



Hochschule
Albstadt-Sigmaringen
Albstadt-Sigmaringen University

University of Applied Science

Faculty of Informatics

Bachelor's Thesis

Business Process Management and Process Mining

Author: Guillermo Gutiérrez Román

Director: Professor Dr. Bernd Stauss

February 22nd, 2018

1. INTRODUCTION	3
2. BUSINESS PROCESS MANAGEMENT	3
2.1. BUSINESS PROCESS	3
2.2. WHAT IS BUSINESS PROCESS MANAGEMENT	5
2.3. DIFFERENT PERSPECTIVES OF BUSINESS PROCESS MANAGEMENT (BPM)	5
2.3.1. BPM IS A CORE INTERNAL CAPABILITY	5
2.3.2. BPM ADDRESSES THE DELIVERY OF VALUE TO CUSTOMER	6
2.3.3. BPM ADDRESSES END-TO-END WORK AND THE ORCHESTRATION OF ACTIVITIES ACROSS BUSINESS FUNCTIONS	6
2.3.4. BPM ADDRESSES WHAT, WHERE, WHEN, WHY AND HOW WORK IS DONE, AND WHO IS RESPONSIBLE FOR PERFORMING IT	7
2.3.5. BPM MANAGES PROCESS IN A CLOSED-LOOP CYCLE TO ENABLE CONTINUOUS IMPROVEMENT.	7
2.4. HISTORY AND DEVELOPMENT OF BUSINESS PROCESS MANAGEMENT ALONG THE TIME	8
2.5. STATE OF THE ART OF BUSINESS PROCESS MANAGEMENT	11
3. BUSINESS PROCESS MODELS	14
3.1. PETRI NETS	14
3.2. WORKFLOW NETS	15
3.3. SWIM LANES	16
3.4. UNIFIED MODELING LANGUAGE UML	16
3.5. IDEF	17
3.6. BUSINESS PROCESS MODEL & NOTATION BPMN	18
3.7. EVENT PROCESS CHAIN, EPC	19
4. PROCESS MINING	21
4.1. WHY PROCESS MINING IS USEFUL FOR BPM	21
4.2. WHAT IS PROCESS MINING	22
4.3. DATA SCIENCE AND PROCESS SCIENCE	23
4.4. HISTORY OF PM	25
4.5. PROCESS MINING CONCEPTS	27
4.6. TYPES OF PROCESS MINING	31
4.7. DISCOVERY TECHNIQUES, α -ALGORITHM	31
4.7.1. α -ALGORITHM	32
4.7.2. ADVANCED TECHNIQUES	37
5. CONFORMANCE CHECKING	39
5.1. MAIN USES OF CONFORMANCE CHECKING: BUSINESS ALIGNMENT AND AUDITORY	40
5.2. TECHNIQUES	41
5.2.1. COMPARING FOOTPRINTS	41
5.2.2. TOKEN REPLAY	45
5.2.3. ALIGNMENTS	52
6. APPLICATION OF PROCESS MINING TECHNIQUES ON A DATA LOG GIVEN	57
6.1. PRESENTATION OF THE PROBLEM AND DATA	57
6.2. DATA PRE-TREATMENT	58
6.3. APPLICATION OF PROCESS MINING TECHNIQUES.	65
6.4. DISCUSSION OF THE RESULTS AND RECOMMENDATIONS.	76
7. CONCLUSION	76
8. BIBLIOGRAPHY	78

1. Introduction

This Bachelor's Thesis has been written aiming to answer three questions:

- What is Business Process Management?
- What is Process Mining?
- How a Process Mining project is managed?

These two concepts are, maybe, the newest disciplines within the fields of Process Management and Data Science respectively nowadays.

Business Process Management is a broad discipline which surrounds knowledge from information technology, management sciences and industrial engineering, and during the last decade, has become a standard for managing and improving business processes. On the other hand, Process Mining is a discipline belonging to Data Science study field, it was born few years ago and provides, through different software solutions, the tools for discovering business processes, checking the conformance between the real process and the modelled, and enhancing the current business processes in order to get a better explanation of the real business processes' behaviour.

Nevertheless, these two concepts are not independent one another since Process Mining is orientated to analyse and improve business process, making easier the main purpose of Business Process Management, creating value for the enterprises through the proper managing of their business processes.

