characteristics [38]. Our work extends the current process mining framework by allowing business users to identify root causes for business problems interactively. Our method is also an example of abductive reasoning that starts with observation and tries to find a hypothesis that accounts for the observation [36].

Probability-based interestingness measures are functions of a 2 x 2 contingency table [49]. Table 1.3 shows the generic representation of a contingency table for a rule $A \rightarrow B$, where $n(AB)$ denotes the amount of cases satisfying both A and B, and N denotes total amount of cases. An example contingency table for a rule $product = hats \rightarrow durationdays \geq 20$ in a database that contains a total of 10 cases such that 3 cases take long time, 4 cases belong to category $hats$, and one case meets both conditions ie. the product delivered is $hats$ and it took a long time is shown in Table 1.4. Summary of 37 different measures with a clear theoretical background and characteristics has been documented in [28]. However, a typical business

person may not be familiar with the measures and may have difficulties understanding the business meaning for each measure. In this dissertation, we will present three probability-based objective measures derived from three business process improvement levels. Business people will be able to decide the level of improvement they are planning to achieve and select a measure based on that level.

Table 1.3. 2 x 2 Contingency table for rule $A \rightarrow B$

	B	$\bar{B}$	
A	$n(AB)$	$n(A\bar{B})$	$n(A)$
$\bar{A}$	$n(\bar{A}B)$	$n(\bar{A}\bar{B})$	$n(\bar{A})$
	$n(B)$	$n(\bar{B})$	N

Table 1.4. Contingency table for rule $product = hats \rightarrow duration \geq 20d$

	B	$\bar{B}$	
A	1	3	4
$\bar{A}$	2	4	6
	3	7	10

Example questions such as "Which customer characteristics are linked to the occurrence of reclamations?" have been presented in a practical framework for process mining analysis uses cases paper [39]. The diagrams presented in the paper show the limitation of the decision tree diagrams as these only show the positive root causes for the given question. One benefit of our method is that it provides the analysis results in the form of comparative benchmarking, showing both the most influential root causes for the long process lead times as well as the most influential best practices for achieving short lead times. The ability to show the root causes for bad and good behavior simultaneously makes it possible to quickly see whether the problem cases have a clear root cause or whether there is a common root cause for cases of good behavior. Our methodology is based on deviations management [50] by discovering significant root causes using process mining data.

Existing studies on root cause analysis and decision tree analysis in the process mining domain are related to binary conditions where each individual case is regarded as good or bad—for example, the decision tree approach presented in [52]. Wetzstein et al. present a framework for monitoring and analyzing influential factors of business process performance [64]. However, their method also requires the usage of a binary contribution measure. Gröger et al. also demonstrate relevant data mining approaches for manufacturing process optimization [30] using binary and decision tree approach. Advantages of using our method include the ability to analyze root causes for continuous variables such as lead times, and the use of case-specific weights for conducting a more business-oriented analysis.

RQ2: How can process mining be used to identify changes in business operations?

Handling concept drift in process mining has been discussed in detail in [12], [13], [17] and [66]. These papers suggest three main problems to be studied: detection of change points, characterization of change, and insight into process evolution. Although these papers are excellent for imparting an understanding of how the process has changed, they share the challenge that they need completed cases and can each be categorized as an offline analysis of change. If cases take six months to complete, the analysis results based on complete cases are at least six months old. However, as part of a business review, the management is reviewing a fixed period and trying to identify relevant change signals as early as possible, hinting how the processes might be changing at this very moment. Using our method, the change point can be set to the beginning of the review period, and our method then shows the most relevant changes that have occurred. Our method can be categorized as online analysis of changes. A somewhat similar approach has been presented in [47] where the complete event log is divided into pre-drift and post-drift logs, and corresponding process models, which are compared to find the minimum set of change operations needed to explain the change behind a drift.

Concept drift in relation to machine learning has been studied extensively in [65], [43], and [62]. The objective of those studies is to increase the accuracy of predictions by utilizing machine learning algorithms that discover the changes in the process. Instead of making accurate predictions, our method is tailored to discover and explain changes as part of the systematic periodical business review.

A novel Trace Clustering algorithm [32] presents an approach to analyze attribute data from events and cases in addition to the traditional business process data. The approach is based on the Markov cluster (MCL) algorithm [25] for finding similar cases. Although the results look promising, the challenge of this approach is that it uses complete cases and is thus more useful for offline process analysis than for periodical business analysis.

An approach more targeted for online business process drift detection is presented in [42]. It uses the concept of partial order runs to run statistical tests in order to find the exact point in time for the change. A somewhat similar method for concept-drift detection in event log streams has been studied in [7], which presents a method for detecting actual concept-drift time and individual anomalies using histograms and clustering. However, these methods do not take into account the attribute data and are not aimed at providing insight into the business review question of what has changed during the current business review period in comparison to previous operations.

Since it is difficult to detect the changes by using only traditional statis-

tical measures, a set of visual analytic tools enabling interactive process analysis and process mining is presented in [16]. The paper presents a visualization for detecting concept drift and changes in business process and case attribute data by plotting all events to a stacked area graph with calendar time on the horizontal axis. Even though the presented tools are useful, they do not provide insight into what has changed during the past review period in the event attribute level. Presented visualization techniques also have challenges when the amount of case attributes is so large that case attributes cannot all be included in the visualizations at the same time.

An alarm-based prescriptive process monitoring for making business people aware of changes that require active intervention is presented in [58]. The method uses a sophisticated cost model to optimize the generation of alerts for business people. One challenge with their method is that it requires a lot of settings and detailed level knowledge of the importance of various process issues. These settings must be fixed beforehand so that the algorithm can then suggest active intervention when needed. Our experience regarding actual business operations is that this kind of detailed information is not typically available, or is challenging to maintain over time. However, our experience is that all large organizations conduct regular business reviews, which makes it beneficial to present the discoveries as part of the business review meetings.

RQ3: How can business areas that have a significant effect on process flow behavior be discovered using clustering?
One key challenge in process mining is that a single event log may often contain many different process variants, in which case trying to discover a single process diagram for the whole log file is not a working solution. In the process mining context, clustering has been widely studied, with excellent results [45], [54], [39], and [59]. This previous work covers the usage of several distance measures like Euclid, Hamming, Jaccard, cosine, Markov chain, edit distance, and several cluster approaches including partitioning, hierarchical, density-based, and neural networks. However, most of the previous research related to clustering within the process mining field has been directly focused on the process flowchart discovery with the prime objectives categorized as process, variant or outlier identification, understandability of complexity, decomposition, or hierarchization. In practice, this means that clustering has been used as a tool for helping the other process mining methods such as control flow discovery to work better—that is, clustering has divided the event log into smaller sub-logs that have been directly used for further analysis. Our approach enables analysts to use clustering for discovering those business areas that have a significant effect on process behavior.

Research has started to address the challenge of how to explain the

clustering results to business analysts [37]. The role of case attributes is important when explaining the characteristics of clusters to business analysts [53]. Our method provides an easy to understand representation of cluster characteristics based on the difference of densities and case attribute information.

Considerable time has also been spent in the process mining community to discover branching conditions from business process execution logs [40]. This has also led to the introduction of decision models and decision mining [9], and to the use of the standard Decision Model and Notation (DMN) for automating operational decision-making [10]. As the objective of decision modeling is to provide additional details into individual branching conditions, the objective of our approach is to analyze the effect of any business area on the whole process flow, not just one decision branch at a time.

Extensive work has been carried out in the area of applying machine learning techniques in process mining. A total of 55 academic papers are listed in a summary paper [26] about predicting the outcome of an ongoing process instance. Discriminating features have been studied as one possible method for feature selection [18]. Discovering signature patterns from event logs has also been studied for predicting desired or undesired behavior [14]. Using event and case attributes to enhance case-level predictions has been studied in [27] and [41]. Specific machine learning algorithms have been studied for predicting the outcome of an ongoing process instance using long-short-term memory (LSTM) [57], gated recurrent unit (GRU) [46], and recurrent neural networks (RNN) [63]. Our work uses these machine learning algorithms to deliver accurate prediction results in shorter execution times, specifically with a high number of case and event attributes.

1.3   Contributions of this Dissertation

This dissertation presents novel methodologies for analyzing business processes using process mining. Its contributions consist of answering the three research questions, which are briefly summarized below.

RQ1: How can process mining be used for resource allocation to maximize business improvement?
Our novel influence analysis methodology is first presented in Publication I as a root cause analysis for business users, helping them to allocate business improvements resources more effectively. Our contributions include methods for collecting event and case attribute information to form new categorization dimensions; methods for forming a binary classification of cases such that each case is either problematic or successful; selection

of a corresponding interestingness measure based on the desired level of business process improvement effect; and the definitions of presented interestingness measures.

One essential element related to the calculation of our contribution measure is the usage of both the density of analyzed problems and the total size of the business area compared to the whole dataset. Our contribution measure is a combination of both these aspects. As the method presented in Publication I is only applicable for binary variables, the Publication II further extends the influence analysis methodology for continuous variables, as well as case-specific weighting. Continuous variables are particularly useful for analyzing business process lead times. Our contributions regarding lead times also include the idea of analyzing the lead time variance—that is, finding the root causes for durations that are too long or too short. We also show how the case-specific weights can be used to create a business-oriented analysis where each case is given a weight based on measures such as monetary value, priority, business importance, profitability, or work effort.

As a special contribution related to working capital optimization, we show how case-specific weights and continuous process lead times can be used to identify those business areas where the largest amounts of extra working capital are tied-up.

Another contribution of our work is that its analysis can be used for benchmarking. Since our contribution measure has a symmetrical behavior, all the business areas included in the root cause analysis get a positive or negative contribution indicator, showing both the problem areas and the best practice areas at the same time. For example, a discovery related to sales orders in a large organization can be used to benchmark the distinct values of the case attribute *Regional Office*. Each regional office gets a positive or negative contribution value such that the sum of all contribution values is always zero.

RQ2: How can process mining be used to identify changes in business operations?

Publication III shows how to use the influence analysis methodology in a novel way to analyze process mining events instead of cases and use event timestamps for identifying changes in the business process over time. Unlike most process mining change detection algorithms, which operate on the case-level, our method analyzes changes in the individual event level. We show how case-level data can be used to construct features for the event level. Our method detects changes in a timely manner, since there is no need to wait for the cases to be completed. We present two alternative methods—a binary approach and a continuous event-age approach—for dividing events into comparison data and review data for business review purposes. Our contributions are related both to preparing the source data

for the analysis, and to calculating the results and presenting to business users.

RQ3: How can business areas that have a significant effect on process flow behavior be discovered using clustering?
Publication IV introduces a novel method of using clustering analysis to discover business areas that have a significant effect on process flow behavior. Our contributions include a method for using process flow data as features for clustering; parameters for the clustering algorithm; the idea of running the clustering several times with different parameters; usage of influence analysis measures for identifying significant business areas; and a method of consolidating results for finding the most significant case attributes. We show how the contribution measure works in practice for explaining clustering results, discovering individual business areas, and finding the case attributes that correlate most with differences in the process flow.

Contributions applicable to all research questions
One generic contribution of this dissertation is the problem setup documented in Chapter 2, which aims to specify common issues and use cases related to analyzing and improving processes. As an additional contribution, we provide case study analyses using real-life industrial data for all research questions in Chapter 4. These analyses show example results and an intended usage scenario for our methods.

Finally, we include a discussion section related to each research question in Chapter 4 to provide our best practices and experiences of using these methods in our process mining projects.

Formal definition of Influence Analysis
Since the Influence Analysis is the main contribution of this dissertation, we present an overall definition of the algorithm already here in the Introduction Section as follows:

Definition 1. *Influence analysis is an algorithm that takes an event log L and its subset L_p as parameters. The result of the algorithm is a set of <feature, feature_value, contribution> tuples, where feature is any property of the cases in L, feature_value is any existing value for the particular feature in the dataset, and contribution is the outcome of the algorithm that measures the significance of the particular feature and feature_value combination for belonging, or not belonging to the subset of cases L_p.*

2. Problem Setup

2.1 Process mining concepts

All research questions addressed in this dissertation use the process mining data as the source data for the methods. We first give definitions for these process mining concepts:

Definition 2. *Let $E = \{e_1, \ldots, e_n\}$, be a set of events in the process analysis.*

Definition 3. *Let $ET = \{et_1, \ldots, et_N\}$ be a set of event types that represent the activity labels for events in the process analysis. Each event is of one event type such that $\#_{EventType}(e_i) \in ET$, for all $e_i \in E$.*

Definition 4. *Let $\#_{TimeStamp}(e_i)$ be the timestamp of the occurrence of event e_i. Timestamp typically represents the date and time when the actual real-life activity took place. Timestamps are required for ordering the events within cases. They are also for calculating lead times and other KPI measures*

Definition 5. *Let $C = \{c_1, \ldots, c_m\}$, be a set of cases in the process analysis. Each individual case represents a single business process execution instance and it has a unique case identifier. Each event belongs to exactly one case such that $\#_{Case}(e_i) \in C$, for all $e_i \in E$. Each case $c_i \in C$ has a trace which is an ordered sequence of events $\#_{Trace}(c_i) = \langle e_1, e_2, \ldots, e_l \rangle$, where $\forall 1 \leq j \leq l : e_j \in E$ and $\forall 2 \leq j \leq l : \#_{TimeStamp}(e_j) \geq \#_{TimeStamp}(e_{j-1})$.*

We will now define the case attributes that contain relevant business data for cases:

Definition 6. *Let $ATC = \{atc_1, \ldots, atc_N\}$ be a set of case attributes in the process analysis. Each case $c_i \in C$ has a value $\#_{atc_j}(c_i)$ for each case attribute $atc_j \in ATC$.*

Definition 7. *Let $V_{atc_j} = \{v_{atc_j}^1, \ldots, v_{atc_j}^N\}$ be a set of distinct values that the case attribute atc_j has in the process analysis.*

Definition 8. *Let* $CaseAttributeSubgrouping(atc_j) = \{C_{v^1}, \ldots, C_{v^N}\}$ *be a subgrouping of all cases* $c_i \in C$ *so that* $\forall c_i \in C_{v^n} : \#_{atc_j}(c_i) = v^n_{atc_j}$.

CaseAttributeSubgrouping for case attribute atc_j allocates all the cases in process analysis into subgroups based on the value of atc_j for each case. The number of subgroups for each case attribute is the number of distinct values for each case attributes. For example, subgroups for the case attribute *Region* could contain *America*, *Europe*, *Asia*, *Africa*, *Other*, and *Unknown*.

We define the event attributes that contain relevant business data for events similarly:

Definition 9. *Let* $ATE = \{ate_1, \ldots, ate_N\}$ *be a set of event attributes in the process analysis. Each event* $e_i \in E$ *has a value* $\#_{ate_j}(e_i)$ *for each event attribute* $ate_j \in ATE$.

Definition 10. *Let* $V_{ate_j} = \{v^1_{ate_j}, \ldots, v^N_{ate_j}\}$ *be a set of distinct values that the event attribute* ate_j *has in the process analysis.*

Definition 11. *Let* $EventAttributeSubgrouping(ate_j) = \{E_{v^1}, \ldots, E_{v^N}\}$ *be a subgrouping of all events* $e_i \in E$ *so that* $\forall e_i \in E_{v^n} : \#_{ate_j}(e_i) = v^n_{ate_j}$.

The summary of these process mining concepts is presented in Figure 2.1. Each concept is linked to the corresponding definition, as presented in this section. Further introduction to process mining can be found in the books "Process Mining - Discovery, Conformance and Enhancement of Business Processes" [4] and "Process Mining - Data Science in Action" [3]

2.1.1 Example data

In this Section, we present a small example process mining data. The chosen business process is the order-to-delivery process, and the case represents one individual order line.

Table 2.1 shows example data containing ten process mining cases and the case attributes *product* and *customer* with values for each case. Table 2.2 contains an event log for each case specifying the activity name and date of the activity. Event *production* also has the name of the country where production was conducted as an event attribute.

Figure 2.2 shows a process model discovered from the example data using a process mining algorithm where:

- The four rectangles *order*, *orderchange*, *production* and *delivery* represent the discovered Event Types (activities) in the Event Log.

- The percentage value inside each rectangle shows the number of cases visiting the particular Event Type.

- The flows between Event Types represent the transitions of a case moving from one event type to another. Each presented flow shows

Table 2.1. Case Data

case	product	customer
1	hats	male
2	hats	female
3	jeans	female
4	shirts	male
5	hats	female
6	shirts	male
7	shirts	male
8	jeans	female
9	shirts	female
10	hats	female

Table 2.2. Event log data

Case	Trace
1	{order(2015-01-01), orderchange(2015-01-07), production(2015-01-15, Ger), delivery(2015-01-19)}
2	{order(2015-01-01), production(2015-01-07, Ger), delivery(2015-01-10)}
3	{order(2015-01-01), orderchange(2015-01-08), production(2015-01-15, Swe), delivery(2015-01-21)}
4	{order(2015-01-01), production(2015-01-12, Fin), delivery(2015-01-13)}
5	{order(2015-01-01), orderchange(2015-01-10), production(2015-01-20, Fin), delivery(2015-01-27), delivery(2015-02-06)}
6	{order(2015-01-01), production(2015-01-08, Ger), delivery(2015-01-13)}
7	{order(2015-01-01), production(2015-01-06, Ger), delivery(2015-01-12)}
8	{order(2015-01-01), production(2015-01-08, Fin), delivery(2015-01-14), delivery(2015-01-22)}
9	{order(2015-01-01), production(2015-01-12, Ger), delivery(2015-01-17)}
10	{order(2015-01-01), production(2015-01-11, Ger), delivery(2015-01-18)}
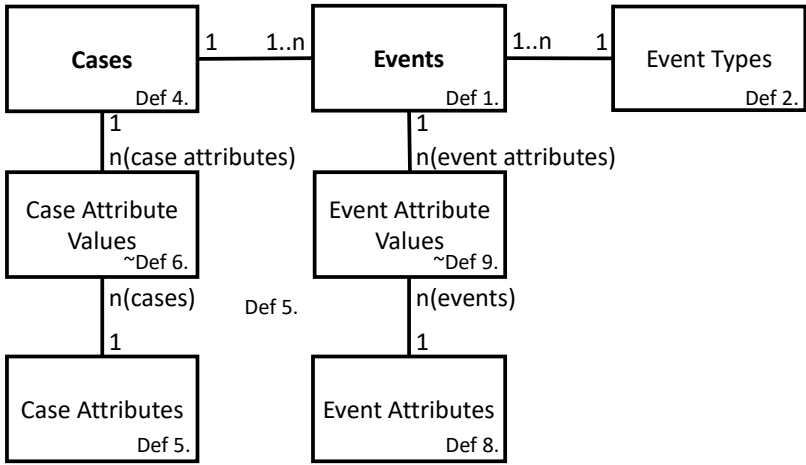
Figure 2.1. Process mining concepts

the volume information (i.e., how many cases contain the transition from source event type to the destination event type) and the average/median duration calculated as the median of the difference of the timestamps.

- Specific start and end symbols represent the beginning and end of the trace.
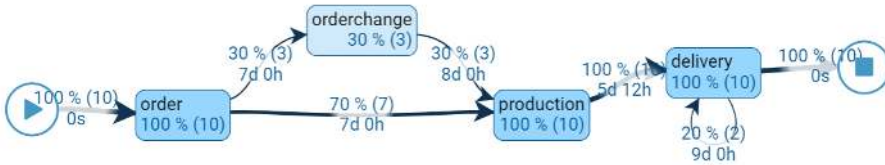
Figure 2.2. Example process flowchart

2.2 RQ1: How can process mining be used for resource allocation to maximize business improvement?

RQ1 is about helping organizations to improve their business operations by providing relevant root causes for discovered process problems. Important

concepts related to RQ1 include:

- **Business Problems.** If the organization/process/system has no problems, then the improvement potential is zero. Process mining operates on the individual process instance level, meaning a system is analyzed based on the individual process instances that have been executed. The analysis uses the event log information containing all recorded process steps with the timestamp information and other event attributes for each event. In this context, the existing process mining algorithms and methods are very efficient in detecting business problems. We want to minimize these problems.

- **Problem Types.** There are two kinds of business problems related to process mining analysis: binary problems, where each case is classified either as problematic or successful; and continuous problems, where each case has a specific numeric value representing the size of the problem. The continuous approach can be used for analyzing lead times where, for example, the actual length of the lead time in minutes, hours, or days can be considered as the size of the problem—the longer the lead time, the bigger the problem.

- **Problem Weighting.** The core idea of Business Process Management (BPM) is to improve business processes by analyzing equally important individual process cases, for example, individual sales orders in the order-to-delivery process. The BPM approach typically tries to find the reason why the process fails to deliver some customer orders on time, using equal importance for each sales order. However, from a business point of view, there are situations when it is beneficial to give specific weighting for individual cases. For example, it would be more important to deliver high-value sales orders in time, making the *order size in currency* a good candidate for case-specific weighting.

- **Development Projects.** Improving real business operations requires development activities to be performed. A development project consists of those activities, aims to reduce problems, and has these properties:

  - *Scope* determines the part of the business that is affected by the development project. The scope for a development project is expressed as a sub-group of cases based on a particular value of a particular case attribute. The characteristics specifying the scope must be known before or, at the latest, during the execution of the case.

  - *Improvement potential* tells how many problems will be fixed by executing the development project.

     – *Investment cost* is an estimate of the amount of resources needed to get the results.

    Finding the best scope for a development project is often tricky since it affects both the improvement potential and investment costs.

- **Amount of resources.** Organizations have an annual minimum amount of development resources available for operational improvement. Additional resources including extra personnel, allocation of time from business operations, purchase of external services, and investments are available when the management accepts an improvement plan with a high return on investment (ROI).

- **Objective.** The aim is to identify those development projects that fix the largest number of problems with the smallest amount of resources.

Here are the formal definitions to these needed concepts:

Definition 12. *Let P_0 be the initial volume of problems in business operations before any development project. If P_0 is zero, then there are no problems in the business operations, and it would be impossible to find any development project that could improve the situation.*

Definition 13. *Let $D = \{d_1, \ldots, d_N\}$ be a set of Development Projects.*

Definition 14. *Let P_{d_i} be the total volume of problems in business operations after the execution of development project $d_i \in D$. P_{d_i} may be smaller or bigger than the initial volume of problems P_0.*

Definition 15. *Let $potential(d_i) = P_0 - P_{d_i}$ be the improvement potential of the development project d_i.*

Definition 16. *Let $cost(d_i) \in \mathbb{R}_{\geq 0}$ be the cost of the development project d_i. According to the assumption that cost of a development project is proportional to scope of the development project, the cost of a development project d_i affecting a subset of cases $C_j \in C$ would be $cost(d_i) \propto n(C_j)$.*

Return on Investment (ROI) is often used to compare the efficiencies of several different investments such as development projects. Using the *potential − costs* as the benefit (or return) of the development project, the generic definition for RQ1 is formulated as finding the development project with highest ROI as:

Definition 17. *Let $RQ1 = \mathrm{argmax}_{d_i \in D} ROI(d_i) = \mathrm{argmax}_{d_i \in D} \frac{potential(d_i) - cost(d_i)}{cost(d_i)}$ be the generic formalization for the question "Identify the development projects with the highest development potential and lowest costs."*

2.3   RQ2: How can process mining be used to identify changes in business operations?

The ability to detect changes is crucial for developing and improving agile business operations. Unwanted changes need to be mitigated quickly, and desired changes need to be reinforced and shared as best practices. During a business review, managers typically review the performance of business operations using Key Performance Indicators (KPIs). One problem is that managers typically do not have an accurate fact-based understanding and analysis of what has changed during the review period. They often rely on subjective comments, views, and suggestions influenced by acute business challenges and crises. There are various details for business managers and analysts to consider:

- **Setting the review period.** Business managers should easily see what has changed during the review period by comparing the new process mining data against data from previous business review periods. For example, a manager or internal auditor who has been away from business operations for a week, month, or year can easily find out what has been changed during that period compared to the previous historical data. During an acute crisis like the coronavirus pandemic [68], process owners and line managers would benefit from an analysis showing how the process execution has changed on a daily level by comparing yesterday's process execution data against the previous month.

- **Fast and slow changes.** Both the fast changes occurring suddenly, as well as the more gradual changes that occur in several years, should be detected to give managers accurate insight into the current situation, changes, and trends.

- **Relevancy.** Identified business process changes should be prioritized based on the combination of the relative amount of the changed cases as well as the absolute scope of the change in relation to the whole process execution. A relatively small change in the context of an activity that occurs in all cases is more important than a major change related to an activity that occurs only once per year.

- **Root causes.** In addition to identifying the relevant changes, it should be possible to find possible root causes for the identified changes. Also it should be possible to show or predict the effect and outcome of each change in relation to a particular business outcome or KPI measure.

- **Data quality detection.** Process mining and KPI reporting rely on continuously updating data, typically the ERP systems. The minimum requirement for the change detection system is that it should

automatically detect major problems in these data integrations. For
example, if the incremental loading of a database table containing
customer delivery timestamps fails for one import, the magnitude of
change should be so high that this issue is reporting very high in the
list of process changes. The change detection method should always
be automatically used as a quality assurance feature for automated
periodical data loads.

Here are the formal definitions to clarify the problem setup:

Definition 18. *Let $A = \{\alpha_1, \ldots, \alpha_N\}$ be a set of process mining event charac-
teristics. The characteristic α_i may be any particular event attribute value,
a combination of several event attribute values, or any other property that
can be defined for an event.*

As an example, if all events have an event attribute *PerformedBy* and
one possible value for this attribute is *User1234*, then the events having
the characteristic α_i have *PerformedBy=User1234*.

Definition 19. *Let $E_{\alpha_i} = \{e^1_{\alpha_i}, \ldots, e^N_{\alpha_i}\}$ be a set of events sharing the same
characteristic α_i. $E_{\alpha_i} \subseteq E$.*

Following the example above, the set of events E_{α_i} where α_i is *Per-
formedBy=User1234* would include all the events that have been *Per-
formedBy* the User *User1234*. The key idea in detecting changes is to
analyze the distribution of events E_{α_i} in relation to the whole set of events.
Since our goal is to identify changes over time, we define a review period
as:

Definition 20. *Let $E_r = \{e^1_r, \ldots, e^N_r\}$ be a set of events belonging to the Review
Period r. $E_r \subseteq E$.*

As a measure for the business significance of any change, we are inter-
ested in the difference in the density of the particular characteristic in the
review period compared to the whole dataset. Here is the generic definition
for RQ2:

Definition 21. *Let $RQ2 = \arg\max_{\alpha_i \in A} abs(\frac{|E_r \cap E_{\alpha_i}|}{|E_r|} - \frac{|E_{\alpha_i}|}{|E|})$ be the generic for-
malization to the question "Identify those process mining event characteris-
tics that are most significantly unevenly distributed over time."*

2.4 RQ3: How can business areas that have a significant effect on process flow behavior be discovered using clustering?

It is often challenging to communicate the detailed process mining findings
to business people. While line managers may not be familiar with process

flow details, they are very interested to learn about differences within the organization's business areas. RQ3 serves as a bridge between BPM and business people, facilitating the knowledge sharing between these groups.

Generic options for analyzing a business process with many process flow variations and exceptions include:

(a) Analyze all business areas separately. This can lead to a large amount of extra work.

(b) Analyze all business areas at the same time. This can result in potentially meaningless results.

(c) Rely on subjective personal information, such as asking business people or merely using intuition to decide which of the business areas should be analyzed separately.

(d) Use the method described in this study for finding those areas of business where the process flow is different to other areas.

Specifically, in order to discover the effects of different business areas on process flow, it is necessary to understand:

- How a business process can be analyzed based on the process flow of individual process instances in order to discover business-relevant clusters in such a way that a business analyst can easily understand the clustering results and use them for further analysis.

- How to find business areas that have a significant effect on process flow behavior.

- How to further consolidate business area results to discover case attributes that have a significant effect on process flow behavior.

Here are the formal definitions for the problem setup:

Definition 22. *Let $B = \{\beta_1, \ldots, \beta_N\}$ be a set of process mining case characteristics. The characteristic β_i may be any particular case attribute value, a combination of several case attribute values or any other property that can be defined for a case.*

As an example, all cases may have a case attribute *Region*, one possible value for this attribute could be *Europe*, and the corresponding characteristic β_i would then be *Region=Europe*.

Definition 23. *Let $C_{\beta_i} = \{c_{\beta_i}^1, \ldots, c_{\beta_i}^N\}$ be a set of cases sharing the same characteristic β_i. $C_{\beta_i} \subseteq C$.*

For example, the set of cases C_{β_i} where β_i is *Region=Europa* would include all the cases whose *Region* is *Europa*.

Definition 24. *Let $P = \{p_1, \ldots, p_N\}$ be a set of clusters each formed by clustering the cases in C using process flow characteristics.*

Definition 25. *Let $CL_p = \{c_{p_1}, \ldots, c_{p_N}\}$ be a set of cases belonging to cluster p. $CL_p \subseteq C$.*

As a measure for the business significance of the business area denoted by the corresponding case characteristic, we are interested in the difference in the density of the particular characteristic in cases belonging to each cluster compared to all cases in the whole dataset. Here is the generic definition for RQ3:

Definition 26. *Let $RQ3 = \arg\max_{\beta_i \in B} \sum_{p \in P} abs(\frac{\left|CL_p \cap C_{\beta_i}\right|}{\left|CL_p\right|} - \frac{\left|C_{\beta_i}\right|}{\left|C\right|})$ be the generic formalization to the question "Identify those process mining case characteristics that are most significantly unevenly distributed in the several representative clustering runs based on process flow characteristics."*

3. Methods

This Chapter describes the approach and methods for solving the problems presented in the previous Chapter 2. Section 3.1 describes the influence analysis methodology, which is used for calculating the interestingness measures needed for root cause analysis, analyzing process changes, and discovering business area effects. Influence analysis serves as the core foundation for this dissertation. It is first presented as a binary version for process mining cases only in Publication I. Section 3.2 extends the influence analysis method to continuous problem functions needed, for example, in process lead time analysis. The section also shows how case-specific weights can be used to take into account the business importance of each individual case as presented in Publication II. Section 3.3 shows how the influence analysis method can be used on the process event level to analyze process changes as presented in Publication III. Finally, Section 3.4 presents our method for discovering business area effects to process mining analysis using clustering based on influence analysis published in Publication IV utilizing the findings in Publication VI, Publication VII, and Publication VIII. Section 3.5 contains our study of the Big Data SQL Frameworks for using distributed computing techniques and frameworks to execute process mining tasks as published in Publication V.

3.1 Influence analysis methodology

3.1.1 Identifying the relevant business process

The first task is to identify a high-level problem in the operations. After identifying the problem, we pinpoint the business process where cases will be classified as successful or problematic based on whether they experienced the problem or not. This is a non-trivial task, and making a wrong decision leads to inaccurate results. For example, when analyzing root causes for failures to deliver customer orders on time, the case can

be selected several ways, such as complete customer order in ERP, individual customer order line in ERP, or individual delivery order in ERP. A complex customer order may include several order lines with various different products and services, each of which may be delivered on several delivery orders according to the agreed schedules. To analyze the business relevant KPI (Key Performance Indicator) *OnTimeDelivery*, it is possible to be interested in:

- the whole customer order to be completely delivered on time,

- each individual item within the order being delivered on time, or

- each agreed delivery of each order item being completed on time.

It is the job of the analyst to select the correct granularity level for the analysis based on the actual business problem that is investigated.

3.1.2   Collecting event and case attribute information

The accuracy of our analysis depends on the amount of data available in event and case logs. Since our goal is to create new insight for business people, we encourage the use of all event and case attribute data available, even though that typically introduces a lot of noise and data that is not always relevant to the problems being analyzed. The generation of suitable log files with extended attributes is a well-studied area [38]. Methods also exist for enriching and aggregating event logs to case logs [56]. These are the key steps for constructing event and case logs:

- Identifying the relational database table whose rows correspond to cases C. In object-oriented systems, this corresponds to identifying the object class whose instances correspond to cases C.

- Finding for each case c_i in C, a set of objects O_i such that every object o_{ij} in O_i is linked to c_i directly. Then add recursively all objects linked to o_{ij} as long as the objects seem to be relevant concerning the analysis objectives. Note that since every table in a relational database is typically linked to all the other tables in the same database recursively, this leads to potentially thousands of relevant linked objects for each case.

- Forming the event log for c_i by including one event for every timestamp attribute of the case c_i and any linked object o_{ij}.

- Forming the case log for c_i by aggregating all attribute values of c_i and every object o_{ij} in O_i, thus creating potentially thousands of case attributes for each case. Suitable aggregation functions include *count*, *sum*, *max*, *min*, *average*, *median*, *concatenate*, *first*, and *last*.

- Augmenting every case c_i by adding external events that have occurred during the lifetime of the case, for example, *machinebreak-started*, *machinebreak-completed*, *weekend*, *strike*, *queuetoolong*, and *badweather*.

3.1.3 Categorization dimensions for cases

The purpose of this step is to create new categorization dimensions for all cases. These dimensions will then be used when identifying the best improvement focus areas, so the more dimensions we have, the larger the coverage of our analysis will be. Table 3.1 shows examples of dimensions that can be created for every event log based on the log itself.

Table 3.1. Illustrative category dimensions for cases

Category dimension	Usage for analyses
The number of events per case	Cases with many events may be complex and contain much rework. Cases with only a few events may be incomplete.
The number of distinct activities per case	Cases containing a large number of different activities have a greater variety of processing than the straightforward cases.
Start and end timestamps of the whole case	Exact calendar date, month, week, and year can be used to detect process changes over the time. *Day of the week* and *month of the year* are useful for discovering periodic and seasonal behavior.
Start and end time of an individual event type	Same rationale as the case-level attribute above. Usage of these measures will create at least one new dimension for each analyzed event type.
Activity profile. The number of event occurrences per each event type	Often the fact that a particular event is executed several times for a case is a root cause for business problems.
Transition profile. The number of flow occurrences per each transition from one event type to the next.	Usage of the transition profile makes it possible to identify root causes caused by performing the activities in a certain order.

  Our method is used to analyze and compare individual cases against each other in order to find subsets of cases that can be used as a scope for process improvement activities. A generic framework for correlating different process characteristics has been presented in [24] where individual event log characteristics like decision points, events, and resources correlate with identified process problems. One important aspect of our method is to communicate the results to business analysts and decision-makers who are not familiar with process execution details but instead are very familiar

with business entities like organization units, product groups, customers, and suppliers. For this reason, we use feature selection for creating new dimensions to the case level.

Structural feature selection

Publication Publication VI includes a more detailed analysis of selecting structural process mining features as categorization dimensions for prediction purposes. The objective is to extract and condense the process flow information into case level features. Challenge is that the number of structural features easily grows high, making the analysis difficult. This is a real risk for the prediction purpose, since the resulting machine learning model may become overfitting and thus fails to generalize the predictions. For the root cause analysis, this is not such a big problem, since we are indeed trying to find causes for something that has already happened. Many of the structural features correlate with each other, such as missing an activity always results in missing all transitions where the current activity would be present, resulting in multiple root cause items for the same process issues.

3.1.4 Forming a binary classification of cases such that each case is either problematic or successful

The purpose of this step is to find a binary classification expression that specifies whether a particular case is problematic or successful. In practice, a wide range of process mining methods can be used to make process discoveries as described in the process mining manifesto [2]. Typical business problems discovered using process mining methods include

- problems in the process execution, such as missing activities, extra processing steps, processing in the wrong order.

- performance issues, such as long lead-times for certain transitions, service-level agreement breaches

- quality issues, such as end customers returning the products, repetitions and rework

- compliance issues, such as breaches in the 4-eye principle where the same person approves the purchase requisition and the purchase order

Table 3.2 shows some example business problems that have been discovered using process mining methods and the corresponding illustrative binary classification expressions.

Table 3.2. Illustrative discovered business problems with corresponding binary classifications expressions

Business problem	Binary classification expression
Problematic cases are not completed within the agreed service level agreement 7 days	$c.totalduration() > 7days$
Problematic cases should not include multiple 'AddressChanged' activities	$c.activitycount('AddressChange') > 1$
March 2015 was a problematic month	$c.startmonth() =' 2015 - 03'$
First AddressChanged event should not be recorded by John	$c.getActivity('AddressChanged').first().recordedBy() =' John'$
In problematic cases the *StartProduction* is done directly after *ReceiveOrderSize*, there should be *CreditCheck* first	$c.flowcount('ReceiveOrder', 'StartProduction') > 0$

3.1.5 Selecting a corresponding interestingness measure based on the desired level of business process improvement effect

In this step, we select which interestingness measure will be used to find the best business improvement areas. Requirements for the interestingness measure include the following:

1. *Easy to understand by business people.* Business people make actual decisions based on the analysis results, so they must understand the results. It is essential to minimize the magic and maximize the clarity of analysis.

2. *Big benefits.* The selected interestingness measure should identify areas that include many problematic cases. This requirement corresponds to the benefit dimension in Table 1.1.

3. *Small effort.* Implementing the change should require only a small amount of resources. This requirement corresponds to the effort dimension in Table 1.1.

Regarding the first requirement of being *easy to understand by business people*, we have identified three corresponding target levels for operational business improvements that business people are familiar with:

1. *Ideal.* The improvement project will be ideal, all problems of the identified type will be removed, and all future cases will be completed without any of these problems.

2. *Other average.* The focus area can be improved so that it reaches the current average performance of other areas. After the improvement

project, the share of problematic cases in the focus area will be equal to the average share of problematic cases in the other business areas before the improvements.

3. *As-is average.* The focus area can be improved so that it reaches the current average performance of all areas. After the improvement project, the share of problematic cases in the focus area will be equal to the average share of problematic cases in the whole business before the improvements.

Regarding the second requirement, *big benefits*, we calculate the overall density of problematic cases after the improvement. Table 3.3 shows these overall density measures calculated for the three identified change types when A is the set of cases selected as a target for business process improvement, B is the set of problematic cases before improvement, and B' is the set of problematic cases after improvement.

Table 3.3. Change types

Change type	To-be density of problematic cases for the selected segment A after the change $P(B'|A)$	Overall to-be density of problematic cases after the change $P(B') = P(B'|A)P(A) + P(B|\bar{A})P(\bar{A})$	Change in overall density of problematic cases $P(B') - P(B)$
ideal	Zero density $= P(\emptyset)$	$P(\emptyset)P(A) + P(B|\bar{A})P(\bar{A}) = P(B) - P(AB)$	$-P(AB)$
other average	Average of current cases excluding this segment $= P(B|\bar{A})$	$P(B|\bar{A})P(A) + P(B|\bar{A})P(\bar{A}) = P(B|\bar{A})$	$P(B|\bar{A}) - P(B)$
as-is average	Average of current cases including this segment $= P(B)$	$P(B)P(A) + P(B|\bar{A})P(\bar{A}) = P(A)P(B) + P(B) - P(AB)$	$P(A)P(B) - P(AB)$

Regarding the third requirement, *small effort*, we say that the effort needed to improve a segment is relational to the size of the segment $P(A)$)—that is, the larger the segment, the more effort is needed to make improvement.

Table 3.4 summarizes the identified change types according to the three requirements. Change type *ideal* sorts the results by the number of problematic cases, thus maximizing benefits. Since it does not take into account the size of the segment, it performs poorly against the small effort requirement. Change type *other average* performs well regarding the benefits, but

it fails to make a difference between different sized segments, including all problematic cases. It is also difficult for business people to understand since the benefit potential of each segment is related to the average performance of all other segments, which needs to be realized separately for each segment. Change type *as-is average* performs well regarding the benefits, is easy enough to understand for business people, and takes into account the cost needed to implement the change. As shown in Table 3.4, we have given equal evaluation for the benefits for change types *other average* and *as-is average* for the volume of benefits since both approaches provide moderately high benefits for the development projects even as the *other average* yields slightly higher benefits than *as-is average*.

Table 3.4. Change types by requirements

Change type	Easy to understand	Big benefits	Small effort to achieve
Ideal	+ + +	+ + +	-
Other average	+	+ +	+ +
As-is average	+ +	+ +	+ + +

Based on Table 3.4, we propose to use the change type *as-is average* as the target level for operational business improvements. We thus select the corresponding interestingness measure from Table 3.3 as $P(AB) - P(A)P(B)$, which is also known as $Leverage(A \rightarrow B)$. Business meaning of this measure is that if the segment specified covered by the antecedent of a rule is improved so that it reaches average performance, then the change in the total density of problematic cases is reduced by $P(AB) - P(A)P(B)$.

Let B be a set of problematic cases and A be a set of cases that will be improved in order to reach an as-is-average density of problematic cases. For influence analysis we define the following measures.

Definition 27. *Let $Contribution(A \rightarrow B) = n(AB) - \frac{n(A)n(B)}{N}$, where $n(AB)$ is the number of problematic cases in segment A before improvement, $n(A)$ is the number of cases in segment A, $n(B)$ is original amount of problematic cases, and N is total number of cases. This measure tells how many cases will be improved when business improvement is focused on segment A.*

Definition 28. *Let $Contribution\%(A \rightarrow B) = Contribution(A \rightarrow B)/n(B)$, where $n(B)$ is the amount of problematic cases before business improvement. This measure tells how big share of the total business problem is improved when business improvement is focused on segment A.*

Contribution is an interestingness measure that can now be used to answer our research question 1 as formally defined in Definition 17 as:

$RQ1 = \mathrm{argmax}_{d_i \in D} ROI(d_i) = \mathrm{argmax}_{A,B} Contribution(A \rightarrow B)$, where A is the scope of the development project specified as any addressable subset of

cases, and B is the set of problematic cases before improvement. Other measures typically used in associative rule mining include support: $supp(X) = P(X)$ to be used as frequency constraint, confidence: $conf(X \rightarrow Y) = P(Y|X)$ to be used for measuring probability and lift: $lift(X \rightarrow Y) = \frac{P(X \cap Y)}{P(X)P(Y)}$ to measure dependency [15]. The *Contribution* measure can be seen as a combination of support, lift, and confidence such that a high contribution can only be achieved when segment A is large enough, and the density of problems in that segment A is higher than on average.

Definition 29. *Let* $AttributeContribution\%(At \rightarrow B) =$
$\frac{1}{2}\sum_{A_i \in AttributeValues(At)} Abs(Contribution\%(A_i \rightarrow B))$, where
AttributeValues(At) is the set of all the sets of cases such that each individual set of cases contains all the cases having one specific attribute value for At. AttributeValues(At) has thus one set of cases for every separate value for At. The AttributeContribution% measure shows the correlation between the set of problematic cases B and the particular case attribute At on a scale of 0 to 1. A strong correlation indicates a high potential for business process improvement, while a low correlation closer to zero indicates lower improvement potential. The division by 2 is used to ensure that AttributeContribution% is always between 0 and 100%.

Attribute contribution is used to identify case attributes that contribute most to the finding. If there are large differences in the distribution of problematic cases for the different values of At, then the attribute contribution for At is high. If attribute contribution is low for attribute At, then we know that At does not include relevant causes for the problematic cases.

3.1.6   Finding the best categorization rules and attributes

The analysis is performed by running a rule learning algorithm for a set of rules $A \rightarrow B$ where B is the binary classification value using the information defined in previous steps. The results show how much the overall density of problematic cases changes when a selected business change is targeted to the segment covered by the antecedent A of the rule.

Based on the author's empirical evaluation, we have used very straightforward rules where the antecedent A consists of only one-dimensional attribute (=Case Attribute) and one exact category value (=Value for the particular case attribute). Usage of a simple antecedent makes the analysis easy to understand for business analysts and allows them to further continue the analysis by limiting the analysis scope to the cases included in antecedent A and discovering the next antecedent A' using other available case attributes as dimensions. This dissertation uses a brute force rule mining algorithm that simply calculates the selected interestingness measure *contribution%* separately for all possible antecedents. This approach

provides complete benchmarking results for all individual case attributes. If the dataset is huge and performance becomes an issue, the authors recommend conducting the analysis first with a random sample to get approximate results fast.

It would also be possible to construct antecedents based on multiple conditional attributes and use available association rule learning algorithms, such as Apriori [5] and Eclat [67], to find combinations with a high contribution. However, when antecedents may contain multiple attributes, the amount of possible rules grows in a very high combinatorial growth rate. Holte [31] has also presented the idea that simple rules perform very well on most business datasets.

3.1.7 Presenting the results

A full influence analysis report shows all discovered rules sorted by the selected interestingness measure. The top of the list contains the problematic cases (=best improvement areas), and the bottom of the list contains the best practice examples.

Large dimensionality is typically a big problem when finding causes from several thousand or more features. Our methodology solves this during the presentation step by only showing a fixed amount of top and bottom rules. For example, an analysis may contain 1 000 dimensions with a total of 100 million distinct single dimension antecedents. Our suggestion is to only show, for example, the top 100 and bottom 100 antecedents. In this way, the interesting dimensions are likely to have at least some values in the top or bottom ranges, and the user can continue checking that attribute in more detail.

Another possibility is to show the report first only for the dimensions. In the previous example, where we have 1 000 dimensions, we first show them ordered by the *AttributeContribution%*, and the user then selects one attribute for more details.

The influence analysis report for one attribute shows the antecedents for one case attribute at a time. Based on the author's empirical evaluation, this kind of view is beneficial for business people allowing them to benchmark the problem and best practice areas for the selected case attribute and validate results.

3.1.8 Example analysis

We show a quick example of influence analysis results with the example data defined in Section 2.1.1. Comprehensive analyses with real-life data are shown in Chapter 4. Table 3.5 shows new categorization dimensions that have been derived from source data. *Duration days* is the total duration of the case in days. *#Del* represents the number of events of type

delivery occurring during the case. The *Region* is the value of *production country* taken from the events of type *production*. *Weekday* represents the day of the week when the *production* event was conducted. *#Order changes* represents the number of events of type *order change* occurring in the case and *Trace* is the full event type sequence of the whole case

Table 3.5. Example derived case data

case	durat. days	#del	region	week-day	#order changes	trace
1	18	1	Ger	Fri	1	order-orderchange-production-delivery
2	9	1	Ger	Thu	0	order-production-delivery
3	20	1	Swe	Fri	1	order-orderchange-production-delivery
4	12	1	Fin	Tue	0	order-production-delivery
5	36	2	Fin	Wed	1	order-orderchange-production-delivery-delivery
6	12	1	Ger	Fri	0	order-production-delivery
7	11	1	Ger	Wed	0	order-production-delivery
8	21	2	Fin	Fri	0	order-production-delivery-delivery
9	16	1	Ger	Tue	0	order-production-delivery
10	17	1	Ger	Mon	0	order-production-delivery

Problematic cases are identified with a binary classification B such that $B = true$ if $duration days \geq 20$ else *false*. With this classification the cases 3, 5, and 8 have $B = true$, so the original density of problematic cases is $P(B) = 3/10 = 0.3$ Table 3.6 shows the rule mining algorithm results for each of the presented three change types: *as-is average*, *other average* and *ideal*. Results are sorted by the change type *as-is average* effects. n(A) is the total number of cases meeting the Antecedent criteria in the first column, and n(AB) is the number of those problematic cases that also meet the problem rule criteria $'duration days \geq 20'$. Δn shows the change in the number of problematic cases for each change type and $\Delta P(B)$ shows the change in the density of problematic cases. According to these results, the business improvement efforts should be targeted to segments # *deliveries = 2* and *product = jeans*, since in both of these segments the number of problematic cases will drop by 1.4 as shown in the column $\Delta_3 n$.