This literature will provide a background of both disciplines, which is vital at time to get familiar with the current trends about management and data science.

In order to achieve this goal, the thesis is structured as follows:

Firstly, Business Process Management is introduced, its different perspectives, development of the discipline along the time, as well as the enablers that allow making this discipline, a key instrument at time to manage the business processes.

Secondly, it will be introduced the concept of Process Mining in depth. The three types of process mining, Discovery, Conformance Checking and Enhancement will be discussed emphasising on Discovery and Conformance Checking, moreover, it will be explained the main techniques about both of them.

Last but not least, the thesis will conclude with the analysis of a real Event Log extracted from SAP where it will be tried to discover a business process.

2. Business Process Management

2.1. Business Process

A business process is a linked set of activities and task that, once completed, accomplish an organizational goal. An illustrative example of business process is the following one:

SALES PROCESS

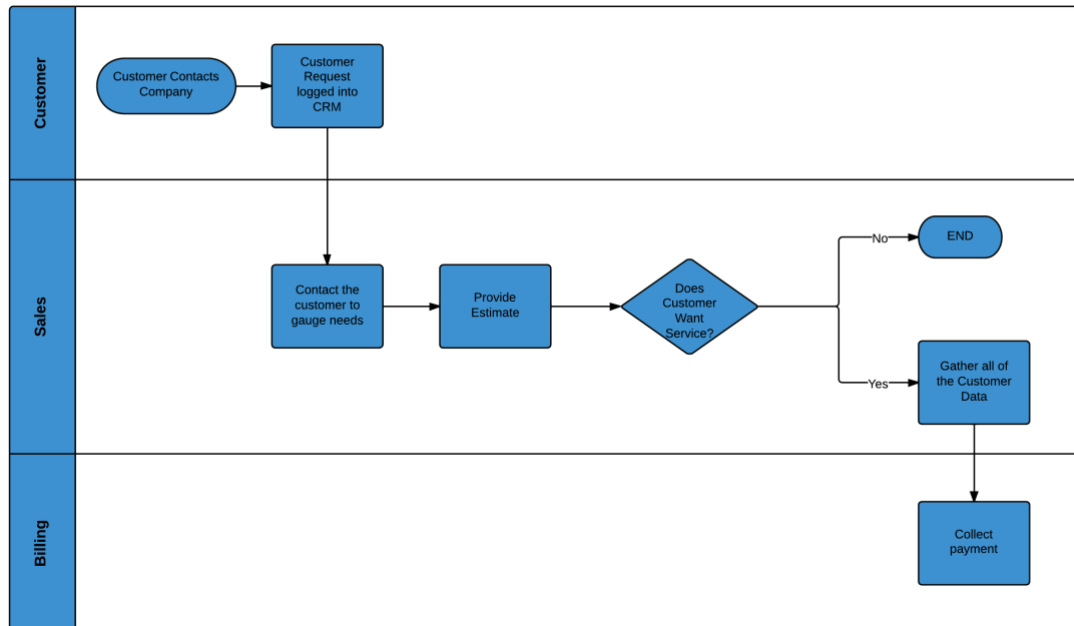


Figure 1. Example of a business process. Source: [1]

In the example, it is shown the different activities, as well as the different stakeholders that intervene during a sales process.

Within the organizations, business processes are split into different categories: Operational or primary processes, supporting or secondary processes, and management processes.

Name	Description	Example
Operational processes	Operational processes represent the core business activities and value chain. They will create value for the customer by producing products or services.	Take customer order Develop the products
Supporting processes	Supporting processes are vital for the company since they guarantee a proper performance of the operational processes. However, they do not provide value to the customer directly.	Human resources Accounting
Management processes	Management processes cover all activities related with leading, setting goals, and organizing within the company.	Budgeting Strategic planning Decision making

Table 1 Types of business processes. Source: Own elaboration based on: [2].

2.2. What is Business Process Management

Business Process Management is a management discipline which presumes that enterprise's organizational goals can be achieved through the definition, engineering, control and dedication to continuous improvement of business process [3]. The main targets BPM try to achieve are [4]:

Target to improve	Definition
Business Agility	Capacity of an enterprise to front and adapt for changes in its environment through changing their integrated process.
Effectiveness	Capability for achieving the strategic aims.
Efficiency	Relation between the obtained results and needed resources.