Table 3.6. Contribution values for problem $'durationdays \geq 20'$

Antecedent	n(A)	n(AB)	ideal		average		as-is avg	
			Δn	$\Delta P(B)$	Δn	$\Delta P(B)$	Δn	$\Delta P(B)$
# deliveries = 2	2	2	-2	-0.2	-1.75	-0.18	-1.4	-0.14
product = jeans	2	2	-2	-0.2	-1.75	-0.18	-1.4	-0.14
customer = female	6	3	-3	-0.3	-3	-0.3	-1.2	-0.12
# order changes = 1	3	2	-2	-0.2	-1.57	-0.16	-1.1	-0.11
Region = Finland	3	2	-2	-0.2	-1.57	-0.16	-1.1	-0.11
ProdDay = Fri	4	2	-2	-0.2	-1.33	-0.13	-0.8	-0.08
Region = Sweden	1	1	-1	-0.1	-0.78	-0.08	-0.7	-0.07
trace = order-orderchange-production-delivery-delivery	1	1	-1	-0.1	-0.78	-0.08	-0.7	-0.07
trace = order-production-delivery-delivery	1	1	-1	-0.1	-0.78	-0.08	-0.7	-0.07
ProdDay = Wed	2	1	-1	-0.1	-0.5	-0.05	-0.4	-0.04
trace = order-orderchange-production-delivery	2	1	-1	-0.1	-0.5	-0.05	-0.4	-0.04
product = hats	4	1	-1	-0.1	0.33	0.03	0.2	0.02
ProdDay = Mon	1	0	0	0	0.33	0.03	0.3	0.03
ProdDay = Thu	1	0	0	0	0.33	0.03	0.3	0.03
ProdDay = Tue	2	0	0	0	0.75	0.08	0.6	0.06
# order changes = 0	7	1	-1	-0.1	3.67	0.37	1.1	0.11
customer = male	4	0	0	0	2	0.2	1.2	0.12
product = shirts	4	0	0	0	2	0.2	1.2	0.12
# deliveries = 1	8	1	-1	-0.1	7	0.7	1.4	0.14
Region = Germany	6	0	0	0	4.5	0.45	1.8	0.18
trace = order-production-delivery	6	0	0	0	4.5	0.45	1.8	0.18

3.2   Analysis types for influence analysis

As pointed out in the problem setup for RQ1 in Section 2.2, business problems can be categorized as binary or continuous. It is also possible to use case-specific weights. In this section, we extend our influence analysis method to accommodate these problem types by introducing the four different analysis types *binary contribution (BiCo)*, *continuous contribution (CoCo)*, *weighted binary contribution (wBiCo)* and *weighted continuous contribution (wCoCo)* shown in Table 3.7. Depending on the problem type, the contribution formula can be either binary or continuous. Depending on the relative importance of cases, the contribution formula can be weighted or not. In many business process analysis situations, the actual business problem can often be formulated with any of these four formulas. Since the formulas give potentially different results, it is essential to understand that seemingly small differences in formulating the problem may lead to substantial differences in the analysis results. Thus it is often beneficial to use multiple contribution formulas for double-checking that suggested business process improvement areas are correct.

Table 3.7. Analysis types

		Problem type	
		Binary	Continuous
Case weights	Equal weights	Binary contribution (*BiCo*)	Continuous contribution (*CoCo*)
	Different weights	Weighted binary contribution (*wBiCo*)	Weighted continuous contribution (*wCoCo*)

Business problem size

The calculating of the problem size before and after the development project is not trivial. Using the given data, we have identified four different kinds of problem sizes, as summarized in Table 3.8. Development resources should be allocated to improving issues where the problem size is large, and the size of the required investment is small. Problem size and an example lead time process for each contribution formula are shown in Table 3.8. When considering business process lead times, we typically want to make the process generally faster (continuous variable), or want to ensure that the lead time of each instance is shorter than a given target (binary variable). The continuous approach is useful when faster performance is desirable without any lower bound. The binary approach is appropriate

when, for example, each process instance is categorized as successful if it meets a service level agreement (SLA) and unsuccessful if it exceeds SLA. Following the principles of power-law distributions in empirical data [19], we can use the binary approach by selecting around 20% of the worst performing cases to find explanations for bad performance.

Table 3.8. Problem size and example lead time for analysis types

Analysis type	Problem size	Example lead time
BiCo	Number of problematic cases	In the service desk process, a lead time longer than seven days may be considered as a problem case.
wBiCo	Sum of the value of problematic cases	Free-of-charge pizza if delivery takes more than 45 minutes. Problem size is equal to the monetary value of late pizza deliveries.
CoCo	Sum of positive overtime compared to average lead time	Lead time from the customer calling helpdesk to the moment the call is answered. The shorter the lead time, the better.
wCoCo	Sum of overtime for each case compared to the weighted average lead time multiplied by the weight separately for each case	Lead time from sending an invoice to the moment the payment arrives. When this lead time is multiplied by the value of the invoice, we get the working capital, i.e., using the value of the invoice as the weight for each case.

Common Definitions
Here we present the common definitions used in all contribution formulas based on previous definitions.

Definition 30. *Let $C_p = \{c_{p_1}, \ldots, c_{p_N}\}$ be a set of problematic cases. $C_p \subseteq C$.*

Definition 31. *Let $C_a = \{c_{a_1}, \ldots, c_{a_N}\}$ be a set of cases belonging to business process improvement segment A. $C_a \subseteq C$.*

Definition 32. *Let d_{c_j} be the duration of the case c_j.*

Definition 33. *Let w_{c_j} be the weight of the case c_j. We consider linear weights so that double weight always means double importance. If $w_{c_j} = 0$ then case c_j will have no effect in the analysis when calculating weighted results.*

Definition 34. *Let pr be the problem size in the original situation before any business process improvement: BiCo: the number of problem cases, wBiCo: the sum of weights of problem cases, CoCo: the sum of overtime compared to average duration, wCoCo: the sum of overtime per case multiplied with the weight of the case compared to the weighted average duration.*

The summary of contribution formula definitions is presented in Table 3.9. Formulas are explained in more detail after the table.

Table 3.9. Problem size, average function and contribution% definitions

Type	Total Problem size	Average function	Average function for subset C_a	*Contribution%*
BiCo	$pr_{BiCo} =$ $$\sum_{c_j \in C_p} 1 \quad (3.1)$$	Average Problem density: $$\rho = \frac{\sum_{c_j \in C_p} 1}{\sum_{c_j \in C} 1} \quad (3.2)$$	$$\rho_a = \frac{\sum_{c_j \in (C_p \cap C_a)} 1}{\sum_{c_j \in C_a} 1} \quad (3.3)$$	$$\frac{(\rho_a - \rho) \sum_{c_j \in C_a} 1}{pr_{BiCo}} \quad (3.4)$$
wBiCo	$pr_{wBiCo} =$ $$\sum_{c_j \in C_p} w_{c_j} \quad (3.5)$$	Weighted Average Problem density: $$\rho_w = \frac{\sum_{c_j \in C_p} w_{c_j}}{\sum_{c_j \in C} w_{c_j}} \quad (3.6)$$	$$\rho_{w_a} = \frac{\sum_{c_j \in (C_p \cap C_a)} w_{c_j}}{\sum_{c_j \in C_a} w_{c_j}} \quad (3.7)$$	$$\frac{(\rho_{w_a} - \rho_w) \sum_{c_j \in C_a} w_{c_j}}{pr_{wBiCo}} \quad (3.8)$$
CoCo	$pr_{CoCo} =$ $$\frac{1}{2} \sum_{c_j \in C} \left| d_{c_j} - \bar{d} \right| \quad (3.9)$$	Average lead time: $$\bar{d} = \frac{\sum_{c_j \in C} d_{c_j}}{\sum_{c_j \in C} 1} \quad (3.10)$$	$$\bar{d}_a = \frac{\sum_{c_j \in C_a} d_{c_j}}{\sum_{c_j \in C_a} 1} \quad (3.11)$$	$$\frac{(\bar{d}_a - \bar{d}) \sum_{c_j \in C_a} 1}{pr_{CoCo}} \quad (3.12)$$
wCoCo	$pr_{wCoCo} =$ $$\frac{1}{2} \sum_{c_j \in C} w_{c_j} \left| d_{c_j} - \bar{d}_w \right| \quad (3.13)$$	Weighted Average lead time: $$\bar{d}_w = \frac{\sum_{c_j \in C} w_{c_j} d_{c_j}}{\sum_{c_j \in C} w_{c_j}} \quad (3.14)$$	$$\bar{d}_{w_a} = \frac{\sum_{c_j \in C_a} w_{c_j} d_{c_j}}{\sum_{c_j \in C_a} w_{c_j}} \quad (3.15)$$	$$\frac{(\bar{d}_{w_a} - \bar{d}_w) \sum_{c_j \in C_a} w_{c_j}}{pr_{wCoCo}} \quad (3.16)$$

Binary contribution - BiCo

For binary contribution, the problem size is the number of problematic cases. Every case needs to be classified as problematic or successful. Definitions for *BiCo* have already been presented in Section 3.1. However, we now present a new style for the definitions to be used for all analysis types.

The total problem size for *BiCo* is the number of problematic cases $pr_{BiCo} = |C_p| = \sum_{c_j \in C_p} 1$ as shown in equation 3.1 in Table 3.9. Average function for *BiCo* is the average problem density $\rho = \frac{|C_p|}{|C|} = \frac{\sum_{c_j \in C_p} 1}{\sum_{c_j \in C} 1}$ as shown in equation 3.2. Similarly, the average problem density for *BiCo* of subset C_a is $\rho_a = \frac{|C_p \cap C_a|}{|C_a|} = \frac{\sum_{c_j \in (C_p \cap C_a)} 1}{\sum_{c_j \in C_a} 1}$ as shown in equation 3.3. Finally the *Contribution%* for *BiCo* of subset C_a is $con_{BiCo} = \frac{(\rho_a - \rho) \sum_{c_j \in C_a} 1}{pr_{BiCo}} = \frac{|C_p \cap C_a|}{|C_p|} - \frac{|C_a|}{|C|} = \frac{\sum_{c_j \in (C_p \cap C_a)} 1}{\sum_{c_j \in C_p} 1} - \frac{\sum_{c_j \in C_a} 1}{\sum_{c_j \in C} 1}$ as shown in equation 3.4

Weighted binary contribution - wBiCo

wBiCo extends the previous sigma-based formulas by replacing the static equal weight with case-specific weights w_{c_j}. Problem size as defined in equation 3.5 in Table 3.9 is the sum of weights of all problem cases. Average problem density as defined in equation 3.6 in Table 3.9 is the sum of weights of all problem cases divided by the sum of weights of all cases, and correspondingly the average problem density in equation 3.6 in Table 3.9 is the sum of weights of all problem cases in subset C_a divided by the sum of weights of all cases in subset C_a.

Continuous contribution - CoCo

CoCo allows analyzing the root causes of continuous problems without the need to have a binary value for each case. When analyzing the lead time as a continuous problem, the cases are no longer separated into two binary categories of just long and short cases. For continuous analysis, each case is considered problematic if the value of the continuous target variable is bigger than the average for all cases, as shown in equation 3.9 in Table 3.9. The bigger the positive difference is, the worse the behavior. Conversely, if the continuous target value is less than average, then the case is better than average. Using this approach, the sum of positive deviations is always the same as the absolute value of the sum of negative deviations, meaning that the problem size for the whole population C is always zero. The problem size of any subset C_a may be nonzero, meaning that the cases in subset C_a either have higher or lower values for the target variable than the whole population. For analyzing lead times, the

continuous target variable is any lead time variable of the business process cases—for example, the total end-to-end lead time or any partial lead time from one activity to another. The average function for continuous analysis types is the average lead time, which is defined for the whole population with equation 3.10 and subset using equation 3.11 in Table 3.9.

The contribution measure for each possible subset C_a for *CoCo* is calculated as follows: subtract the average lead time of whole population C from the average lead time of the subset C_a, and multiply this by the number of cases in subset C_a. This gives an absolute value of how much more or less time is spent on the cases in subset C_a as a total compared to the average. The final step is to divide this figure by the problem size—that is, by the total sum of positive (or negative) cases in the population, giving the definition for equation 3.12 in Table 3.9.

Weighted continuous contribution - wCoCo
In this Section, we extend the previously defined continuous contribution formulas to support case-specific weights. Using lead time as an example the the average weighted lead time using case-specific weights is calculated with equation 3.14 in Table 3.9, where the lead time of each case is multiplied by the case-specific weight, and the result is divided by the total sum of weights. This weighted lead time is then used to calculate the total problem size according to equation 3.13 in Table 3.9 so that the absolute difference of lead time for each case is multiplied by the case-specific weight and then summed up.

The calculations for *wCoCo* analysis are similar to those for non-weighted analysis: subtract the weighted average duration of the subset C_a from the total weighted average and multiply this by the sum of weights in subset C_a. Dividing this by the total problem size gives the amount of weighted time that would be saved if the lead times for cases C_a could be reduced to the average weighted lead time in the whole population C as shown in equation 3.16 in Table 3.9.

It is notable that *wCoCo* corresponds to the working capital needed in a business process when the weight used for each case is the amount of capital tied-up to each case during the transition representing the lead time. As an example, let us consider the process of building houses where each case is one house. Working capital needed is proportional to the total cost of each house and the lead time from starting the construction to selling the house. *wCoCo* gives this measure when the cost of the house is used as the case-specific weight, and the building time is used as the lead time. The business improvement activity for reducing working capital for this construction company then corresponds to conducting influence analysis using weighted continuous contribution analysis type to identify those subsets that should be the focus for process improvements. In business terms, this corresponds to calculating the extra working capital (positive)

or unneeded working capital (negative) for each case and then summing them together. According to our approach, if the lead time for every case is equally long, then the problem size is zero, and there is no extra working capital in the process.

Strengths and weaknesses of analysis types
Deciding which analysis type should be used in a particular business process analysis situation is not a trivial decision. Table 3.10 shows the strengths and weaknesses for binary and continuous analysis types and Table 3.11 respectively for weights. It is often desirable to select one analysis type as the baseline for each specific business analysis and then use the other analyses types for reviewing, double-checking, and confirming results from different perspectives.

Table 3.10. Strengths and weaknesses of analysis types

Type	Strengths	Weaknesses
Binary	• Can be applied to every lead time problem by separating cases into good cases and bad cases based on a selected cut-off threshold. • Can be controlled by setting the cut-of threshold for duration. • Manages outliers very well because every case is just considered good or bad and the amount of extra lead time is not considered at all.	• Requires decision for the cut-off threshold. If a customer of the process would like the result in ten days and average duration currently is six days, should we consider all cases taking more than ten days as bad, or should the cut-off threshold be nine days in order to improve cases that are close to being missed; or should the threshold be 20 days to allow identification of areas that have severe problems? Influence analysis gives potentially different results with every cut-off threshold.
Continuous	• Does not need any separate cut-off threshold. Continuous variables such as lead time are used directly by the algorithm and overtime is calculated from the average duration.	• Is sensitive for outliers. If one case takes a million times longer than the other cases, then the whole analysis is likely to suggest improvement in all the subsets C_a containing that particular case.

Table 3.11. Strengths and weaknesses for using weights

Type	Strengths	Weaknesses
Equal weights	- No need to define and calculate weights. - Every case simply has equal weight. - Not sensitive to outliers regarding weights.	It is often a major problem to lose a large customer, fail the *service level agreement* of an important customer request, or have quality issues with expensive products. Using equal weights takes customer importance and order size into account.
Different weights	Using the sales value, profit, importance or similar as weights makes results more aligned with business value and importance.	- It is not easy or straightforward to define the weight for each case. Some cases may have small or even zero value, but they could be part of an important project for a major customer. - Using weights makes the analysis sensitive to outliers.

3.3   Analyzing business process changes

In this Section, we provide a method for answering RQ2. We explain how to analyze the variance and deviations in business processes on the individual transaction level to discover and explain changes that have taken place. Our objective is to find areas that have more variance compared to average areas. If there are no changes in the business operations, then the data in the ERP system for the review period is similar to the data for the previous periods. Conversely, if there are changes, then the data will be different from the past data. Our method for analyzing business process changes is based on the influence analysis methodology presented in Sections 3.1 and 3.2 with two major exceptions: 1. The analysis is done on the event level, meaning that feature selection and creation of new categorial dimensions are made for events. 2. Instead of problematic vs. non-problematic cases, the analysis is done over time comparing new and old events. The following sections illustrate the main steps in our method.

3.3.1   Identifying the relevant business process

Our approach detects changes from one business process at a time. A large organization with multiple processes needs to run the analysis separately for each business process to detect the changes in all business operations.

Typically, business reviews are based on consolidated data. For example, a dashboard report can contain several KPIs. The ERP systems in large organizations can easily contain 1 billion new database-level transactions (i.e., database rows) per month. A traditional business review could contain 10 KPIs with 100 consolidated drill-down measures each. Only 0.0001% of the available data, i.e., 1 000 consolidated data elements derived from 1 billion transaction-level data elements, would be available for making findings in business review. Using our method to set up ten process mining models containing an average of 1 million transaction-level events per business review period of 1 month would make 1% of the data available for analysis, i.e., 10 million events from 1 billion transactions. In this example, our approach would give 10 000 times more data available in business review as potential root cause elements than the approach based on consolidated KPI data. We recommend organizations to analyze as many processes as possible and include as many events as possible in order to get a comprehensive view of the changes in business operations. We also suggest that the data to be prepared so that it covers as much of the end-to-end processes as possible in order to facilitate identifying root causes for the discovered process changes.

3.3.2   Collecting event and case attribute information

Typical process mining analysis consolidates data from the event level to the case level. In our method, we further copy the consolidated case-level data into each event occurring in the case, which gives the analysis an augmented set of attributes for each event. After collecting the data by identifying cases, events, and their attributes according to Section 3.1.2, conduct the following steps to prepare the data for analysis done on the event level:

- Use the properties of each event e_i in E as event attributes.

- Identify for each event e_i a corresponding case c_i and copy all case attributes as additional event attributes.

- Form an event path for each event e_i by concatenating the event type names of events linked to the same case, sorted from oldest to newest. The event path can be expressed in many ways—for example, as a single event attribute containing the full path, as several attributes containing single predecessor values or as the full activity and transition profiles copied from the case level to each event up to the point of the event occurrence time stamp.

- Identify for each event e_i in E a set of objects O_i such that every object o_{ij} in O_i is linked to e_i. Use the properties of objects o_{ij} as additional event attributes for events e_i.

- Augment every event e_i by adding external events that have occurred at the same time. Examples of external events include machine-break, weekend, strike, queue-too-long, and bad-weather. Adding external events makes it possible to use this same approach for detecting changes in external circumstances as well.

3.3.3 Categorization dimensions for events

The purpose of this step is to create new categorization dimensions for the events. This method is otherwise similar to the Categorization dimensions for cases in Section 3.1.3, except that the dimensions are formed on the event level. All these dimensions can then be used for detecting the changes, so the more dimensions there are, the more extensive the coverage of the analysis will be. Table 3.12 shows examples of dimensions that can be created for every event log based on the log itself.

Table 3.12. Illustrative categorization dimensions for events

Dimension	Amount	Dimension identifier	Value
Event types	One	"Event type name"	Event type name
Case attributes	One for each case attribute	"CA1:" + Case attribute name	Case attribute value
Case attributes by event type	One for each combination of event type and case attribute	"CA2:" + Event type name + Case attribute name	Case attribute value
Event attributes	One for each event attribute	"EA1:" + Event attribute name	Event attribute value
Event attributes by event type	One for each combination of event type and event attribute	"EA2:" + Event type name + Event attribute name	Event attribute value + Event type name
Predecessor name	One	"Predecessor1"	Predecessor event type name
Predecessor name by event type	One for each event type	"Predecessor2:" + Event type name	Predecessor event type name
Process path	One	"Path1"	Full event type path including the event itself
Process path by event type	One for each event type	"Path2:" + Event type name	Full predecessor event type path without the event itself

Categorization dimensions form the bases for influence analysis when discovering the business process changes. Categorization dimensions are needed to discover any root causes for changes. Having the *event types* dimension enables us to detect changes in the amounts of particular event types. For example, we could find out that there were more *ontime delivery* events and fewer *customer complaint* events during the review period as compared to the comparison period. The *case attributes* dimension in Table 3.12 is often very useful since it allows the detection of changes in the background data of active cases. For example, in November, there could

have been more events from *Region* with value *Finland* compared to the previous six months. Case attribute changes may be analyzed as specific to particular event types using the event type name in the dimension identifier or as global case attributes without the event type name, or both. Similarly, all the dimensions in Table 3.12 can be added to the analysis. The total number of dimensions—that is, feature vectors for analysis—can easily grow large if all the dimensions are to be used in the analysis. For example, with 30 event types, 50 case attributes, and 10 event attributes, the total amount of dimensions from Table 3.12 would be *1 + 50 + 1500 + 10 + 300 + 1 + 30 + 1 + 30 = 1 923*. Our influence analysis method works well with this high dimensional data since it only shows those dimensions where the changes are most significant. If the performance becomes an issue, it is possible to select only those dimensions that seem to be essential for review purposes. The benefit of this is that business people are not overloaded with data that they cannot understand, and the risk is that some relevant root causes are not reported. Advanced feature selection algorithms provide possibilities for limiting the number of features. We have studied this in Publication VI.

3.3.4   Defining data for review and comparison periods

In Sections 3.1 and 3.2, we discovered root causes for already known business problems. In order to identify changes, we will now define the business problem component of influence analysis in a different way specifying for each event, whether it belongs to the historical period or the review period.

The change window divides the time into different parts: the review period and the comparison period. The review period is the new data corresponding to the new events in the business operations. The comparison period is the old data containing old events. To identify changes, we need to identify the differences in the events in the review period versus the comparison period.

Binary approach
Figure 3.1 shows how the analysis data is divided into four different periods in order to identify process changes.

- **Review period (c)**. All events occurring in this period are taken into consideration when discovering changes. If these events, their quantities, and event attributes are similar to the comparison period events, then there are no significant changes. In real life, something is always changing, so our target is to detect the most important changes. An example review period could be one calendar month.

- **Comparison period (b)**. All events occurring in this period are also taken into consideration when discovering changes. A typical
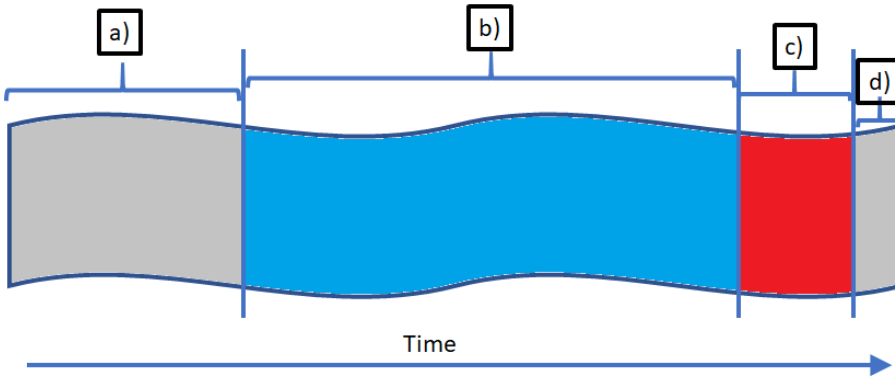
Figure 3.1. Business review periods using fixed periods

setup would be to use the six months prior to the review period as a comparison period and one month as the review period. If a business is very seasonal, then one option is to use a year-to-year comparison period so that the comparison period can be compared to the same month in the previous year.

- **History period (a)**. The events that occurred during the history period are not used as separate events in the review or comparison sets. However, these events should be used for constructing the business process path (trace) for each event in both review and comparison periods. For example, the review and comparison periods both contain events *OntimeDeliveryFailed*. In order to understand the root causes of these failures, we want to include a full process path for each *OntimeDeliveryFailed* event so that we can analyze the difference in the activities leading to the *OntimeDeliveryFailed* process step. For this reason, the predecessor events from the history period need to be used when constructing this path for review and comparison period events.

- **Most recent data period (d)**. All events occurring after the review period are excluded from the analysis. As an example, the typical business review for *November* is done in early *December* when the data from *November* is complete. We do not want to use the recent data from *December* as it becomes available because that data will be analyzed in next month's business review. It is also possible to set up a review period such as the *last 30 days*, so that the review data contains all recent data from the last 30 days, while the *most recent data* period would then be empty.

The benefit of using a binary approach is that it is typically easy to use for business people who have prior knowledge of the operations for both review period and comparison period. The binary approach also

guarantees that all discovered changes have indeed taken place during the well-defined review period, since the *history* and the *most recent data* periods are excluded from the analysis results.

Continuous approach
Another approach for defining the review and comparison periods is to use a continuous measure to determine to which period any particular event belongs. For example, an ad-hoc analysis can be performed by a process analyst using the continuous approach, in order to see how the process has generally changed over time.

Figure 3.2 shows an example of how the analysis data can be divided into four different periods using the continuous approach:
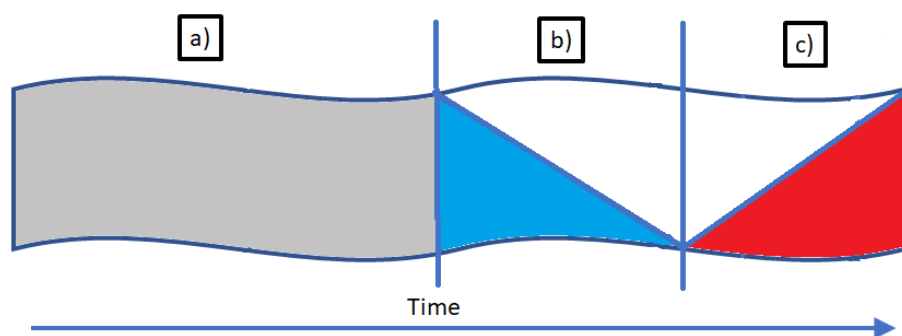
Figure 3.2. Business review periods using continuous periods

- Use half of the available data for the history period (a). This ensures that all analyzed events have a proper history and consistent set of predecessor events. To be exact, one can limit the history of each analyzed event in comparison period (b) and review period (c) to be exactly the length of the history period (a), starting from the actual timestamp of the occurrence of the analyzed event.

- The other half of the data can then be used as a 50% comparison period (b) and 50% review (c) period. This approach gives a neat 50% ratio, so that for each dimension, event type, and analysis finding, there should be an equal number of events in both the comparison and review periods.

- To detect changes that have occurred in the long run, the oldest and newest events can be given higher weight than the events that took place when the comparison period ended and the review period started. One way to achieve this is to calculate an *age* attribute for each event. *Age* would be equal to the elapsed time between the actual time of the event and the time of the most recent data refresh

to our analysis. The *age* attribute can then be used as the weighting measure for continuous contribution formulas, as defined in Table 3.9. In practice, the continuous approach using *age* gives the most significant weight for the events that take place at the beginning of the comparison period and at the end of the review period. Events that take place in the middle have only a small weight, so this is a particularly good approach for analyzing small gradual changes that occur over a long time.

3.3.5 Detecting changes

Definition 35. *Let $E_a = \{e_{a_1}, \ldots, e_{a_N}\}$ be a set of events sharing the same characteristics as defined in segment A. $E_a \subseteq E$. These characteristics are derived from different values for the Categorization Dimensions.*

Definition 36. *Let $E_p = \{e_{p_1}, \ldots, e_{p_N}\}$ be a set of Review Period events. $E_p \subseteq E$ to be used in the Binary Approach.*

Definition 37. *Let d_{e_j} be the age of the event e_j. Age is calculated as the difference between a reference timestamp t_0 and $\#_{TimeStamp}(e_j)$. Reference timestamp can be any fixed timestamp as long as it is used for all events. Typical reference timestamp values include the runtime timestamp of the influence analysis algorithm and the latest data extraction timestamp. Age is used in the Continuous Contribution approach by comparing each events' Age to the average Age.*

Business problem size
Business Problem Size is an important parameter for the Influence Analysis, and it has been explained in Section 3.2 and defined in Definition 34. Table 3.8 contains typical examples of problem size. However, the definition and usage of problem size is slightly different when detecting process changes. For process change detection using *Binary Contribution*, the problem size is the total amount of events in the *Review Period*, and our analysis identifies root causes for the changes in the density of certain kind of events in *Review Period* compared to the *Comparison Period*. Similarly, for the Continuous Contribution, the problem size is calculated using the Age - see Definition 37 - of each event occurrence as described below.

Binary change window
In Binary Change Window analysis, each event is either included in the set of Review Period events or the set of Comparison period events. Note that events belonging to the *History* period and the *Most Recent Data period* have already been excluded from the analysis.

Converting the formulas from Table 3.9 from cases to events, the total problem size for BiCo is the number of events in the review period

$pr_{BinaryChange} = |E_p| = \sum_{e_j \in E_p} 1$ derived from Equation 3.1. Average function

for BiCo is the density of events with property a, i.e., the events that belong to the set E_a compared to the total amount of events in E as follows: $\rho = \frac{|E_a|}{|E|} = \frac{\sum_{e_j \in E_a} 1}{\sum_{e_j \in E} 1}$ derived from Equation 3.2. Similarly, the problem

density for BiCo of subset E_a is the density of events belonging to the set E_a within the review period events E_p as $\rho_a = \frac{|E_p \cap E_a|}{|E_p|} = \frac{\sum_{e_j \in (E_p \cap E_a)} 1}{\sum_{e_j \in E_p} 1}$ de-

rived from Equation 3.3. Finally the *Contribution%* for BiCo of subset E_a

is $con_{BiCo} = \frac{|E_p \cap E_a|}{|E_p|} - \frac{|E_a|}{|E|} = \frac{\sum_{e_j \in (E_p \cap E_a)} 1}{\sum_{e_j \in E_p} 1} - \frac{\sum_{e_j \in E_a} 1}{\sum_{e_j \in E} 1}$ derived from Equation 3.4 in

Table 3.9.

Continuous change window

For Continuous Change Window analysis, the formulas from Table 3.9 are written as: Total problem size for CoCo is the sum of the distance between *Age* and average *Age* for each event separately $pr_{ContinuousChange} = \frac{1}{2} \sum_{e_j \in E} |d_{e_j} - \bar{d}|$ derived from Equation 3.9. The average function for CoCo is

the average age $\rho = \bar{d} = \frac{\sum_{e_j \in E} d_{e_j}}{\sum_{e_j \in E} 1}$ derived from Equation 3.10. Similarly the

average problem density for CoCo of subset E_a is $\rho_a = \bar{d}_a = \frac{\sum_{e_j \in E_a} d_{e_j}}{\sum_{e_j \in E_a} 1}$ derived

from Equation 3.11. Finally the *Contribution%* for CoCo of subset E_a is

$con_{CoCo} = \frac{(\bar{d}_a - \bar{d}) \sum_{e_j \in E_a} 1}{pr_{CoCo}}$ derived from Equation 3.12 in Table 3.9.

Both binary and continuous methods give the possibility of using weights for individual events. For example, events occurring during the past 12 months may have standard weights, and events occurring 13-24 months ago may have slightly lower weights. Alternatively, event types may be prioritized to give the most important events higher weights.

3.4 Discovering business area effects

In this section, we present our methodology for discovering business area effects to process mining analysis using clustering and influence analysis as published in Publication IV. Our approach is to first cluster the cases using available process flow features, and then conduct influence analysis using case attribute data to identify those business areas that have the highest contribution with clustering results. If all values for any particular case attribute are distributed randomly to all clusters, the contribution measure

for each corresponding business area is very small, and the information for the analyst is that the particular case attribute does not correlate with process flow variations. According to our methodology, this means that the particular case attribute has no influence on the process flow behavior. In summary, our method uncovers those business areas and case attributes that have the highest contribution to the process flow behavior.

3.4.1 Clustering cases

Clustering is an unsupervised machine learning method used to group similar objects into same clusters. Clustering can be done by using a variety of data, and different clustering algorithms with their specific parameters.

To identify those business areas that have the most substantial effect on the process execution, we first perform clustering using relevant features representing the process execution characteristics. These features have been widely studied in trace clustering papers [54] and [59]. We discuss feature selection in more detail in Publication VI and Publication VII. Clustering is a trade-off between quality and performance. As the amount of features increases, the quality of the results potentially improves to a certain degree while clustering takes more time.

Feature selection
The methods presented in Section 3.4 rely on clustering as the unsupervised learning method. We have studied the usage of process mining results as features for machine learning algorithms in Publication VI. The aim of the methods presented in Publication VI is to find a set of features that should be used as the baseline data for machine learning algorithms when making predictions regarding the outcome of individual cases. By comparing the seven different feature selection algorithms - Random, Blanket, Cluster, ClusterImp, LASSO1se, mRMREns5, and Rec2S, with five different structural feature types (*activity*, *transition*, *starter*, *finisher*, and *ordering*) we come into the conclusion that the Cluster algorithm used with activity occurrence counts within a process mining case yielded the best results for classification.

Feature selection results are shown in Figure 3.3. Best accuracy was achieved by using activity profile (*act*) together with sequential ordering of event types (*order*). The feature selection study presented in Publication VI is also linked to the influence analysis experiments we have conducted with methods presented in Sections 3.1 and 3.2. We used the structural feature types *activity profile*, *transition profile*, *ordering*, *starter*, and *finisher* as new categorization dimensions, and found out that they often provide meaningful root causes for identified real-life process mining problems. The influence analysis results motivated us to use the machine learning
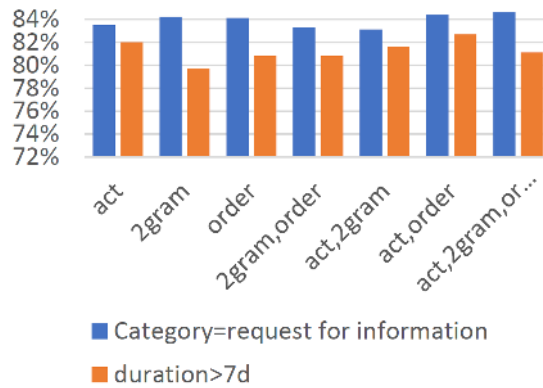
Figure 3.3. Clustering accuracy based on selected features

framework for a more detailed study about the significance of structural features.

A comparative analysis of process instance cluster techniques is presented in [59] and shows how various clustering techniques have been used to separate different process variants from a large set of cases as well as reducing the complexity by grouping similar cases into same clusters. With our method, the main functional requirement for the clustering algorithm is that it needs to put cases with similar process flow behavior into the same clusters, and all 20 approaches listed in [59] meet this requirement.

Encoding of event data
The amount of event data available from our analyses with real-life data urged us to find more efficient ways of selecting the features for clustering. Publication VII and Publication VIII present our findings for using Recurrent Neural Networks (RNN) for learning the relevant features automatically from the large datasets. Publication VII presents a method for encoding event activity sequences into input vectors. Publication VIII improves the analysis even further by presenting a methodology for using event attribute values for classification. This is particularly useful for business-oriented analysis where attributes like *customer*, *employee*, *product*, or *weekday* may have a large impact on business outcomes.

Proposed feature selection
Using our findings in Publication VI and Publication VII and based on the author's empirical evaluation, we propose the following structural features to be used for clustering:

- Activity profile. This profile contains one feature for each event type label in the data. The value of this feature is related to the number

of occurrences of that particular event type within the case. If the number of occurrences is used as an exact value, then the clustering algorithm somehow needs to take into account the continuous values—that is, repeating any activity seven times is more similar to repeating it six or eight times than repeating it only twice. One approach is to use value *zero* if the event log contains no occurrences of the event type for the given case and *one* if the log contains one or more occurrences. Based on the author's empirical evaluation, this approach provides good clustering results with a majority of the event logs. However, if it is essential to distinguish cases where a particular activity has occurred several times, it is recommended to use value *zero* for no occurrences of the event type, *one* for only one occurrence and two for *two* or more occurrences.

- Transition profile. This profile captures all process flows from each activity to the next activity. In effect, it contains the process control flow information. Transition profile potentially provides a large number of features up to the square of the number of event types plus one for start and end transitions. For example, in the sample analysis presented in Section 4.3.1, there are 42 distinct event types, giving potentially $43^2 = 1849$ distinct transitions. Fortunately, the control flow for 251 734 cases only contains 676 distinct transitions. Because the potential number of transition features is high, it is practical to use the coding *zero* if the transition does not occur in the case and *one* if it occurs once or more frequently.

Proposed clustering algorithm

If a particular clustering algorithm produces meaningful results and if there indeed is a correlation with a particular business area, then our method gives very high contribution values for that business area. The essential non-functional requirement for the clustering algorithm is performance—that is, the ability to produce results fast with a small amount of memory. Using our findings in Publication VI and Publication VII and based on the author's empirical evaluation, we propose the following algorithms and parameters for clustering:

- *One-hot encoding*. Since our activity and transition feature profiles only include categorical values *zero*, *one*, and *two*, it is possible to use an efficient one-hot encoding. The total number of feature vectors is the number of transitions plus twice the sum of activities.

- *Hamming distance* is the best choice for the distance function with binary data such as one-hot encoded features, because it completely avoids the floating-point distance calculations needed for the common Euclidian distance measure.

- *The k-modes* clustering algorithm is suitable for categorical data. In our tests, k-modes produced well-balanced clusters and was fast to execute. The result of k-modes depends on the initial cluster center initialization. Agglomerative clustering algorithms took more time to execute and produced highly unbalanced clusters. As the Influence Analysis method can be used with any clustering algorithm, the choice of algorithm, its parameter values and the actual input data may have a major effect on the discovered business areas.

- *Number of clusters* has a significant effect on clustering results. When the number of clusters is less than five, the large business areas correlate more with the clustering. While clustering to 10 or more clusters, the smaller business areas like *Vendor*, *Customer*, *Product* having more distinct values correlate more with the clusters. Running the clustering several times is also an easy way to mitigate the random behavior of k-modes coming from initialization.

- *Number of clustering runs* Clustering should be done several times with different number of clusters to discover significant business areas within a potentially large number of case attributes. We found out that clustering four times with cluster sizes 2, 3, 5, and 10 clusters gave enough variation in the results providing meaningful results about the business areas. For example, if the clustered cases originated from three distinct *ProductGroups* and five distinct *Locations*, then clustering runs of three and five are likely to discover differences significant differences in those dimensions if such differences exist.

3.4.2 Influence Analysis

Business Areas

All the case attributes that are relevant to business can be used as business area dimensions—for example, *product code*, *company code*, *product line*, *sales unit*, *delivery team*, *geographical location*, *customer group*, *product group*, *branch offices*, *request category*, and *diagnosis code*. However, a large organization may have thousands of low-level product codes in their ERP system, so it is beneficial to have access to product hierarchy and use each level as a separate business area dimension. Another example of a derived business area dimension is when a case attribute such as *logistics manager* can be used to identify the *delivery team*. We suggest having both the *logistics manager* and *delivery team* as business area dimensions; if any particular *logistics manager* has a major effect on process flow behavior and enough cases, then our method will show that person as the most significant business area in the *logistics manager* dimension. The third example of derived business areas is to utilize the event attributes. For

example, the *logistics manager* may be stored as an attribute value for the *delivery planning* activity. If there is always a maximum of one *delivery planning* activity, then the attribute value can be used as such in the case level. If there are multiple *delivery planning* activities, then typical options include using the first occurrence, using the last occurrence, or using a list of all distinct values from activities as the value on the case level. The outcome of forming business area dimensions is a list of case-level attributes that contain a specific (possibly empty) business area value for each case. To continue our formal methodology, we now consider these business area dimensions as case attributes and the case attribute values as the corresponding business areas.

Interestingness Measures

We now present the definitions for interestingness measures used for finding the business areas that correlate with the clustering results. Let $C = \{c_1,\ldots,c_N\}$ be the set of cases in the process analysis. Each case represents a single business process execution instance. Let $P = \{p_1,\ldots,p_N\}$ be a set of clusters each formed by clustering the cases in C. $C_p = \{c_{p_1},\ldots,c_{p_N}\}$ is the set of cases belonging to cluster p. $C_p \subseteq C$. Similarly, $C_a = \{c_{a_1},\ldots,c_{a_N}\}$ is the set of cases belonging to the same business area a, ie. they have the same value for the case attribute a.

Definition 38. *Let Density $\rho(a,C) = \frac{n(C_a)}{n(C)}$ where $n(C_a)$ is the total amount of cases belonging to the business area a and $n(C)$ is the total amount of all cases in the whole process analysis. Similarly, the density $\rho(a,C_p) = \frac{n(C_p \cap C_a)}{n(C_p)}$ is the density of cases belonging to the business area a within the cluster P.*

Definition 39. *Let $Contribution\%(a \to p) = \rho(a,C_p) - \rho(a,C) = \frac{n(C_p \cap C_a)}{n(C_p)} - \frac{n(C_a)}{n(C)}$ is the extra density of cases belonging to the business area a in the cluster p compared to average density.*

If business area a is equally distributed to all clusters, then the $Contribution\%(a \to p)$ is close to zero in each cluster. If the business area a is a typical property in a particular cluster p_i and rare property in other clusters, then the $Contribution\%(a \to p_i)$ is positive and other $Contribution\%(a \to p_j, where\, j <> i)$ values are negative. Calculating the sum of all Contribution values for all clusters is always zero, so the extra density in some clusters is always balanced by the smaller than average density in other clusters.

We now want to find the business areas that have a high contribution in many clustering. We define:

Definition 40. *Let $BusinessAreaContribution(a) = \sum_{p_i \in P} \frac{n(C_{p_i})}{n(C)}(max\{Contribution\%(a \to p_i), 0\})^2$.*
Here we sum the weighted squares of all positive contributions the business area a has with any clustering p_i. Positive values of $Contribution\%(a \to p_i)$

indicate a positive correlation with the business area a and the particular cluster i, while negative values indicate that the business area a has smaller than the average density in the cluster i. We found out that using only the positive correlations gives more meaningful results when consolidating to the business area level. Since a few high contributions are relatively more important than many small contributions, we use the Variance of the density differences, i.e., taking the square of the $Contribution\%(a \rightarrow p_i)$. Since a contribution within a small cluster is less important than contribution in a large cluster, we also use the cluster size based weight $\frac{n(C_p)}{n(C)}$.

Any particular business area a may have a substantial contribution in some clusters and small contribution in other, so the sum of all these clusterings is giving the overall correlation between business area a and all clusters $p_i \subseteq P$

We use the term *Business area* for any combination of a process mining case attribute and a distinct value for that particular case attribute. *BusinessAreaContribution* thus identifies the individual case attribute and value combinations that have the highest effect on clustering results. It is then also possible to continue and consolidate the results further to Case Attribute level:

Definition 41. *Let $AT = \{at_1,\ldots,at_N\}$ be a set of case attributes in the process analysis. Each case $c_i \in C$ has a value $at_{j_{c_i}}$ for each case attribute $at_j \in AT$. $at_{j_{c_i}}$ is the value of case attribute at_j for case c_i and $V_{at_j} = \{v_{at_{j_1}},\ldots,v_{at_{j_N}}\}$ is the set of distinct values that the case attribute at_j has in the process analysis.*

Definition 42. *Let $CaseAttributeContribution(at)$ be a sum of all BusinessAreaContributions from all the business areas corresponding to the given case attribute at as $\sum_{v_{at_{j_i}} \in V_{at_j}} BusinessAreaContribution(at_{j_{v_{at_{j_i}}}})$*

3.5 Distributed computing

The methods presented in Sections 3.1, 3.2, and 3.3 have originally been implemented in QPR ProcessAnalyzer product using SQL language queries. In order to facilitate the analysis of large process mining datasets, we researched the benefits of using distributed computing to meet the computational challenges. As business data is typically stored in SQL databases, we performed a comparative study for evaluating the feasibility of Big Data Processing frameworks. We present a method for assessing the capability and performance of distributed computing frameworks Hive, Presto, and Spark compared to traditional relational databases in Publication V using several test environments. The flow analysis and trace analy-
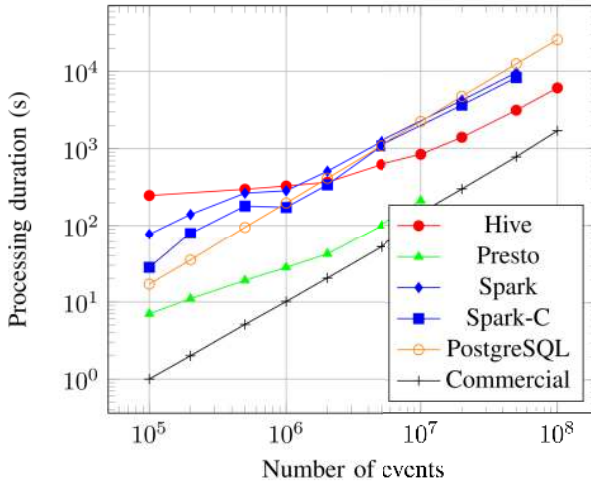
Figure 3.4. Distributed Computing Framework assessment results

sis test queries used in Publication V were heavily used in our original implementation of influence analysis.

The analysis results of performing typical process mining queries using several different distributed frameworks are shown in Figure 3.4. Tests have been performed in Amazon AWS EC2 based cluster using m1.large computing instances. It is interesting to notice that commercial SQL Server database outperformed distributed computing frameworks in all tests up to the processing of the maximum size of 100 million process mining events that completed in about 30 minutes.

Based on these results, we propose to use a Commercial SQL framework when executing the methods presented in this dissertation using SQL language. However, the performance of distributed computing frameworks is likely to improve, and they should be regarded as a viable alternative for a commercial implementation. Since the performance requirements are very important for customers using the methods, the latest versions of QPR ProcessAnalyzer use in-memory computing techniques utilizing several processor cores in parallel.

4. Results

This Chapter contains the results of applying our methods in publicly available real-life data and provides a discussion summary of using the methods in industrial projects. Further insight and case studies about using the merhod is found in Appendix A A. Supplementary details about the implmentation of these methods with the QPR ProcessAnalyzer tool is shown in Appendix B B.

4.1 RQ1: How can process mining be used for resource allocation to maximize business improvement?

4.1.1 Case Study: Rabobank Group ICT

We evaluated the influence analysis with publicly available data from Rabobank Group ICT used in BPI Challenge 2014 [21]. The data consists of 46.616 cases and a total of 466 737 events. Using a process mining analysis, we discovered that the average duration for cases is five days, and the median duration is 18 hours. We decided to consider all cases that took more than one week to complete as problematic, resulting in a total of 7.400 (16%) cases. Table 4.1 shows that the biggest contributor to this finding is *Impact=5*. There is a total of 16.741 cases with *Impact=5*, out of which 3.535 (21%) are problematic. As a *contribution%* this corresponds to 12% of the total amount of problematic cases. For the process performance point of view, this is intuitive since it is probably acceptable to have low (5=lowest on scale 1..5) impact cases taking a long time compared to higher impact cases. Table 4.1 shows that 29% of cases having *ServiceComp WBS (CBy) = WBS000091* are completed in more than one week, which makes *WBS000091* a candidate for business process improvements. If *WBS000091* would reach the average level of performance, then there would 4% less problematic cases.

Table 4.2 shows rules that have the biggest negative contribution. These

Table 4.1. Top positive contributors

Business Area	n(A)	n(AB)	Δn	ΔP(B)
Impact = 5	16741	3535	877	0,12
Urgency = 5	16779	3538	874	0,12
Priority = 5	16486	3473	856	0,12
# Related Interactions = 2	2736	1108	674	0,09
# Update from customer = 1	1692	793	524	0,07
Closure Code = Other	16470	3137	522	0,07
# Reassignments = 2	5378	1340	486	0,07
# Reassignments = 3	2191	814	466	0,06
# Reassignments = 4	1606	701	446	0,06
Category = request for information	8846	1810	406	0,05
CI Type (CBy) = computer	3404	865	325	0,04
ServiceComp WBS (CBy) = WBS000091	2453	700	311	0,04
CI Type (CBy) = application	29456	4979	303	0,04

can be regarded as the reasons why cases are completed within one week more often than average. If *#Reassignments = 0*, then only 6% of cases will take more than one week. If these cases would take the same time as average cases, then there would be 37% more problematic cases. Another observation from Table 4.2 is that only 11% of cases having *ServiceComp WBS (CBy) = WBS000073* are completed late, which makes *WBS000073* a positive benchmark.