Table 2 Improvements reached with the application of BPM. Source: own elaboration based on: [3].

Although the definition of Business Process Management is a good start, Business Process Management includes many perspectives which must be explained for the comprehension of the concept in its fullness. [3]

2.3. Different perspectives of Business Process Management (BPM)

2.3.1. BPM is a Core Internal Capability

A successfully implementation of Business Process Management, should provide the organizations the capability to reach the strategic objectives through the adoption of and creation of three vital points.

Point	Function	Example
New process	The creation of processes which support the management of business process.	The monitoring and control of business process execution. The continuous improvement of business processes over time in response of external changes.
Specific roles	New roles and people who engaged in the management of business process.	Process Architects responsible for definition and design of process. Process Analyst responsible for build, deployment, monitoring and optimization of business process.
Specialized technologies	Supporting and making easier the management of business process.	Design business process for deployment. Execute business process in operations.

Table 3 The three pillars of BPM Source: own elaboration based on [3].

2.3.2. BPM addresses the delivery of value to customer

Regardless of whether an organization is for-profit, not-for-profit, or a government entity, the main organization's purpose is to deliver value to customer, in form of products and services, either external to the organization customers or customers between functions within an enterprise.

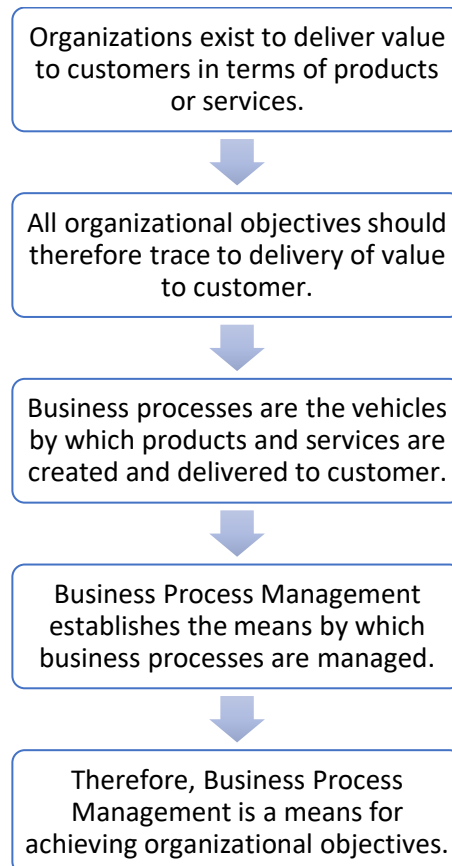


Figure 2. Source: own elaboration based on [3]

Organizations successful in business process management instil and foster a culture of customer focus at the enterprise level, the functional level, and down through the role level.

Business Process Management is about optimizing the means by which this value is delivered.

2.3.3. BPM addresses end-to-end work and the orchestration of activities across business functions

Business Process Management and Functional Management disciplines must cohabit and work together for the organization to remain competitively viable.

Discipline	Description
Functional Management	Ensures execution of the functional disciplines required to produce the organization's products and services.
Business Process Management	Ensures work is coordinated across these functions, in order to deliver products and services in the most effective and efficient manner possible.

Table 4 Disciplines within organizations Source: own elaboration based on [3].

2.3.4. BPM addresses What, Where, When, Why and How work is done, and Who is responsible for performing it

A Business Process Management discipline must accommodate the means by which What work is done, How work is done, When, Where, Why, by Whom.

A well-structured process definition will provide the right amount of visibility and detail to the various consumers of this information, potentially across all levels of the organization. That is why a process definition should be fit for purpose and fit for use.

Objective of Information	Description
Fit for Purpose	The process definition contains all necessary information to answer the Who, What, Where, When, Why, and How questions it is intended to address.
Fit for Use	The process definition is structured to represent this information in the most efficient and effective manner possible, considering the needs of the recipient.

Table 5 Amount of detail should be aligned with these objectives. Source: own elaboration based on [3].

2.3.5. BPM manages process in a closed-loop cycle to enable continuous improvement.

Organizations with mature BPM capabilities lead their processes through a closed-loop cycle that addresses the planning, design, implementation, execution, measurement, control, and continuous improvement of business processes.