Table 4.2. Contribution analysis on attribute value level – top negative contributors

Busines Area	n(A)	n(AB)	Δn	ΔP(B)
CI Name (aff) = SUB000456	3050	138	-346	-0,05
CI Type (aff) = subapplication	7782	841	-394	-0,05
Category = incident	37748	5582	-410	-0,06
Closure Code = User error	3554	152	-412	-0,06
CI Type (CBy) = subapplication	7711	800	-424	-0,06
Urgency = 3	6536	607	-431	-0,06
Priority = 3	6703	620	-444	-0,06
Impact = 3	6591	602	-444	-0,06
Service Component WBS (aff) = WBS000073	13342	1437	-681	-0,09
ServiceComp WBS (CBy) = WBS000073	13173	1401	-690	-0,09
Reopen Time = (blank)	44332	6285	-752	-0,1
# Related Interactions = 1	43058	5907	-928	-0,13
# Reassignments = 0	27468	1628	-2732	-0,37

ServiceComp WBS (CBy) was identified both as having a high positive and negative contribution. Based on the author's empirical evaluation, it is often beneficial to continue the analysis by showing the contribution of all distinct values of this case attribute in one list ordered by contribution, as shown in Table 4.3. If the amount of distinct values is large, we propose presenting only the top-10 most significant positive and negative contributions.

Table 4.3. Benchmark of distinct values of ServiceComp WBS (CBy)

ServiceComp WBS (CBy)	Contribution
WBS000091	4%
WBS000072	3%
WBS000088	2%
WBS000162	2%
WBS000263	1%
WBS000296	1%
WBS000271	1%
WBS000092	1%
WBS000187	1%
WBS000089	0%
WBS000318	0%
WBS000219	0%
WBS000172	0%
WBS000096	0%
WBS000223	0%
WBS000125	0%
WBS000292	0%
WBS000146	0%
WBS000128	-1%
WBS000094	-1%
WBS000307	-1%
WBS000152	-1%
WBS000016	-1%
WBS000228	-1%
WBS000095	-2%
#N/B	-2%
WBS000073	-9%

In this Section, we used contribution analysis with real case data. We were able to identify causes for cases lasting more than one week. We observed a benchmark for a particular case attribute that seems to contribute a lot to the finding. All the results have been shown in easy-to-understand lists ordered by the contribution metric. Showing these results to the people working in the process gives them more fact-based insight and enables them to combined this information with their tacit knowledge to discover even deeper underlying cause-effect relationships.

Analyzing process lead times and using weights
We now continue our case study using the Rabobank Group ICT data [21]. As the lead time measure for each case, we used the total case duration

from the first event to the last event within the case. Typical process mining analysis discovers that the average duration for cases is 5.07 days, and the median duration is 18 hours. If the duration was longer than seven days, the case is categorized as problematic, which results in a total of 7 400 (15.9%) problematic cases for the *Binary Analysis*.

As the weighting for cases, we use a formula $w_{c_j} = (6 - Impact_{c_j})(6 - Urgency_{c_j})(6 - Priority_{c_j})$ where *Impact*, *Urgency* and *Priority* all have values in (1,2,3,4,5) where 1 means highest importance and 6 is the lowest importance. With this formula, the highest possible weight $\max w = (6 - 1)(6-1)(6-1) = 125$, lowest possible weight $\min w = (6-5)(6-5)(6-5) = 1$, and the weighted average lead time drops to 3.97 days. Since the weighted average lead time is shorter than the equal weight lead time, we conclude that the average lead time is shorter for more important cases than for less important cases. The same finding can also be made from the binary results since the average problem density is 15.9% compared to the weighted average problem density of 11.1%.

Contribution results for individual case attribute values
The most significant positive and negative root causes according to different Analysis Types are shown in Table 4.4. The analysis is limited to the distinct values of case attribute *ServiceComp WBS(CBy)*, and the first column shows the case attribute value. Two columns are shown for each analysis type. Column *Cont* is the contribution% measure calculated using our methods, which shows how a large portion of the problematic cases can be explained with this case attribute value. Positive numbers mean there are more than average problems in these cases, and negative numbers represent best practice areas with a smaller amount of problems. The second column for each analysis type is the rank of the case attribute value within the full results for that analysis type. The rank **+1** means the most significant root cause, and the rank **-1** means the most significant best practice area.

For BiCo the highest contribution% is 4.2% for case attribute value *WBS000091* and lowest contribution% is -9.3% for case attribute value *WBS000073*. When considering the most beneficial focus area for process improvement reducing the lead time most we see that BiCo results in *WBS000091* and all other Contribution Formulas result in *WBS000088*. According to the Table 4.4 the best performing *ServiceComp WBS(CBy)* regarding the lead time is *WBS000073* in all analysis types except that in Weighted Binary Contribution the best performing area is *#N/B*.

Some interesting results include the behavior of cases whose attribute *ServiceComp WBS(CBy)* has the value *WBS000091* which contributes to 4.2% (Top 1) of the total problem in BiCo, 3.4% (Top 2) in wBiCo, but only 1.7% (Top 8) in CoCo and only 1.2% (Top 7) for wCoCo. The reason for the higher contribution in BiCo and lower in CoCo is that average lead

Table 4.4. Comparison of root causes for all analysis types

ServiceComp WBS(CBy)	BiCo Cont	R	wBiCo Cont	R	CoCo Cont	R	wCoCo Cont	R
WBS000091	+4,2%	+1	+3,4%	+2	+1,7%	+8	+1,2%	+7
WBS000072	+2,8%	+2	+0,8%	+10	+3,4%	+4	+0,6%	+12
WBS000088	+2,4%	+3	+3,7%	+1	+10,4%	+1	+12,7%	+1
WBS000162	+2,2%	+4	+2,9%	+3	+8,6%	+2	+10,0%	+2
WBS000055	+0,7%	+9	+1,2%	+7	+3,5%	+3	+4,2%	+4
WBS000043	+0,2%	+21	+1,2%	+8	+1,3%	+10	+4,8%	+3
.....								
WBS000228	-1,0%	-4	+0,1%	+50	-1,7%	-3	-0,2%	-39
WBS000146	-0,5%	-9	-2,2%	-3	-0,7%	-11	-2,8%	-4
WBS000095	-1,7%	-3	-0,6%	-8	-2,5%	-2	-0,9%	-8
#N/B	-1,7%	-2	-8,3%	-1	+2,6%	+5	-5,9%	-2
WBS000073	-9,3%	-1	-4,1%	-2	-17,4%	-1	-10,9%	-1

time for area *WBS000091* is only 6.14, which is only a little longer than the average for the whole process 5.07. This means that there are many *WBS000091* -cases that have lead time a little bit longer than 7 days. On the other hand, the behavior of area *WBS000088* is the opposite, since it only contributes 2.4% (Top 3 value) of the total problem in BiCo, 3.7% (Top 1) in wBiCo, much more 10.4% (Top 1) in CoCo and even more 12.7% (Top 7) in wCoCo. Reason for this behavior is that the average lead time for cases in area *WBS000088* is 39.2 days, which is much longer than the average lead time 5.07

Exciting results include the behavior of area *#N/B*, which is listed as a best practice area with negative contribution -1.7% in BiCo (Top -2), -8.3% in wBiCo (Top -1) and -5.9% in wCoCo (Top -2). However, it is listed as a problem area with a positive contribution 2.6% in CoCo as the 5th most important problem area. There are at least two reasons for this result: first, the high weight cases in *#N/B* perform much better than the low weight cases, i.e., BiCo contribution gets 6.6 percentage points better with weighting than without and CoCo contribution gets 8.5 percentage points better. The second reason is that area *#N/B* performs consistently worse in *Continuous Analysis* compared to the *Binary Analysis*, which is caused by the higher than average lead time of 6.2% in CoCo, which again is caused by some specific long-duration cases in *#N/B* which may also be regarded as outliers in the data.

Contribution results for activity occurrences
In this Section, we present the Rabobank root cause analysis for long-lasting cases using activity occurrence data. As a preprocessing step, we add a new case attribute for each different activity name and use the

number of activity occurrences as the value for that case attribute in each case. For example, if activity *Status Change* occurs twice for a certain case, then the value of case attribute Status Change is *2* for that particular case.

Table 4.5 shows the top-5 positive and negative root causes for binary analysis types and Table 4.6 the same for the continuous analysis types leading us to the following observations:

- Lack of reassignments is the most important negative root cause for a case to exceed the seven day SLA (BiCo analyses) or generally take a long time (CoCo analysis). In other words, having zero reassignments makes a case very fast.
- Contribution values for activity occurrence numbers are much higher than they are for the case attribute *ServiceComp WBS(CBy)*, which means that these activity amounts correlate more with the total duration than the case attribute *ServiceComp WBS(CBy)*.
- *Update from customer(1)* is the most important positive root cause for long case duration, as can be seen in continuous contributions in Table 4.6. However, for binary contributions in Table 4.5, the *Status Change(2)* is the most important positive root cause, which means that having two occurrences of *Status Change* causes a bigger risk than getting an *update from a customer* for failing SLA.

Table 4.5. Comparison of root causes based on activity profiles for *binary analysis*

BiCo		wBiCo	
Activity occurrences	Contrib.	Activity occurrences	Contrib.
n(Closed)=2	9.7%	n(Status Changes)=2	10.3%
n(Status Changes)=2	8.9%	n(Comm. with customer)=1	10.1%
n(Comm. with customer)=1	8.8%	n(Closed)=2	9.9%
n(Reopen)=1	7.8%	n(Update from customer)=1	9.5%
n(Update from customer)=1	7.1%	n(Assignment)=3	9.3%
...	.	...	.
n(Status Change)=0	-20.8%	n(Status Change)=0	-22.0%
n(Assignment)=1	-26.4%	n(Assignment)=1	-26.8%
n(Update)=0	-29.6%	n(Update)=0	-32.0%
n(Operator Update)=0	-35.8%	n(Operator Update)=0	-41.5%
n(Reassignment)=0	-37.9%	n(Reassignment)=0	-43.2%

Contribution results for case attributes

In this Section, we will show the results of calculating the summary contribution values for case attributes. We use both the original case attributes provided in the source data as well as the activity occurrences features as case attributes.

Table 4.7 shows that the count of Reassignments is the most influential case attribute for explaining the root causes for long durations for all

Table 4.6. Comparison of root causes based on activity profiles for *continuous analysis*

CoCo		wCoCo	
Activity occurrences	Contrib.	Activity occurrences	Contrib.
n(Update from customer)=1	14.9%	*n(Update from customer)=1*	15.6%
n(Closed)=2	13.1%	*n(Status Change)=2*	12.8%
n(Status Change)=2	12.2%	*n(Update from customer)=2*	12.5%
n(Reopen)=1	11.8%	*n(Update)=2*	12.1%
n(Description Update)=1	11.1%	*n(Description Update)=1*	12.0%
...	...		
n(Update from customer)=0	-38.6%	*n(Assignment)=1*	-43.0%
n(Assignment)=1	-45.8%	*n(Update from customer)=0*	-46.3%
n(Update)=0	-51.3%	*n(Update)=0*	-50.1%
n(Operator Update)=0	-53.2%	*n(Operator Update)=0*	-55.8%
n(Reassignment)=0	-62.0%	*n(Reassignment)=0*	-65.0%

Table 4.7. Case Attribute analysis results for all analysis types

Case Attribute	BiCo		wBiCo		CoCo		wCoCo	
	Cont	*R*	*Cont*	*R*	*Cont*	*R*	*Cont*	*R*
n(Reassignment)	15.9%	1	20.8%	1	42.0%	1	46.5%	1
n(Operator Update)	14.5%	2	19.4%	2	31.0%	2	34.1%	2
n(Assignment)	10.2%	3	12.5%	3	27.5%	4	27.9%	4
n(Update)	9.8%	4	11.8%	4	29.3%	3	28.9%	3
n(Status Change)	6.8%	5	8.1%	5	12.5%	6	11.8%	6
n(Update from customer)	2.4%	9	5.0%	6	18.4%	5	26.2%	5
n(Comm. with customer)	3.5%	6	5.0%	7	5.9%	8	5.7%	13
Closure Code	1.3%	18	2.3%	12	4.9%	12	7.7%	9
Service Component WBS (aff)	1.5%	16	1.4%	20	7.4%	7	8.8%	8
# Related Interactions	2.5%	8	2.3%	13	5.3%	10	2.9%	21
Impact	2.0%	10	3.3%	8	1.2%	25	5.9%	12
n(Closed)	2.6%	7	2.1%	14	5.3%	11	2.7%	23
ServiceComp WBS (CBy)	1.4%	17	1.5%	18	5.6%	9	5.0%	17
Priority	1.9%	13	3.2%	9	1.1%	26	5.6%	14
Cost	1.9%	12	3.0%	10	1.1%	27	5.5%	15
Urgency	2.0%	11	3.0%	11	1.1%	28	5.4%	16
n(Reopen)	1.6%	15	1.7%	16	4.1%	14	2.7%	22
CI Name (aff)	0.7%	22	0.9%	25	4.8%	13	6.9%	11
n(Description Update)	1.0%	19	1.9%	15	3.2%	20	3.9%	20
Related Interaction	1.6%	14	1.5%	17	3.6%	16	2.1%	27
KM number	0.4%	28	1.2%	21	3.8%	15	7.3%	10

analysis types. We see that the number of occurrences of *Operator Update*, *Assignment*, *Update*, *Status Change*, *Update from customer*, and *Communication with customer* all appear as very influential case attributes. This means that the activity occurrence amounts have a high correlation with long lead times. The first original case attributes in the list are *Service Component WBS (aff)* and *Closure Code*. They both are more relevant in the Continuous Contribution analyses compared to binary analyses. Some interesting findings from these results include:

- Several activity occurrence counts have a much stronger influence on long lead times than the original case attributes. This means that what happens during the process instance execution is often more relevant than original case attributes.

- *Service Component WBS (aff)* is the most important original case attribute that could be used to predict long lead times. However, the contribution for *Service Component WBS (aff)* is clearly higher for *Continuous Contribution* analysis types (Rank for *CoCo* is 7 and Rank for *wCoCo* is 8) compared to the *Binary Contribution* analysis (Rank for *BiCo* is 16 and rank for *wBiCo* is 20). This means that as a whole, the individual *Service Component WBS (aff)* values are not such important root causes for determining if a case takes more than one week. However, some individual *Service Component WBS (aff)* values are significant root causes for those cases that have a very long duration, since *CoCo* analysis gives high weight for very long cases.

- *Impact*, *Priority*, and *Urgency* are all more relevant root causes for weighted analysis than for non-weighted analysis. The reason for this is because the case-specific weights were calculated based on these three case attributes.

4.1.2 Discussion

This section is based on the author's empirical evaluation of using our method in industrial process mining projects. Influence Analysis was first introduced in the commercial product QPR ProcessAnalyzer [1] version 3.7 on 27th April 2012, showing both the change type *ideal* and change type *as-is average* results [2].

Influence analysis has been successfully used in more than 200 customer process mining projects. Vanjoki presents a summary of analyzing automated purchase to pay process value modeling and comparative process

---

[1] https://www.qpr.com/products/qpr-processanalyzer
[2] https://devnet.onqpr.com/pawiki/index.php/QPR_ProcessAnalyzer_Release_Archive

speeds using influence analysis [60]. In practice, problem areas and best practice areas have been accurately identified using influence analysis.

Interactive usage in workshop meetings has proven to be very valuable, and it motivates business people in the same meeting to share their tacit knowledge to deepen the influence analysis findings. A typical scenario is that participants first try to guess the most influencing factors, and when they then see the results their own hypotheses are strengthened or weakened. This process further facilitates participants' thinking and collaboration with each other. Based on the discussion, the organization then selects the focus areas for business process improvements and starts monitoring the performance on monthly intervals using the same contribution measures.

Our method is able to identify root causes for problems ranging from very rare to very common. However, a certain number of successful cases are necessary in order to discover root causes with our method. In practice, the Pareto principle percentages of 20% problematic cases versus 80% successful cases seem to be the ideal ratio, and anything between 0.1% and 50% problematic cases still works well. If noticeably more than 90% of cases are problematic, then the discovered best practice areas are often more insightful than the discovered root cause areas.

Influence analysis also has an important application in deciding whether the organization should improve the whole process design or improve certain problem areas. If the contribution values for all rules are relatively low, then there is no clear root cause that should be fixed. Thus, if no focus area is found and business still needs to be improved, there is a need to improve the whole process design. Also, the method can be used to evaluate potential risks in any given segment by checking those areas that have a low density of problematic cases in the current as-is situation, since the number of problems will increase if those areas become like others, on average.

Actual root cause analysis is an iterative method for identifying the underlying root causes of problems [6]. Depending on the problem and the case attribute data available, the influence analysis method may return a significant root cause already in the first query. However, the first query may return a higher-level finding like *Region=Europe*. To further discover the root cause for the original problem in *Europe*, we suggest creating a process mining filter so that our method can be run for all *Europe* cases. This second influence analysis could now return a significant factor *ProductLine=Hats*, which now suggests that the *ProductLine=Hats* in *Europe* is a particularly significant root cause for the original problem. Further multi-level queries of this kind can pinpoint very specific root causes for the problems, since while every round reduces the number of cases, it is possible to increase the number of case attributes while providing fast response times for the business analyst.

Summary of our results:

1. The influence analysis methodology is able to find root causes for long lead times.

2. Root causes for long lead times may be substantially different when using a predefined lead time limit for problematic/successful cases (binary) compared to when using continuous lead time values.

3. Case-specific weighting can be easily used when analyzing both binary and continuous contributions.

4. When weighting is used in the continuous contribution analysis, the results can be directly used for discovering opportunities for reducing working capital.

5. Contribution values can be summarized to the case attribute level by calculating the sum of squares of individual *contribution%* values. This gives a quick overview of interesting case attributes for any given lead time finding.

4.2 RQ2: How can process mining be used to identify changes in business operations?

4.2.1 Case Study: BPI Challenge 2017 Dataset

In this Section, we show a real-life example of using the presented methodology on the loan applications process data from a Dutch Financial Institute. The data is publicly available as BPI Challenge 2017 Dataset [22] and contains 31 509 cases and a total of 1 202 267 events. The original dataset has been prepared in a way that it contained full cases. Since the purpose of our analysis is to show business process changes within a continuous monitoring situation, we have taken the following steps in preparing a setup for business review analysis.

- November 2016 is selected as the business review month. The data contains 104 946 events whose timestamp belongs to November so that the Review period will consist of these events.

- All events occurring later than November belong to the *Most Recent Data* period and are excluded from analysis, consisting of 120 568 events. These events would naturally be included in later business review periods.

- The comparison period has been chosen to include the six months before the review period, i.e., from May 2016 to October 2016 containing 647 406 events.

- History period contains 329 347 events occurring before May 2016. These events are used for constructing the process path and predecessor dimensions for History and Review events, but they are not included in the analysis as actual events belonging to either Comparison or Review sets.

- The total number of events in the analysis is 752 352 consisting of 104 946 events for the Review period (13.95% of all events) and 647 406 events for the Comparison period (86.05% of all events).

| Total | | 752352 | 104946 | 647406 | 13.95% | |
Dimension	Value	Events	Review	Comparison	Density %	Contribution %
org:resource	User_133	7995	3728	4267	46.63%	2.49%
org:resource	User_65	3015	2033	982	67.43%	1.54%
org:resource	User_67	5812	2326	3486	40.02%	1.44%
org:resource	User_131	4675	2149	2526	45.97%	1.43%
org:resource	User_100	12069	2787	9282	23.09%	1.05%
org:resource	User_66	2863	1481	1382	51.73%	1.03%
lifecycle:transition	ate_abort	54945	8686	46259	15.81%	0.97%
org:resource	User_78	4000	1556	2444	38.90%	0.95%
org:resource	User_3	15379	3133	12246	20.37%	0.94%
Action	Deleted	92748	13854	78894	14.94%	0.87%
org:resource	User_134	986	986	0	100.00%	0.81%
org:resource	User_69	941	941	0	100.00%	0.77%
lifecycle:transition	suspend	134262	17984	116278	13.39%	0.71%
Event Type	W_Call after offers - suspend	40235	4865	35370	12.09%	-0.71%
Predecessor	[W_Validate application - resume] - [W_Validate application - suspend]	18465	1780	16685	9.64%	-0.76%
Event Type	W_Validate application - resume	18603	1792	16811	9.63%	-0.77%
org:resource	User_45	6104	0	6104	0.00%	-0.81%
Action	Obtained	158671	21243	137428	13.39%	-0.85%
org:resource	User_112	7610	76	7534	1.00%	-0.94%
lifecycle:transition	resume	77702	9700	68002	12.48%	-1.09%
org:resource	User_60	9144	0	9144	0.00%	-1.22%
org:resource	User_116	9506	0	9506	0.00%	-1.26%

Figure 4.1. Changes for BPI Challenge 2017 Applications. Changes for November 2016 compared to the previous six months

Results using Binary Change Window

Figure 4.1 shows the top-10 most important changes in the business process and related data for the review period. We see that there are many user changes in event attribute *org:resource*, so it seems like employees are changing a lot. *User_133* has conducted 3 728 events during the review period and only 4 267 in the comparison period, so 47% of his events have taken place during the review period, which makes him as the biggest process change, taking into account the size of his total activity (7 995 events) and the difference 33% from the average 13.95%.

Figure 4.2 shows the changes in only the event type dimension. The event types *W_Call incomplete files - suspend* and *W_Call after offers - ate_abort* occur more often during the Review period, whereas the event types *W_Validate application - resume* and *W_Call after offers - suspend* occur less often during the Review period than in Comparison period.

Considering the business process related changes where the order of activities is changing, we limit the analysis to only the predecessor changes where a specific event takes place immediately after another event, as

Total		752352	104946	647406	13.95%	
Dimension	Value	Events	Review	Comparison	Density %	Contribution %
Event Type	W_Call incomplete files - suspend	38342	5926	32416	15.46%	0.55%
Event Type	W_Call after offers - ate_abort	19842	3247	16595	16.36%	0.46%
Event Type	W_Call after offers - schedule	20422	3293	17129	16.12%	0.42%
Event Type	W_Call incomplete files - resume	24829	3761	21068	15.15%	0.28%
Event Type	W_Validate application - ate_abort	15269	2412	12857	15.80%	0.27%
Event Type	W_Call incomplete files - ate_abort	13102	2078	11024	15.86%	0.24%
Event Type	A_Incomplete - complete	14892	2325	12567	15.61%	0.24%
Event Type	W_Call incomplete files - schedule	14892	2325	12567	15.61%	0.24%
Event Type	W_Call incomplete files - start	14982	2333	12649	15.57%	0.23%
Event Type	W_Validate application - schedule	24774	3621	21153	14.62%	0.16%
Event Type	O_Created - complete	27165	3700	23465	13.62%	-0.09%
Event Type	A_Accepted - complete	19880	2683	17197	13.50%	-0.09%
Event Type	O_Sent (mail and online) - complete	25115	3407	21708	13.57%	-0.09%
Event Type	A_Concept - complete	19906	2678	17228	13.45%	-0.09%
Event Type	W_Complete application - schedule	19917	2679	17238	13.45%	-0.09%
Event Type	A_Create Application - complete	19906	2676	17230	13.44%	-0.10%
Event Type	W_Validate application - suspend	34439	4198	30241	12.19%	-0.58%
Event Type	W_Call after offers - resume	20035	2100	17935	10.48%	-0.66%
Event Type	W_Call after offers - suspend	40235	4865	35370	12.09%	-0.71%
Event Type	W_Validate application - resume	18603	1792	16811	9.63%	-0.77%

Figure 4.2. Changes in Event Types for BPI Challenge 2017 Applications. Changes for November 2016 compared to the previous six months

Total		752352	104946	647406	13.95%	
Dimension	Value	Events	Review	Comparison	Density %	Contribution %
Predecessor	[W_Call after offers - schedule] - [W_Call after offers - ate_abort]	625	617	8	98.72%	0.50%
Predecessor	[W_Call after offers - ate_abort] - [W_Call after offers - suspend]	11302	1923	9379	17.01%	0.33%
Predecessor	[W_Call incomplete files - suspend] - [W_Call incomplete files - resume]	22739	3496	19243	15.37%	0.31%
Predecessor	[W_Call incomplete files - resume] - [W_Call incomplete files - suspend]	24457	3713	20744	15.18%	0.29%
Predecessor	[W_Call incomplete files - schedule] - [A_Incomplete - complete]	14811	2321	12490	15.67%	0.24%
Predecessor	[W_Call incomplete files - start] - [W_Call incomplete files - schedule]	14850	2323	12527	15.64%	0.24%
Predecessor	[A_Incomplete - complete] - [W_Validate application - suspend]	7815	1324	6491	16.94%	0.22%
Predecessor	[W_Validate application - ate_abort] - [W_Call incomplete files - start]	8009	1344	6665	16.78%	0.22%
Predecessor	[W_Call incomplete files - suspend] - [W_Validate application - ate_abort]	7886	1324	6562	16.79%	0.21%
Predecessor	[W_Validate application - start] - [W_Validate application - schedule]	24713	3617	21096	14.64%	0.16%
Predecessor	[O_Create Offer - complete] - [A_Accepted - complete]	19013	2572	16441	13.53%	-0.08%
Predecessor	[A_Create Application - complete] - START	12983	1728	11255	13.31%	0.08%
Predecessor	[W_Call after offers - start] - [W_Call after offers - schedule]	19729	2668	17061	13.52%	-0.08%
Predecessor	[O_Created - complete] - [O_Create Offer - complete]	27165	3700	23465	13.62%	-0.09%
Predecessor	[O_Returned - complete] - [A_Validating - complete]	10736	1392	9344	12.97%	-0.10%
Predecessor	[A_Incomplete - complete] - [W_Validate application - resume]	3565	330	3235	9.26%	-0.16%
Predecessor	[W_Validate application - suspend] - [W_Validate application - resume]	12174	1167	11007	9.59%	-0.51%
Predecessor	[W_Call after offers - resume] - [W_Call after offers - suspend]	19010	1997	17013	10.50%	0.62%
Predecessor	[W_Call after offers - suspend] - [W_Call after offers - resume]	19914	2094	17820	10.52%	-0.65%
Predecessor	[W_Validate application - resume] - [W_Validate application - suspend]	18465	1780	16685	9.64%	-0.76%

Figure 4.3. Changes in Predecessors for BPI Challenge 2017 Applications. Changes for November 2016 compared to the previous six months

shown in Figure 4.3. During the review period, the control flow transition from event *W_Call after offers - ate_abort* to *W_Call after offers - schedule* occurs more often and the transition from event type *W_Validate application - suspend* to *W_Validate application - resume* less often than during the comparison period.

Results using Continuous Change Window

Figure 4.4 shows the continuous approach versions of the same overall analysis as the previous binary approach Figure 4.1. The continuous analysis is configured to discover differences in events from mid-August to November with events from May to mid-August. The results of Continuous analysis are the result of giving each event a weight based on the *Age* of the event. The bigger the distance from average *Age*, the bigger the weight of that particular event. Using the case study data, the average *Age* of events is 103.97 days. An event occurring 100 days before or after the average data have 100 times the weight compared to an event taking place one day after or before the average date. Similarly, an event that

Dimension	Value	Events	Avg Age	Contribution %
Total		752352	103.97	
org:resource	User_133	7995	33.88	-2.89 %
org:resource	User_67	5812	36.44	-2.02 %
lifecycle:transition	ate_abort	54945	97.63	-1.80 %
org:resource	User_131	4675	34.41	-1.68 %
Action	statechange	226453	102.67	-1.51 %
org:resource	User_123	18244	87.92	-1.51 %
Action	Deleted	92748	100.90	-1.47 %
org:resource	User_65	3015	24.07	-1.24 %
EventOrigin	Application	151446	102.41	-1.22 %
org:resource	User_56	5578	64.83	-1.13 %
Action	Released	134262	106.38	1.67 %
lifecycle:transition	suspend	134262	106.38	1.67 %
org:resource	User_87	12652	130.91	1.76 %
org:resource	User_117	4251	184.97	1.78 %
Event Type	W_Validate application - suspend	34439	114.58	1.89 %
Action	Obtained	158671	106.69	2.22 %
org:resource	User_116	9506	151.05	2.31 %
Predecessor	[W_Validate application - resume] - [W_Validate application - suspend]	18465	130.28	2.51 %
Event Type	W_Validate application - resume	18603	130.25	2.52 %
lifecycle:transition	resume	77702	111.02	2.83 %

Figure 4.4. Changes for BPI Challenge 2017 Applications. Changes for November 2016 compared to the previous six months using continuous comparison approach

takes place exactly in the average *Age* has zero weight as it does not belong either to the old period or new period.

As shown in Figure 4.4, the most significant positive change over time has been the number of events with *User_133* as the org:resource. The average Age of these events is 70.09 days shorter than the average (103.97 days - 33.88 days). The Age is defined as relative to a fixed timestamp in the analysis, and the absolute zero point for the Age is not relevant. The total problem size for the analysis is the total sum of positive differences from the average Age calculated from each event, which in this case study model gives a total of 19382661.6 days. The contribution of *User_133* is thus calculated as $-70.09 days * 7995 events/19382661 days = -2.89\%$

Continuous analysis results are well in line with the binary approach results, and differences are based on the different setup of Review and Comparison periods and a different weighting approach as described. For example, *User_133* as the new value for *org:resource* is still the biggest change, and both *org:resources User_67* and *User_65* are included in top-10 changes for both Binary and Continuous approaches as is visible in Figures 4.1 and 4.4.

In this dissertation we have presented a method for detecting business process changes. The method is based on previously published Influence Analysis and it uses the conformance measure to scale different types of changes in order to present various kind of changes sorted by their significance. We have shown how to use Influence Analysis on the event level instead of business process case level. Operating on the event level makes it possible to use all available data from the review period for detecting changes instead of having to wait until a business process case is completed. Summary of our key experiences when using the analysis with real-life cases include:

4.2.2   Discussion

The following observations can be made based on the previous case study results and the author's empirical evaluation of using the method in industrial process mining projects:

- Changes in business operations can be analyzed by comparing review period events to comparison period events using influence analysis.

- Operating on the event level makes it possible to use all available data from the review period for detecting changes instead of having to wait until a business process case is completed.

- Business people quickly learn to read the influence analysis results on a monthly basis. Detecting the top 10 or top 50 changes gives an excellent starting point for a more detailed periodical analysis of business process changes.

- Detected changes may also be a result of incorrect data integration between the process mining system and the actual ERP system(s). Our method serves as an easy to use quality assurance tool for evaluating the correctness of periodical data loads and integration. For example, after each monthly, weekly, or daily data import, the system can notify business analyst of the top 10 changes so that a potential technical integration problem is detected and corrected before other business users waste time analyzing incorrect data.

4.3   RQ3: How can business areas that have a significant effect on process flow behavior be discovered using clustering?

4.3.1   Case Study: Purchase Order process

In this Section, we apply our method to the real-life purchase order process data from a large Netherlands multinational company operating in the area of coatings and paints. The data is publicly available as the BPI Challenge 2019 [23] dataset. We made the following choices:

- **Source data** We imported the data from the XES file as such without any modifications. To keep the execution times short, we experimented with the effect of running the analysis with a sample of the full dataset. Our experiments showed that the results remained consistent for sample size 10 000 cases and more. With a sample size of 1 000 cases, the *K-modes* clustering results were changing a lot of due to a large number of features and a small number of cases.

- **Clustering algorithm** We used the k-modes clustering as implemented in Accord.Net Machine Learning Framework [55] with one-hot encoding and hamming distance function. To take into account the different clustering sizes, we performed clustering four times, fixed to two, three, five, and ten clusters.

- **Activity profile features for clustering** We used our default boolean activity profile, which creates one feature dimension for each activity, and the value is *zero* if the activity does not occur in a case, value *one* if the activity occurs once, and value *two* if it is repeated multiple times. There were 37 different activities in the sample, and the Top 20 activity profile is shown in Table 4.8.

Table 4.8. Activity profile: Top 20 activities ordered by unique occurrence count

Name	Unique Count	Count
Create Purchase Order Item	10 000	10 000
Record Goods Receipt	9 333	13 264
Record Invoice Receipt	8 370	9 214
Vendor creates invoice	8 310	8 901
Clear invoice	7 245	7 704
Remove Payment Block	2 223	2 272
Create Purchase Requisition Item	1 901	1 901
Receive Order Confirmation	1 321	1 321
Change Quantity	707	853
Change Price	443	498
Delete Purchase Order Item	338	339
Cancel Invoice Receipt	251	271
Vendor creates debit memo	244	253
Record Service Entry Sheet	232	10 326
Change Approval for Purchase Order	194	319
Change Delivery Indicator	112	128
Cancel Goods Receipt	109	136
SRM: In Transfer to Execution Syst.	42	57
SRM: Awaiting Approval	42	50
SRM: Complete	42	50

- **Transition profile features for clustering** Using a typical process mining analysis to discover the process flow diagram, we discovered 376 different direct transitions, including 13 starting activities, 22 ending activities, and 341 direct transitions between two unique activities. All of these 376 features were used as dimensions for clustering in a similar way as the activity profile, i.e., the value *zero* if transition did not occur in a case, and *one* if it occurred once or multiple times.

- **Business area dimensions** Since we did not have any additional information or hierarchy tables concerning possible business areas, we are using all available 15 distinct case attributes listed in Table 4.11 as business area dimensions. These case attributes have a total

of 9901 distinct values, giving us 9901 business areas to consider when finding those business areas that have the most significant effect on process flow.

Clustering Results for Individual Clustering

Table 4.9 shows the results of clustering to fixed five clusters. We see that the first cluster contains 48% of cases, the second cluster 33%, third 17%, and both fourth and fifth one percent each. Here we show the five most important business areas based on the contribution%, which is calculated as the difference between Cluster specific density of that business area and Total Density. These results already give hints about the meaningful characteristics in the whole dataset, i.e., the first cluster contains many *Standard* cases from spend areas related to *Sales*, *Products for Resale*, and *NPR*. On the other hand, the second cluster contains more than the average amount of cases from *spend area Packaging*, related to *Labels* and *PR*. *VendorID_0120* seems to be highly associated with the process flow characteristics of the second cluster. The third cluster is dominated by *Consignment* cases. The fourth cluster contains many *Metal Containers & Lids* cases as well as cases from *VendorID*s *0404* and *0104*. Further analysis of the top five business areas listed as characteristics for each cluster confirms that these business areas indeed give a good overall idea of the cases allocated into each cluster.

Discovering Business Areas

We clustered four times for fixed cluster amounts of 2,3,5 and 10 - yielding a total of 20 clusters, and then consolidating the results into business area level using Definition 40. The top 20 of all these 9901 business areas ordered by their respective Business Area Contribution is shown in Table 4.10. Clearly, the business areas *Item Category = Consignment* and *Item Type = Consignment* have the most significant effect on the process flow. Looking at the actual process model, we see that *Consignment* cases completely avoid three of the five most common activities in the process, namely *Record Invoice Receipt*, *Vendor creates invoice*, and *Clear Invoice*. Similarly, the business area *Spend area text = Packaging* also has a high correlation with process flow characteristics. Analysis of the process model shows that, for example, 23% of *Packaging* cases contain activity *Receive Order Confirmation* compared to only 5% of the other cases. Further analysis of all the business areas listed in Table 4.10 shows that each of these areas has some distinctive process flow behavior that is more common in that area compared to the other business areas.

Clustering Summary for Case Attributes

Finally, Table 4.11 consolidates individual business areas into the Case Attribute level. *Item Type* having six distinct values and *Item Category*

Table 4.9. Clustering results based on Contribution

Cluster	Business Area a	Cluster Density	Total Density	Contri-bution
	Spend area text = Sales	0.36	0.26	0.11
Cluster1	Sub spend area text = Products for Re-sale	0.34	0.24	0.11
48% cases	Spend classification text = NPR	0.41	0.32	0.10
	Item Type = Standard	0.96	0.87	0.09
	Item Category = 3-way match, invoice before GR	0.95	0.88	0.07
	Spend area text = Packaging	0.65	0.44	0.21
Cluster2	Sub spend area text = Labels	0.39	0.24	0.16
33% cases	Spend classification text = PR	0.79	0.66	0.13
	Name = vendor_0119	0.14	0.05	0.08
	Vendor = vendorID_0120	0.14	0.05	0.08
	Item Category = Consignment	0.33	0.06	0.27
Cluster3	Item Type = Consignment	0.33	0.06	0.27
17% cases	Name = vendor_0185	0.09	0.02	0.08
	Vendor = vendorID_0188	0.09	0.02	0.08
	Item = 10	0.33	0.26	0.07
	Sub spend area text = Metal Contain-ers & Lids	0.19	0.08	0.11
Cluster4	Name = vendor_0393	0.09	0.01	0.08
1% cases	Vendor = vendorID_0404	0.09	0.01	0.08
	Name = vendor_0104	0.11	0.04	0.07
	Vendor = vendorID_0104	0.11	0.04	0.07
	Spend classification text = NPR	0.59	0.32	0.27
Cluster5	Spend area text = Sales	0.41	0.26	0.15
1% cases	GR-Based Inv. Verif. = TRUE	0.21	0.06	0.15
	Item Category = 3-way match, invoice after GR	0.21	0.06	0.15
	Sub spend area text = Products for Re-sale	0.38	0.24	0.14

Table 4.10. Top 20 Business areas with major effect to process flow

Business Area a	Contribution	nCases $n(C_a)$
Item Category = Consignment	0.051	576
Item Type = Consignment	0.051	576
Spend area text = Packaging	0.040	4382
Spend classification text = NPR	0.024	3175
Sub spend area text = Labels	0.022	2351
Spend area text = Sales	0.021	2574
Item Type = Standard	0.021	8740
Sub spend area text = Products for Resale	0.021	2390
Spend classification text = PR	0.019	6574
Item Category = 3-way match, invoice before GR	0.017	8760
Spend area text = Logistics	0.013	210
Item Type = Service	0.013	244
Item = 1	0.012	342
GR-Based Inv. Verif. = TRUE	0.012	623
Item Category = 3-way match, invoice after GR	0.012	625
Name = vendor_0119	0.007	549
Vendor = vendorID_0120	0.007	549
Sub spend area text = Road Packed	0.006	145
Name = vendor_0185	0.004	163
Vendor = vendorID_0188	0.004	163

with four distinct values, have the most significant effects on process flow characteristics. To confirm the validity of these results, we further analysed the materials provided in the BPI Challenge 2019 website, including the background information and submission reports [23]. It is clear that the *Item Type* and *Item Category* indeed can be regarded as the most important factors explaining the process flow behavior as they are specifically mentioned to *roughly divide the cases into four types of flows in the data*. It is also interesting to see that both the *Spend area text* and *Sub spend area text* have a significant effect on the process flow even though they have a much higher number of distinct values (19 and 115) compared to *Spend classification text*, which only has four distinct values.

4.3.2 Discussion

Summary of findings based on the previous case study results and the author's empirical evaluation of using our method in industrial process mining projects:

- Our method is capable of discovering those business areas that have the most significant effect on process execution. It provides valuable information to business people who are familiar with case attributes and attribute values, but not so familiar with the often technical event type names extracted from transactional system log files.

- Our method supports any available trace clustering method. Our case study shows that using the k-modes clustering algorithm with

Table 4.11. Case Attributes ordered by the effect on process flow

Case Attribute at	Contribution	Distinct Values $n(V_{at})$
Item Type	0.086	6
Item Category	0.080	4
Spend area text	0.077	19
Sub spend area text	0.056	115
Spend classification text	0.043	4
Name	0.025	798
Vendor	0.025	840
Item	0.016	167
GR-Based Inv. Verif.	0.012	2
Purchasing Document	0.002	7937
Document Type	0.000	3
Goods Receipt	0.000	2
Company	0.000	2
Source	0.000	1
Purch. Doc. Category name	0.000	1

activity and transition profiles provides good results.

- Clustering makes analysts aware that not all the cases in the process model are similar. Using the *Contribution%* measure to explain clustering results works well for explaining the clustering results to business people.

- The presented case study confirms that the identified business areas do indeed have distinctive process flow behavior—for example, missing activities, higher than average amounts of some particular activities, or a distinctive activity execution sequence. Using our method, the business analyst is able to divide the process model into smaller subsets and analyze them separately. The analysis of any process subset can be started by running the clustering to see if the cases are similar enough from process flow point of view.

- Clustering reduces the need for external subject matter business experts. It would ideal if there was someone on hand to explain everything, but in reality the subject matter experts are very busy, and some essential details are always likely to be forgotten by busy business people.

5. Conclusions

This dissertation studies business-related process analysis methods based on process mining data, with a focus on resource allocation, root cause analysis, change identification, and analysis of business area effects on process flow. The dissertation consists of eight publications and this introduction.

The methods presented in this dissertation can be used to:

1. Allocate business improvement resources

2. Identify process-related changes in business operations

3. Discover business areas that have a significant effect on process flow behavior

5.1 RQ1: How can process mining be used for resource allocation to maximize business improvement?

This dissertation presents an influence analysis method that can be used for effective resource allocation to maximize business improvement. Publication I presents the original influence analysis method as a root cause analysis method for focusing business improvements. Publication II extends the method to cover continuous problem variables and case-specific weighting, specifically aiming at reducing lead times and working capital analysis. The objective of this first research question is to determine the optimal scope for an improvement project to reduce the number of problems already discovered by other means, as presented in Section 1.1.1. The related problem setup is formalized with concepts *business problem*, *problem type*, *problem weighting*, *development project*, and *amount of resources* in Section 2.2.

The basic influence analysis method containing data preparation steps, options for the desired level of business improvement, and the calculation of corresponding interestingness measures is presented in Section 3.1 for binary problems, and further extended to cover continuous problems and

case-specific weights in Section 3.2. Publication V presents results from a distributed computing framework assessment aimed at finding options for using our method with massive amounts of data, briefly summarized in Section 3.5. Section 4.1 presents a case study with real-life data and summarizes our experiences in using the method.

5.2 RQ2: How can process mining be used to identify changes in business operations?

As presented in Publication III, the influence analysis presented in Publication I and Publication II can be extended to identify changes in business operations. Our objective is to discover various process changes and sort them in order of business significance, as shown in 1.1.2. The problem setup is formalized with concepts for *setting the review period*, *fast and slow changes*, *relevancy*, *root causes*, and *data quality detection* in Section 2.3.

Although the method for analyzing business process changes is based on influence analysis, it has two fundamental differences compared to the original influence analysis used in RQ1. As presented in Section 3.3, these extensions are the analysis of event-level data instead of process mining cases, and the discovery of changes related to time instead of the discovery of root causes related to a previously detected business problem. Section 4.2 presents a case study with real-life data and summarizes our experiences in using the method.

5.3 RQ3: How can business areas that have a significant effect on process flow behavior be discovered using clustering?

Publication IV presents a novel approach for discovering business areas that have a significant effect on process flow. The goal is to use clustering on historical data in order to group similar kinds of process instances into the same clusters, and then finding the business areas that correlate most with these identified clusters, as presented in 1.1.3. The problem setup is formalized with concepts *communicating clustering results to business analysts*, *finding the business areas that have a significant effect on process flow behavior*, and *further consolidating business area results to discover most significant case attributes* in Section 2.4.

Our method is based on clustering analysis for grouping the cases followed by influence analysis for discovering the business areas, as presented in Section 3.4. In Publication VI we present results for effective feature selection related to clustering and in Publication VII and Publication VIII we examine the idea of learning the relevant features using recurrent neu-

ral networks, briefly summarized in Sections 3.4.1 and 3.4.1. Section 4.3 presents a case study with real-life data and summarizes our experiences in using the method.

5.4 Future Work

To further increase the usability and applicability of our influence analysis method, several research ideas could be explored. Current influence analysis uses discrete variables for discovering root causes and business areas, and testing different techniques for discretization and grouping of individual values into subsets would provide new insights. This would be specifically beneficial for using the lead times of transitions between any particular event types as potential root causes. It would also provide further insight if decision mining techniques that form multiple component rules based on several attributes were combined with our influence analysis measures.

Individual business analysts often consider some of the discoveries generated by our methods very interesting, and some not so important. It would offer further insight if our method was combined with a machine learning based system that would learn which findings are considered relevant and which are ignored by the analyst. This kind of learning could take place in many levels, including: 1. Individual user using a specific process mining model; 2. Anybody using the same model; 3. One user using any process mining model; 4. All users within one organization using any models; 5. All users from hundreds of organizations internationally using the same cloud-based process mining environment.

Performance optimization is continuously needed to meet the growing customer requirements for big data analytics. In-memory computing should be explored in more detail to find methods for analyzing large datasets such as 100TB with considerably smaller RAM sizes such as 1TB.

As presented in this dissertation, the influence analysis reports correlations between various aspects. The user of the method is responsible for understanding whether the correlation is indeed a proper root-cause dependency based on the causality of the cause and effect or whether both aspects are just effects of a third cause. Causality in process mining has been studied utilizing the timestamps of the events to see how a change in original aspects at one time may have caused effects on other aspects at later times [34]. It would be interesting to include the causality in more robustly into the *Contribution* measure.

From the economic point of view, one important future work is to take the methods presented in this dissertation into active use in process mining analyst community for making processes more efficient, ensuring compliance, supporting robotic process automation, and driving successful digital

transformation. A survey exploring benefits and challenges related to the usage of our influence analysis based methods would provide useful insight and best practices.

Appendices

A. Customer Case Studies

This Appendix contains supplementary material from three industrial customer cases where the influence analysis has been in extensive use.

Metsä Board
1. "The process analysis delivered the needed visibility for Metsä Board to focus their improvement activities to the right areas."

2. "The overall goal of the development work was to improve customer satisfactionthrough better delivery accuracy, production efficiency, optimised stock rotationand reduced number of changes."

3. "The process insight and facts delivered by QPR ProcessAnalyzer were priceless. We were immediately able to focus our process improvement activities to the right things to reach the results our business needed. And not wasting time on trial and error." - Jari Vuori, Vice President Supply Chain, Metsä Board

4. Reaching optimal process performance requires continuous work and process owners now at Metsä Board regularly monitor their processes with QPR ProcessAnalyzer. They base their developmnent activities on facts, not on hunches. ... With the knowledge of where to target development efforts, Metsä Board can now concentrate on customer needs and deliver improved service experience.

EY
1. Using QPR ProcessAnalyzer we also get a comprehensive understanding of root causes and how to take corrective actions for improving operations.

2. "Root Cause analysis really works well in QPR ProcessAnalyzer", Stewart Wallace, Director, Risk Analytics, EY UK.

3. QPR ProcessAnalyzer's latest machine learning and artificial intelligence (AI) based advancements such as clustering analysis and

case level prediction lets business understand transaction-based data much faster.

4. "My favourite feature is conformance analysis, and business process modelling and then correlate it with the root cause analysis", Ricky Vachhani, Manager, Risk & Data Analytics, EY UK.

KBC

1. "With Root Cause Analysis, we found out why certain tasks are often done incorrectly." - Sander Van Lombeek, Team Lead, Commercial Credits, KBC Group

2. "QPR ProcessAnalyzer allows us to keep track of SLA situation in every subtask and find the root cause of inefficiency. To illustrate, we discovered a bottleneck caused by deficiency of the e-form system, doubling the promised processing time." - Christof De Groote, Service Manager Life Insurance, KBC Insurance

3. By using the holistic process mining view including bottlenecks and root causes, KBC is able to reduce throughput times

4. "I really really love the influence analysis. I think that's one of the major key benefits for me from the QPR ProcessAnalyzer tool." - Sander Van Lombeek, Team Lead, Commercial Credits, KBC Group

5. "I am also experimenting with clustering and predictive models." - Maaike Roekens, Credit Risk Model Manager, KBC Bank

With QPR ProcessAnalyzer Metsä Board knows where the bottlenecks are in real processes

Improving supply chain based on data - analysis of the real process

In 2011, Metsä Board started the initiative to improve predictability and harmonise operating models to better respond to customer needs and requests. As a firststep, the company wanted to understand the real status of their Order to Cash process. With the help of QPR ProcessAnalyzer they were able to get facts of process performance based on SAP data. The process analyses delivered the needed visibility for Metsä Board to focus their improvement activities to the right areas and ultimately deliver excellent customer experience.

Metsä Board is a leading European folding boxboard and white fresh forest fibre linerboard producer as well as a market pulp supplier. The company's sales network serves brand owners, carton printers, corrugated packaging manufacturers, printers and merchants. Metsä Board is headquartered in Finland. In 2016, the company's sales totaled EUR 1.7 billion, and it has approximately 2,500 employees. Metsä Board, part of Metsä Group, is listed on the NASDAQ OMX Helsinki.
www.metsaboard.com

The road to process harmonization

Metsä Board was looking to transform the way they manage their Order to Cash process and needed process insight from SAP data to plan relevant improvement activities. To support the goal of understanding how their Order to Cash process works in reality, Metsä Board got help from QPR and QPR ProcessAnalyzer in order to get an end-to-end process analysis from SAP.

Such analysis was not obtainable from the existing SAP reporting tools. The findings were priceless.

The process visualization showed clearly that the process performance was affected by changes that were made to orders mostly due to ad-hoc requests and as internal adjustments to already made changes.

This insight provided the basis for next steps. For the following two years Metsä Board focused internally on harmonizing their supply models and defining the target state of the Order to Cash process as well as how people should work around the process to ensure optimal efficiency.

The overall goal of the development work was to improve customer satisfaction through better delivery accuracy, production efficiency, optimised stock rotation and reduced number of changes.

www.qpr.com

Target state vs. the real process

In 2014, with the target state of the order to cash process well defined and deployed alongside the supply models, Metsä Board was ready to start measuring the impact they will have on their delivery accuracy and customer satisfaction

For measuring Metsä Board implemented QPR ProcessAnalyzer and the visibility gained to the real process gave the company the means to see how the process was adopted across organisation and how they could support change management. Process metrics were defined for effective monitoring of the process performance. With access to these facts, Metsä Board has been able to improve customer experience by keeping the delivery promise and drive internal efficiency by ensuring people work according to the process guidelines.
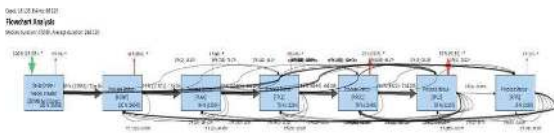
*The real process

Results of continuous process improvement

Reaching optimal process performance requires continuous work and process owners now at Metsä Board regularly monitor their processes with QPR ProcessAnalyzer. They base their development activities on facts, not on hunches.

The changes triggered by the analysis findings have all impacted the way of working rather than the support system itself. Facts delivered by QPR ProcessAnalyzer have also acted as an effective tool for change management and a means to engage the management. To ensure the optimal process performance, Metsä Board has focused on training their personnel on the agreed process guidelines not needing to invest in further SAP development.

With the knowledge of where to target development efforts, Metsä Board can now concentrate on customer needs and deliver improved service experience.
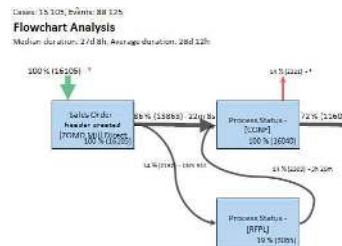
Finding: First confirmed delivery date is unsatisfactory for 19% of the orders

About QPR

QPR Software Plc offers the best services and software for measuring, analyzing, and improving business processes. QPR has more than 1,500 private and public sector customers across the globe in more than 50 countries. QPR's shares are listed on the NASDAQ OMX Helsinki Ltd.
QPR's software products offer customers innovative and efficient tools to discover any processes based on actual event data, to analyze root causes for problems and to continuously measure process performance. www.qpr.com

QPR provides insight to your business operations

Contact us for more information:
QPR Software Plc HQ (Finland): | Tel. +358 290 001 150

EY is using Process Mining for Risk Management and Internal Audit powered by QPR ProcessAnalyzer

Challenge

- Increasing pressure from regulators to use analytics for testing controls
- Reactive approach to errors
- Unaware of weaknesses and bottlenecks
- Risk management landscape changes
- How to know if controls are relevant
- Lack of precise internal control
- Slow process analytics

Benefits

- Obtain efficient view on internal process
- Assess quality of KPIs
- Give stakeholders greater confidence to make right business decisions
- Predict errors in preventive controls
- Prioritize actions based on risks
- Real-time monitoring of data
- Automated Root Cause analysis
- Identification of corrective actions

Risk Management Landscape

The risk management landscape is changing rapidly. The enormous growth of data demands internal control and audit team to be able to give strategic advice (on risk management and control) for stakeholders to make right business decisions and take advantage of opportunities. On the other hand, financial regulators and other industry bodies urge business to disclose its internal control process.

In order to stay one step ahead, businesses need to adapt fast. *"That is why understanding your processes is absolutely critical, and that's where process mining steps in."* – says Stewart Wallace, Director, Risk Analytics, EY UK.

EY

One of the biggest professional services firms in the world, headquartered in UK, providing assurance, tax, consulting and advisory services to businesses from their 700 offices in over 150 countries.

Employees: **270,000+**
Revenue: **36 billion USD**

www.ey.com

"Process mining allows you to look at every transaction, 100 percent coverage."

Stewart Wallace, Director, Risk Analytics, EY UK

QPR

"Textbook" process: Most simplified version of a process — **20%**

What people think Most commonly displayed and simplified process — **40%**

The reality: Actual complexity un-manageable without Data Mining — **100%**
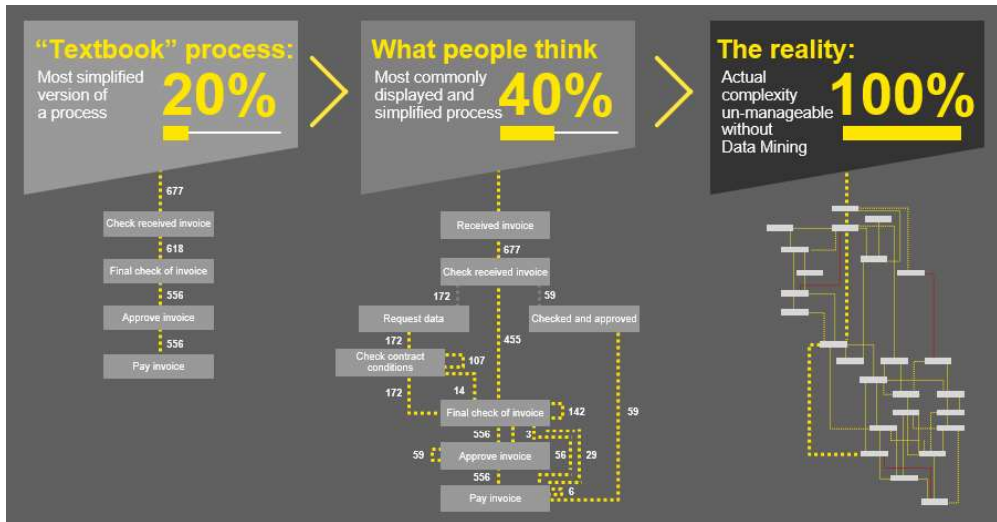
Give stakeholder confidence in making business decisions

Compared to traditional risk and control methodology, process mining proves to be an exceptional technological breakthrough. *"Before, you looked at the process documentation and talked to the people in order to form an opinion about what was happening. And we all know that would give you maybe 40 percent of the truth,"* says Stewart, *"but process mining is using data to tell you what's actually happening, 100 percent coverage. Process mining gives you a holistic view of all business processes showing what is going right or wrong. Using QPR ProcessAnalyzer we also get a comprehensive understanding of root causes and how to take corrective actions for improving operations."*

Analysis in less than one week

Using QPR ProcessAnalyzer, the EY UK Risk Analytics team has been able to complete an end-customer process mining analysis in less than one week. This includes time from initial data acquisition to model building, creating dashboards and presenting results to the customer. Complex end-to-end processes naturally take longer but typically one to two weeks is enough for any well-defined process mining analysis or iteration.

"QPR ProcessaAnalyzer helps us design data models and dashboards the fastest so that time to insight is the shortest." – says Stewart.

QPR

"We are able to identify areas where themes within the audit reports did not match with what was recorded in the audit tracking system."

Ricky Vachhani, Manager,
Risk & Data Analytics, EY UK

Obtain efficient view on internal control

As the business shifts to become more agile, resulting in constant changes in processes and operations, only the dynamism of process mining can keep up with these rapid changes.

"Traditional analytics is static and hard-coded. If we now have a new event in the middle of our process, process mining picks that up, but traditional analytics won't." – says Stewart.

★ Favourite QPR ProcessAnalyzer feature

"QPR is the one that I always turn to for ease of use and fastest time to insight. Root Cause analysis really works well in QPR ProcessAnalyzer"

Stewart Wallace, Director,
Risk Analytics, EY UK

Real-time data monitoring

QPR ProcessAnalyzer's latest machine learning and artificial intelligence (AI) based advancements such as clustering analysis and case level prediction lets business understand transaction-based data much faster. This helps business to identify almost instantly when transactions are going down a suboptimal path and alert to the process owner to take immediate corrective actions. From the internal control and audit perspectives, this approach brings much more practical and monumental benefits, compared to traditional way of working.

"With QPR ProcessAnalyzer we now embed process mining in real time, identify bottleneck instantly, and take actions much earlier, so the business gets more benefit, more actionable insights on a continuous basis" – says Stewart.
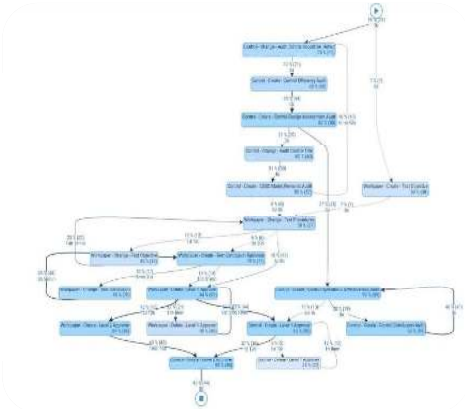
⭐ Favourite QPR ProcessAnalyzer feature

"My favorite feature is conformance analysis, and business process modelling and then correlate it with the root cause analysis"

Ricky Vachhani, Manager,
Risk & Data Analytics, EY UK

Internal Audit Process

Process mining can detect inefficiencies and bottlenecks also in the internal audit process itself. *"A large banking customer of EY asked our team to perform mining based analysis of their internal audit process. We were able to detect issues like audit reports showing no issues whilst event logs data showed multiple rounds of testing and approval"* says Ricky Vachhani, Manager , Risk & Data Analytics, EY UK.

QPR ProcessAnalyzer solution

- Risk Management and internal audit solutions built based on QPR ProcessAnalyzer dashboards
- Licenses covering EY UK employees and end customers
- QPR ProcessAnalyzer hosted in high security Azure datacenter
- Event log data from multiple systems including SAP, Oracle and audit tracking systems

---

QPR Software Plc

QPR Software Plc (Nasdaq Helsinki) provides process mining, performance management and enterprise architecture solutions for digital transformation, strategy execution and business process improvement in over 50 countries. QPR software allows customers to gain valuable insights for informed decisions that make a difference.

Dare to improve.

Read more at
www.qpr.com

KBC Group is reducing risks, automating processes, and improving customer satisfaction with QPR ProcessAnalyzer

Challenge

- Unknown reasons for process failures
- Too much manual work
- Unidentified costs in the back office
- High number of process variations
- Long process lead times
- Failing Service Level Agreements

Benefits

- Reduce operational risks
- Identify RPA opportunities
- Removed bottlenecks to meet SLAs
- Maximize efficiency front and back office
- Faster credit acceptance process
- Harmonized process for home loans

Reduce operational risks

Many risk management operations in KBC are centralized, making it quite difficult to find reasons for process failures occurring in local offices in 30 countries. *"QPR ProcessAnalyzer allows us to follow up worldwide execution of processes from the headquarter office. With Root Cause Analysis, we found out why certain tasks are often done incorrectly. These insights substantiate our decisions to automate said tasks and implement a four-eye principle."* - says Sander Van Lombeek, Team Lead, Commercial Credits, KBC Group.

KBC

A multi-channel bank-insurer, focusing on private clients and small and medium-sized enterprises headquartered in Belgium and operating in some 30 countries

Employees:	**42,000+**
Revenue:	**8 billion EUR**
Assests:	**290+ billion EUR**
Clients:	> **11.000.000**

www.kbc.com

> **"Process Mining is the technology that helps us to make data powerful."**
>
> Sander Van Lombeek, Team Lead, Commercial Credits, KBC Group

Identify RPA opportunities

QPR ProcessAnalyzer is helping KBC to detect unnecessary manual steps in credit acceptance process. *"Based on process mining analysis, we have started to automate the credit application checking process, so that a software robot collects more information to speed up credit approval decision."* - says Maaike Roekens, Credit Risk Model Manager at KBC Bank.

QPR

Remove bottlenecks to meet SLAs

"QPR ProcessAnalyzer allows us to keep track of SLA situation in every subtask and find the root cause of inefficiency. To illustrate, we discovered a bottleneck caused by deficiency of the e-form system, doubling the promised processing time." - says Christof De Groote, Service Manager Life Insurance, KBC Insurance.

Increase sales

QPR ProcessAnalyzer gives insights into client behavior on company level, which generates more end-customer leads to increase sales.

By using the holistic process mining view including bottlenecks and root causes, KBC is able to reduce throughput times. End customers now get fast service which also increases customer satisfaction.

"Process mining gives us real insights into the sales potential of our employees based on conversation ratios and working time, which helps us to automate bottleneck tasks and increase sales." – says Sander.

"With QPR ProcessAnalyzer, we were able to analyze the process in two or three hours, unlike three weeks in the past."

Christof De Groote, Service Manager, Life Insurance, KBC Insurance

Save processing costs

By leveraging QPR ProcessAnalyzer to assess their two top-selling branches of KBC Autolease, KBC identifies huge differences in processing costs. Root Cause Analysis in the software helps KBC maximize efficiency in both front and back office.

"Some dealers are extremely intransparent. The cost of that intransparency is high and has caused us quite a lot of work", says Sander. *"With the insights from process mining, we can now tackle inefficiency at the headquarters."*

"We now avoid unnecessary steps and waiting time between local branch offices and headquarters, which means better service for our customers."

Maaike Roekens, Credit Risk Model Manager KBC Bank

"I was immediately impressed. What I like about QPR Software is first of all it's extremely user-friendly."

Sander Van Lombeek, Team Lead, Commercial Credits, KBC Group

Faster credit acceptance

Understanding the actual process with variations makes it possible to streamline manual tasks, leading to faster credit acceptance process.

"We now avoid unnecessary steps and waiting time between local branch offices and headquarters, which means better service for our end customers." - says Maaike.

QPR

Better employee performance

In the past, KBC measured employee performance based on only the number of signed contracts. With QPR ProcessAnalyzer, they now can integrate time factor on individual process mining task level into performance measurement.

"We evaluated the performance of 1000 KBC employees with QPR ProcessAnalyzer to find out what are the time-consuming activities in the whole process." – says Sander.

⭐ Favourite QPR ProcessAnalyzer feature

"QPR ProcessAnalyzer is very intuitive - after 2-3 hours I was able to do experiments of my own. It's a very user-friendly product."

Christof De Groote, Service Manager, Life Insurance, KBC Insurance

Easy to collect data

QPR ProcessAnalyzer eases the process of collecting the required data from our in-house built banking systems. By dividing credit acceptance process activities into five phases, local branch activities are efficiently benchmarked with similar activities in the head office.

"QPR's process mining tool is a great way to get your facts together and ensure we all are talking on the same level" - says Maaike.

⭐ Favourite QPR ProcessAnalyzer feature

"I really really love the influence analysis. I think that's one of the major key benefits for me from the QPR ProcessAnalyzer tool."

Sander Van Lombeek, Team Lead, Commercial Credits, KBC Group

⭐ Favourite QPR ProcessAnalyzer feature

"Process discovery, Chart View and Dashboards are my favorite features. I am also experimenting with clustering and predictive models."

Maaike Roekens, Credit Risk Model Manager, KBC Bank

QPR ProcessAnalyzer solution
- Used in KBC Bank and KBC Insurance
- 20+ process mining models including main processes such as credit acceptance and customer termination
- Event log data loaded from main banking and insurance systems

QPR Software Plc

QPR Software Plc (Nasdaq Helsinki) provides process mining, performance management and enterprise architecture solutions for digital transformation, strategy execution and business process improvement in over 50 countries. QPR software allows customers to gain valuable insights for informed decisions that make a difference.

Dare to improve.

Read more at
www.qpr.com

B. QPR ProcessAnalyzer

This Appendix contains suplementary material regarding the implementation of the methods in the commercial process mining tool QPR ProcessAnalyzer.

QPR ProcessAnalyzer - Webpage
Figure A1 shows the commercial homepage of QPR ProcessAnalyzer product. The page `https://www.qpr.com/products/qpr-processanalyzer` can be used to access public case studies, feature related marketing material, recorded process mining webinars and other marketing related materials.
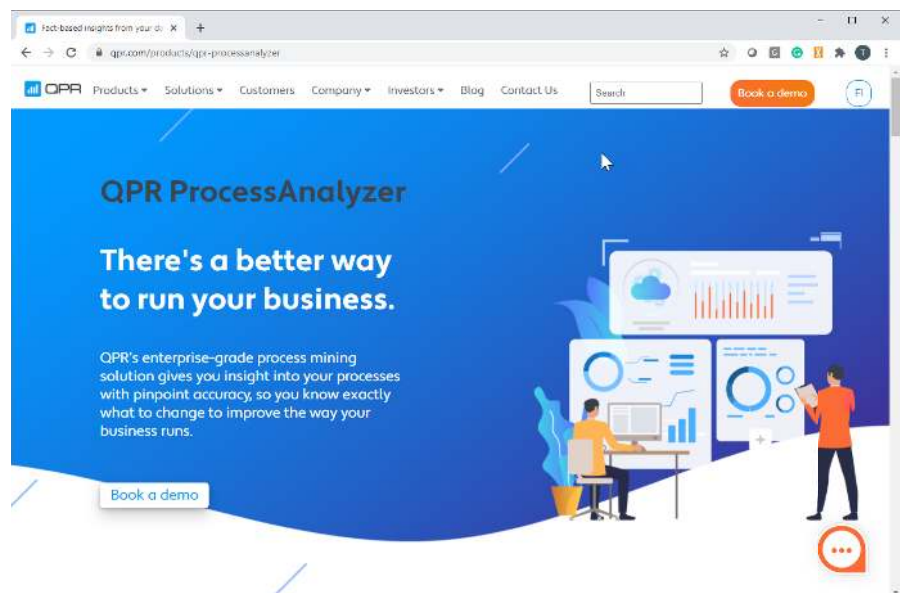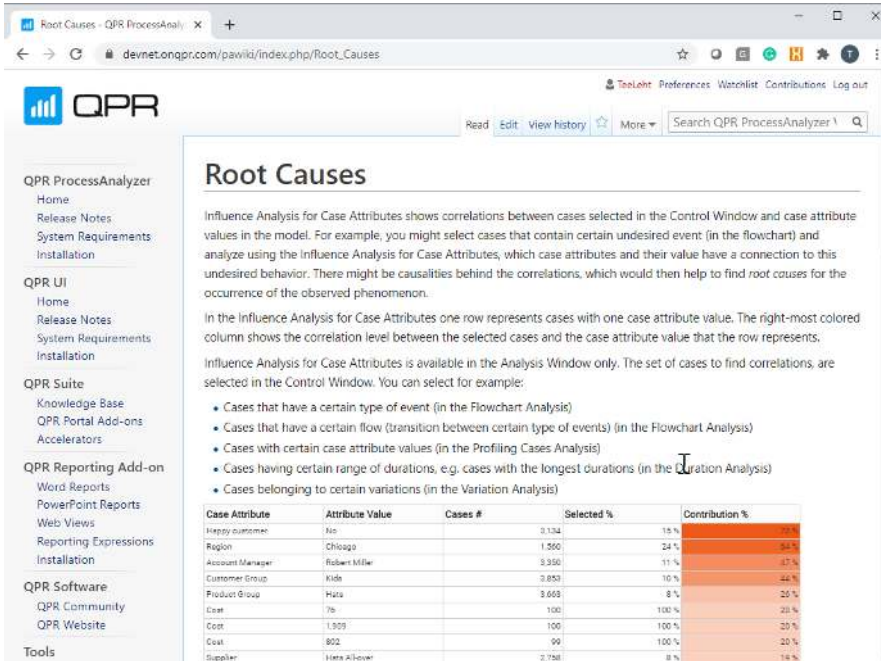

Figure A1. QPR ProcessAnalyzer - Webpage

QPR ProcessAnalyzer - Documentation
Figure A2 shows a documentation page for Influence Analysis taken from the public QPR ProcessAnalyzer wiki site `https://devnet.onqpr.com/pawiki/`

`index.php/Root_Causes`.

Figure A2. Screenshot

Influence Analysis - with details

Figure A3 shows the QPR ProcessAnalyzer user interface with one example of influence analysis feature. The user has selected a particular event type *Returned with Notification* as the criteria for root causes, ie. *Cases going through* the *Event Type Returned with Notification* are considered as problematic cases. The Red and Blue coloring conditional coloring is used for the Contribution values which are presented in the table. Red color correlates with the problem cases, ie. the problem cases are likely to have this property. Blue color correlates with the opposite, ie blue cases are NOT likely to belong to the set of problem cases.

Figure A3. Influence Analysis with details

Influence Analysis with flowchart

In the screenshot Figure A4, the coloring is also used in the flowchart so that red event types occur more often with problematic cases, blue event types less often, and white event type labels occur equal times in both sets. Also the Settings panel is visible showing the internal Analysis Type name *Root Causes*, parameter for the Maximum amount of positive and negtive items to be included (here 15 items), filtering for the case attribute (here *<All>* included), display option to show detailed columns or only the summary columns and the possibility to weight the results using the case-specific weights.
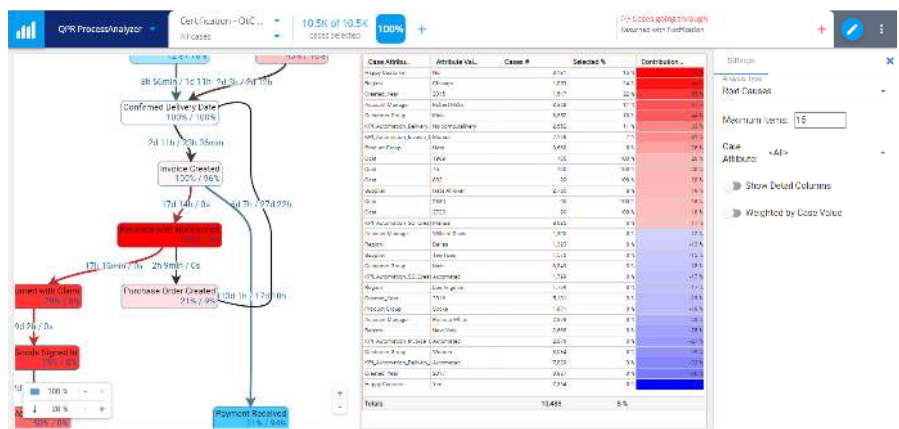
Figure A4. Influence Analysis with flowchart

Influence Analysis with tooltips

Figure A5 shows the Influence Analysis using a multiple line format with pop-up explanation texts. Here the user is analyzing cases where *Confrmed Delivery Date* is before *Delivery Completed*, as shown in the red root causes analysis rule in the top ribbon. Root causes are being shown for both the top negative and positive root causes based on the absolute value of contribution. As the user has clicked the property *Customer Group = Women*, a pop-up text is explaining the finding as: "Customer Group Women has the analyzed feature in 23% of the cases which is 8.0% more (184 cases) than on average. This root cause explains 15% (184 of 1,197 cases) of the occurrence of the analyzed feature, provides that there is a causal relationship.".
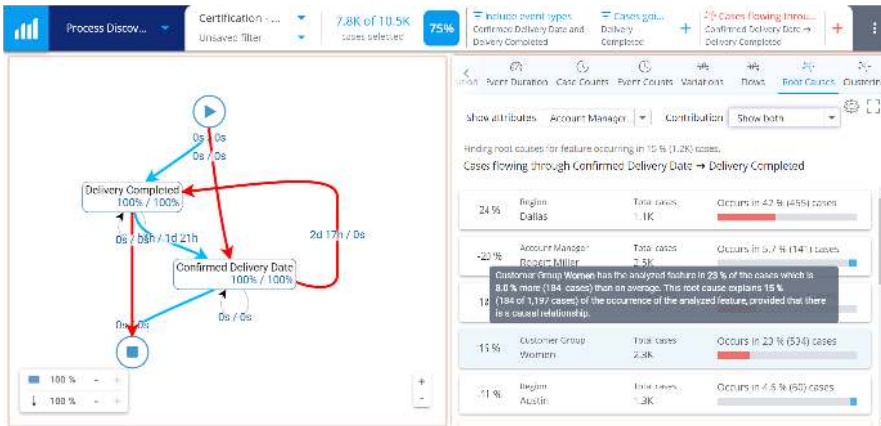
Figure A5. Influence Analysis with tooltips

Influence Analysis in MS Excel

Figure A6 shows the Influence Analysis in Microsoft Excel user interface. The original implementation released on 27.4.2012 executed analysis queries using Microsoft SQL Server computing engine and Microsoft Excel user interface clients.

Influence Analysis - Case Attributes — Cases: 7 832, Events: 15 670
Transitions=Confirmed Delivery Date->Delivery Completed

Total		7832	1197	6635	15%			
Case Attribute	Attribute Value	Cases #	Selected #	Compared #	Selected %	Difference %	Contribution #	Contribution %
Region	Dallas	1091	455	636	42%	26%	288	24%
Happy Customer	Yes	5709	1149	4560	20%	5%	276	23%
Account Manager	Patricia White	1779	491	1288	28%	12%	219	18%
Customer Group	Women	2290	534	1756	23%	8%	184	15%
KPI Automation Invoice Created	No Invoice	303	209	94	69%	54%	163	14%
Created_Year	2015	1072	279	793	26%	11%	115	10%
Cost	1019	100	100	0	100%	85%	85	7%
Cost	902	100	100	0	100%	85%	85	7%
Cost	758	99	97	2	98%	83%	82	7%
Cost	414	98	93	5	95%	80%	78	7%
Region	Los Angeles	1499	179	1320	12%	-3%	-50	-4%
Product Group	Shirts	1887	238	1649	13%	-3%	-50	-4%
Customer Group	Men	2569	341	2228	13%	-2%	-52	-4%
Created_Year	2017	2914	349	2565	12%	-3%	-96	-8%
Region	New York	2080	207	1873	10%	-5%	-111	-9%
Customer Group	Kids	2973	322	2651	11%	-4%	-132	-11%
Region	Austin	1292	60	1232	5%	-11%	-137	-11%
KPI Automation Invoice Created	Automated	2377	139	2238	6%	-9%	-224	-19%
Account Manager	Robert Miller	2479	141	2338	6%	-10%	-238	-20%
Happy Customer	No	2123	48	2075	2%	-13%	-276	-23%

Figure A6. Influence Analysis in MS Excel

Influence Analysis with Weights

Figure A7 shows the Influence Analysis results similar to the previous figure. Results contain an additional column *Case Value* after the case amounts. Contribution measures now explain how much each business area contributes to the total value of the problematic cases, instead of just the total amount of problematic cases..

Influence Analysis - Case Attributes — Cases: 7 832, Events: 15 670
Transitions=Confirmed Delivery Date->Delivery Completed

Total		7832	4730482	888530	3808952	19%			
Case Attribute	Attribute Value	Cases #	Cases Value	Selected Value	Compared Value	Selected %	Difference %	Contribution Value	Contribution %
Happy Customer	Yes	5709	3486495	853258	2633237	24%	5%	187303	21%
Customer Group	Women	2290	1433992	457838	976154	32%	13%	192949	22%
KPI_Automation_Invoice_Created	No Invoice	303	204739	195339	9400	95%	76%	150233	17%
Region	Dallas	1091	617367	269271	348096	44%	25%	151351	17%
Account Manager	Patricia White	1779	1247851	336904	900947	31%	11%	149017	17%
Cost	1019	100	101900	101900	0	100%	81%	82437	9%
Created_Year	2015	1072	653594	205029	448565	31%	12%	80190	9%
Cost	902	100	90200	90200	0	100%	81%	72971	8%
Product Group	Jeans	400	172600	96569	76031	56%	37%	63602	7%
Supplier	Global Jeans	400	172600	96569	76031	56%	37%	63602	7%
KPI_Automation_SO_Created	Automated	1499	942730	140082	802648	15%	-4%	-30983	-4%
Region	Los Angeles	1499	942730	140082	802648	15%	-4%	-30983	-4%
Product Group	Shirts	1887	947245	137880	809365	14%	-6%	-51048	-6%
Customer Group	Men	2569	1720676	251040	1469636	15%	5%	77616	9%
Region	New York	2080	1688034	228532	1459502	14%	-6%	-93889	-10%
Created_Year	2017	2914	1886604	265430	1621174	14%	-5%	-94919	-11%
Customer Group	Kids	2973	1554814	190652	1364162	12%	-7%	-106324	-12%
Account Manager	Robert Miller	2479	1533039	107199	1425840	7%	12%	185618	21%
Happy Customer	No	2123	1222967	46272	1176715	4%	-15%	-187323	-21%
KPI_Automation_Invoice_Created	Automated	2377	1539918	87530	1452388	6%	-13%	-200601	-23%

Figure A7. Influence Analysis with Weights

Duration Influence Analysis

Figure A8 shows the Influence Analysis conducted for the total duration of the case. One benefit of this kind of analysis is that it can be completed for every process mining model without any user selected parameter values. Analysis shows the average duration (here 45,58 days) as points out the Region = Dallas where average duration is 79,57 days for 1393 cases. If all cases in Dallas would have been completed in 45,58days, the total time saving would have been (79,57 - 45,58) x 1393, i.e., 47 353,22 days. This in turn is 44,04% of the total 143 335,58 days of overtime in the whole population. Business analysts can easily focus on the top and bottom rows to identify potential causes for long (or short) durations and lead times.

Figure A8. Duration Influence Analysis

Weighted Duration Influence Analysis

Figure A9 shows the Duration Influence Analysis results now weighted using a case specific value. Dallas is still the most significant reason for long durations. *Product Group = Hats* seems to contain at least some high value cases that took a long time to complete, since the wegted average duration is 54,92 days compared to equal weight average duration of 52,43 days.

Figure A9. Weighted Duration Influence Analysis

Clustering

Figure A10 shows the results of clustering analysis with parameters of five (5) clusters, five (5) rows per cluster, Most of the case attributes and all Event types. First cluster is highly characterized as cases that belong to *Customer Group = Kids* and go thourgh event types *Shipment Sent* and *Delivery Completed*.
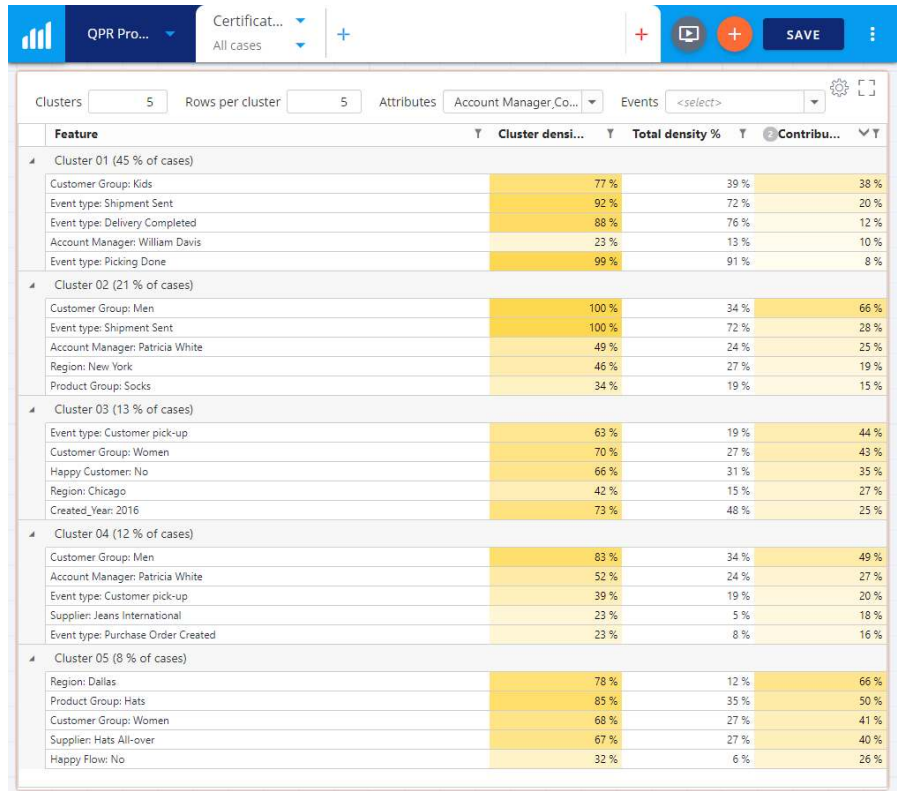
Figure A10. Clustering

Bibliography

[1] Van Der Aalst, W., Adriansyah, A., & Van Dongen, B. (2011, September). Causal nets: a modeling language tailored towards process discovery. In International conference on concurrency theory (pp. 28-42). Springer, Berlin, Heidelberg.

[2] Van Der Aalst, W., Adriansyah, A., De Medeiros, A. K. A., Arcieri, F., Baier, T., Blickle, T., ... & Burattin, A. (2011, August). Process mining manifesto. In International Conference on Business Process Management (pp. 169-194). Springer, Berlin, Heidelberg.

[3] Van Der Aalst, Wil. (2016). Process Mining - Data Science in Action, Second Edition. Springer-Verlag Berlin Heidelberg, ISBN 978-3-662-49850-7.

[4] Van Der Aalst, W. (2011). Process mining: discovery, conformance and enhancement of business processes (Vol. 2). Heidelberg: Springer. ISBN 978-3-642-19344-6

[5] Agarwal, R., & Srikant, R. (1994, September). Fast algorithms for mining association rules. In Proc. of the 20th VLDB Conference (pp. 487-499).

[6] Andersen, B. and Fagerhaug, T. (2006). Root cause analysis: simplified tools and techniques. ASQ Quality Press, 2006.

[7] Barbon Junior, S., Tavares, G. M., da Costa, V. G. T., Ceravolo, P., & Damiani, E. (2018, April). A Framework for Human-in-the-loop Monitoring of Concept-drift Detection in Event Log Stream. The Web Conference 2018 (pp. 319-326).

[8] Bay, S. D. and Pazzani, M. L. (2001). Detecting group differences: Mining contrast sets. Data Mining and Knowledge Discovery 5 (3), 213–246.

[9] Bazhenova, E., & Weske, M. (2016, September). Deriving decision models from process models by enhanced decision mining. In International conference on business process management (pp. 444-457). Springer, Cham.

[10] Biard, T., Le Mauff, A., Bigand, M., & Bourey, J. P. (2015, October). Separation of decision modeling from business process modeling using new "Decision Model and Notation"(DMN) for automating operational decision-making. In Working Conference on Virtual Enterprises (pp. 489-496). Springer, Cham.

[11] Bolt, A., de Leoni, M., & van der Aalst, W. M. (2018). Process variant comparison: using event logs to detect differences in behavior and business rules. Information Systems, 74, 53-66.

[12] Bose, R. J. C., van der Aalst, W. M., Žliobaitė, I., & Pechenizkiy, M. (2011). Handling concept drift in process mining. In International Conference on Advanced Information Systems Engineering (pp. 391-405). Springer, Berlin, Heidelberg.

[13] Bose, R. J. C., Van Der Aalst, W. M., Zliobaite, I., & Pechenizkiy, M. (2014). Dealing with concept drifts in process mining. IEEE transactions on neural networks and learning systems, 25(1), 154-171.

[14] Bose, R. J. C., & van der Aalst, W. M. (2013, April). Discovering signature patterns from event logs. In 2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM) (pp. 111-118). IEEE.

[15] Brin, S., Motwani, R., Ullman, J. D., & Tsur, S. (1997, June). Dynamic itemset counting and implication rules for market basket data. In Proceedings of the 1997 ACM SIGMOD international conference on Management of data (pp. 255-264).

[16] Buijs, J. C. A. M., & van der Aalst, W. M. P. (2017). Enabling interactive process analysis with process mining and visual analytics. BIOSTEC 2017, 573.

[17] Carmona, J., & Gavalda, R. (2012). Online techniques for dealing with concept drift in process mining. In International Symposium on Intelligent Data Analysis (pp. 90-102). Springer, Berlin, Heidelberg.

[18] Cheng, H., Yan, X., Han, J., & Hsu, C. W. (2007, April). Discriminative frequent pattern analysis for effective classification. In 2007 IEEE 23rd International Conference on Data Engineering (pp. 716-725). IEEE.

[19] Clauset, A., Shalizi, C. R., & Newman, M. E. (2009). Power-law distributions in empirical data. SIAM review, 51(4), 661-703.

[20] Devlin, B. A., & Murphy, P. T. (1988). An architecture for a business and information system. IBM systems Journal, 27(1), 60-80.

[21] Van Dongen, B.F. (2014). BPI Challenge 2014. Rabobank Nederland. Dataset. http://dx.doi.org/10.4121/uuid:c3e5d162-0cfd-4bb0-bd82-af5268819c35, 2014

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

[23] Van Dongen, B.F. (2019), Dataset BPI Challenge 2019. 4TU.Centre for Research Data. https://doi.org/10.4121/uuid:d06aff4b-79f0-45e6-8ec8-e19730c248f1

[24] De Leoni, M., van der Aalst, W. M., & Dees, M. (2016). A general process mining framework for correlating, predicting and clustering dynamic behavior based on event logs. Information Systems, 56, 235-257.

[25] Van Dongen, S. (2000). A Cluster Algorithm for Graphs. Technical report, National Research Institute for Mathematics and Computer Science in the Netherlands.

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

[27] Di Francescomarino, C., Dumas, M., Maggi, F. M., & Teinemaa, I. (2017). Clustering-based predictive process monitoring. IEEE transactions on services computing, 12(6), 896-909.

[28] Lqiang, G., & Hamilton, H. J. (2006). Interestingness Measures for Data Mining: A Survey [J]. ACM Computing Surveys (CSUR), (3), 61-93.

[29] Goldratt, E. M. (1990). Theory of constraints. Croton-on-Hudson: North River.

[30] Gröger, C., Niedermann, F., & Mitschang, B. (2012, July). Data mining-driven manufacturing process optimization. In Proceedings of the world congress on engineering (Vol. 3, pp. 4-6).

[31] Holte, R. C. (1993). Very simple classification rules perform well on most commonly used datasets. Machine learning, 11(1), 63-90.

[32] Hompes, B., Buijs, J. C., van der Aalst, W. M., Dixit, P., & Buurman, H. (2015). Detecting Change in Processes Using Comparative Trace Clustering. In SIMPDA (pp. 95-108).

[33] Hompes, B. F., Buijs, J. C., & van der Aalst, W. M. (2016, October). A generic framework for context-aware process performance analysis. In OTM Confederated International Conferences" On the Move to Meaningful Internet Systems" (pp. 300-317). Springer, Cham.

[34] Hompes, B. F. A., Maaradji, A., La Rosa, M., Dumas, M., Buijs, J. C. A. M., & van der Aalst, W. M. P. (2017). Discovering causal factors explaining business process performance.

[35] Inmon, William H. (2005). Building the Data Warehouse. John Wiley & Sons. ISBN: 978-0-764-59944-6.

[36] Kakas, A. C., Kowalski, R. A., & Toni, F. (1992). Abductive logic programming. Journal of logic and computation, 2(6), 719-770.

[37] De Koninck, P., De Weerdt, J., & vanden Broucke, S. K. (2017). Explaining clusterings of process instances. Data mining and knowledge discovery, 31(3), 774-808.

[38] De Leoni, M., Van der Aalst, W. M., & Dees, M. (2014, September). A general framework for correlating business process characteristics. In International Conference on Business Process Management (pp. 250-266). Springer, Cham.

[39] De Leoni, M., van der Aalst, W. M., & Dees, M. (2016). A general process mining framework for correlating, predicting and clustering dynamic behavior based on event logs. Information Systems, 56, 235-257.

[40] De Leoni, M., Dumas, M., & García-Bañuelos, L. (2013, March). Discovering branching conditions from business process execution logs. In International Conference on Fundamental Approaches to Software Engineering (pp. 114-129). Springer, Berlin, Heidelberg.

[41] Leontjeva, A., Conforti, R., Di Francescomarino, C., Dumas, M., & Maggi, F. M. (2016, September). Complex symbolic sequence encodings for predictive monitoring of business processes. In International Conference on Business Process Management (pp. 297-313). Springer, Cham.

[42] Maaradji, A., Dumas, M., La Rosa, M., & Ostovar, A. (2015). Fast and accurate business process drift detection. In International Conference on Business Process Management (pp. 406-422). Springer, Cham.

[43] Maisenbacher, M., & Weidlich, M. (2017, June). Handling concept drift in predictive process monitoring. In Services Computing (SCC), 2017 IEEE International Conference on (pp. 1-8). IEEE.

[44] Mayer-Schönberger, V., & Cukier, K. (2013). Big data: A revolution that will transform how we live, work, and think. Houghton Mifflin Harcourt.

[45] De Medeiros, A. K. A., Guzzo, A., Greco, G., Van Der Aalst, W. M., Weijters, A. J. M. M., Van Dongen, B. F., & Saccà, D. (2007). Process mining based on clustering: A quest for precision. In International Conference on Business Process Management (pp. 17-29). Springer, Berlin, Heidelberg.

[46] Nolle, T., Seeliger, A., & Mühlhäuser, M. (2018, September). Binet: Multivariate business process anomaly detection using deep learning. In International Conference on Business Process Management (pp. 271-287). Springer, Cham.

[47] Ostovar, A., Leemans, S. J., & Rosa, M. L. (2020). Robust drift characterization from event streams of business processes. ACM Transactions on Knowledge Discovery from Data (TKDD), 14(3), 1-57.

[48] Pearl, J. (2009). Causality. Cambridge university press.

[49] Piatetsky-Shapiro, G. (1991). Discovery, analysis, and presentation of strong rules. Knowledge discovery in databases, 229-238.

[50] Piatetsky-Shapiro, G., & Matheus, C. J. (1994, July). The interestingness of deviations. In Proceedings of the AAAI-94 workshop on Knowledge Discovery in Databases (Vol. 1, pp. 25-36).

[51] Quinlan, J. R. (1986). Induction of decision trees. Machine learning, 1(1), 81-106.

[52] Rozinat, A., & van der Aalst, W. M. (2006, September). Decision mining in ProM. In International Conference on Business Process Management (pp. 420-425). Springer, Berlin, Heidelberg.

[53] Seeliger, A., Nolle, T., & Mühlhäuser, M. (2018, September). Finding Structure in the Unstructured: Hybrid Feature Set Clustering for Process Discovery. In International Conference on Business Process Management (pp. 288-304). Springer, Cham.

[54] Song, M., Günther, C. W., & Van der Aalst, W. M. (2008). Trace clustering in process mining. In International Conference on Business Process Management (pp. 109-120). Springer, Berlin, Heidelberg.

[55] Souza, C. R. (2014). The accord .NET framework. São Carlos, Brazil. http://accord-framework.net

[56] Suriadi, Suriadi, Ouyang, Chun, van der Aalst, Wil M.P., & ter Hofstede, Arthur (2013) Root cause analysis with enriched process logs.

Lecture Notes in Business Information Processing [Business Process Management Workshops: BPM 2012 International Work-shops Revised Papers], 132, pp. 174-186.

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

[58] Teinemaa, I., Tax, N., de Leoni, M., Dumas, M., & Maggi, F. M. (2018, September). Alarm-based prescriptive process monitoring. In International Conference on Business Process Management (pp. 91-107). Springer, Cham.

[59] Thaler, T., Ternis, S. F., Fettke, P., & Loos, P. (2015). A Comparative Analysis of Process Instance Cluster Techniques. Wirtschaftsinformatik, 2015, 423-437.

[60] Vanjoki, V. (2013). Automated Purchase to Pay Process Value Modeling and Comparative Process Speeds. Lappeenranta University of Technology, 2013.

[61] Webb, G. I., Butler, S., & Newlands, D. (2003, August). On detecting differences between groups. In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 256-265).

[62] Webb, G. I., Hyde, R., Cao, H., Nguyen, H. L., & Petitjean, F. (2016). Characterizing concept drift. Data Mining and Knowledge Discovery, 30(4), 964-994.

[63] Verenich, I., Dumas, M., La Rosa, M., Maggi, F. M., Chasovskyi, D., & Rozumnyi, A. (2016). Tell me what's ahead? Predicting remaining activity sequences of business process instances.

[64] Wetzstein, B., Leitner, P., Rosenberg, F., Brandic, I., Dustdar, S., & Leymann, F. (2009, September). Monitoring and analyzing influential factors of business process performance. In 2009 IEEE International Enterprise Distributed Object Computing Conference (pp. 141-150). IEEE.

[65] Žliobaitė, I., Pechenizkiy, M., & Gama, J. (2016). An overview of concept drift applications. In Big data analysis: new algorithms for a new society (pp. 91-114). Springer, Cham.

[66] Yeshchenko, A., Di Ciccio, C., Mendling, J., & Polyvyanyy, A. (2019, November). Comprehensive process drift detection with visual analytics. In International Conference on Conceptual Modeling (pp. 119-135). Springer, Cham.

[67] Zaki, M. J. (2000). Scalable algorithms for association mining. IEEE transactions on knowledge and data engineering, 12(3), 372-390.

[68] Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., ... & Niu, P. (2020). A novel coronavirus from patients with pneumonia in China, 2019. New England Journal of Medicine.

Errata

Publication IV

Corrected text in page 10. Should use the subscript E_p instead of incorrect subscript E_a.

Correct version:

Similarly, the problem density for BiCo of subset E_a is the density of events belonging to the set E_a within the review period events E_p as $\rho_a = \frac{|E_p \cap E_a|}{|E_p|} = \frac{\sum\limits_{e_j \in (E_p \cap E_a)} 1}{\sum\limits_{e_j \in E_p} 1}$ derived from Equation 3.3.

Incorrect version in the publication:

Similarly, the average problem density for BiCo of subset E_a is $\rho_a = \frac{|E_p \cap E_a|}{|E_a|} = \frac{\sum\limits_{e_j \in (E_p \cap E_a)} 1}{\sum\limits_{e_j \in E_a} 1}$ derived from Equation 3.3.

The ability to improve processes is essential for every organization. Process mining provides a fact-based understanding of actual processes in the form of discovered process diagrams, bottlenecks, compliance issues, and other operational problems. Organizations need to carry out accurate root cause analysis and efficient allocation of development resources to improve the process and reduce problems.

This work presents a novel influence analysis method to improve the allocation of development resources, detect process changes, and discover business areas that have a significant effect on process flow. The method combines the usage of process mining analysis with probability-based objective measures and analysis of deviations. The method is specially designed for business analysts, process owners, line managers, and auditors in large organizations, to be used as a set of interactive root cause analyses and benchmark reports. Methods and algorithms are presented for analyzing both binary problems where each case is either successful or non-successful, and continuous variables, including process lead times and costs. A method for using case-specific weights to take into consideration the relative business importance of each case is also presented. This work also includes data preparation methods and best practices for acquiring relevant data of business operations in the event log format.

**BUSINESS +
ECONOMY**

**ART +
DESIGN +
ARCHITECTURE**

**SCIENCE +
TECHNOLOGY**

CROSSOVER

**DOCTORAL
DISSERTATIONS**