Because of its simplicity, PDCA Cycle can be used as a good model for the implementation of every single stage.

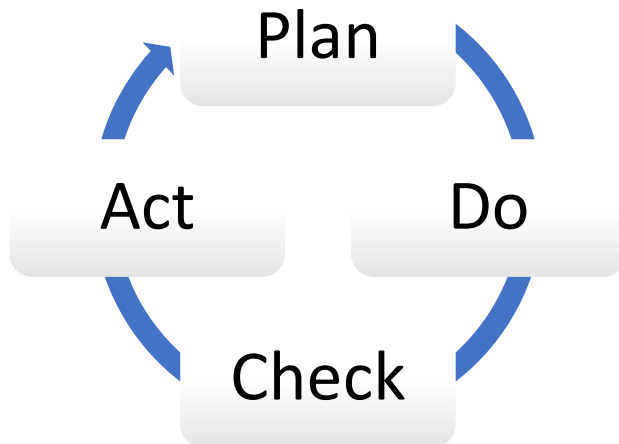


Figure 3. PDCA cycle Source: own elaboration based on: [3]

Phase	Objective
Plan	Aligning business process context and internal process with the organization's strategic objectives
Do	Deployment of the process per the specifications developed in "Plan" phase and commit the process to the operations.
Check	Process performance is compared against performance expectations.
Act	Making determinations and react accordingly to process data collected during the "Check" phase. This phase allows maintenance of process integrity in spite of environmental changes and ensures continual improving to reach new performance goals over time.

Table 6 PDCA applied to BPM Source: own elaboration based on [3].

At this point, it is easy to notice that BPM is much more than a theoretical definition. Therefore, BPM is a complex discipline, which takes many years to adopt and implement, and requires several changes within the organization, such as the commitment from top-to-bottom organization, it means, from the executive leadership which defines and support the practice of BPM, to the functional managers who must collaborate with process owners on the design and execution of business processes, and the operations staff who must work in virtual team in order to ensure value delivery to the end customer. The creation of new roles and the support with the Information Technologies are also vital in order to achieve an effective implementation of BPM.

2.4. History and development of Business Process Management along the time

It is true that BPM is a very recent concept, if organizations nowadays are able to manage, optimize, and adapt their business process with the efficiency they do, it is due to the contribution of many people and the evolution of the different management disciplines along the history. The table below shows what were the most important “waves” of development that have brought BPM what it is nowadays.

Phase	Time	Focus	Business	Technology	Tools/Enablers
Industrial Age	1750 - 1960s	-Specialization of Labour -Task Productivity -Cost Reduction	-Functional Hierarchies -Command & Control -Assembly Line	-Mechanization -Standardization -Record- keeping	-Scientific Management -PDCA Improvement Cycle -Financial Modeling
Information Age					
1st Wave Process improvement	70s-80s	-Quality Management -Continuous Flow -Task Efficiency	-Multi-Industry -Enterprises -Line of Business -Organization -Mergers & Acquisitions	-Computerized Automation -Management -Information Systems -MRP	-TQM -Statistical Process Control -Process Improvement Methods
2nd Wave Process Reengineering	1990s	-Business via Internet -Business process reengineering -Business Innovation	-Flat Organization -End-to-end Processes -Value Propositions – Speed to Market, Customer Intimacy, Operational Excellence	-Enterprise Architecture -ERP -CRM -Supply Chain Mgt	-Activity Based Costing -Six Sigma -Buy vs. build -Process Re-design/ Reengineering Methods
3th Wave Business Process Management	2000+	-Assessment, Adaptability, & Agility -24X7 Global Business -Continual Transformation	-Networked Organization -Hyper Competition -Market Growth Driven -Process Effectiveness over Resource Efficiency -Organizational Effectiveness over Operational Efficiency	-Enterprise Application Integration -Service Oriented Architecture -Performance Management software -BPM Systems	-Balanced Scorecard -Self Service & Personalization -Outsourcing, Co-Sourcing, In-sourcing -BPM Methods

Table 7 BPM's history line. Source: [5].

As far back as 1911, Frederick Winslow Taylor (1856-1915) introduced the basics of the management culture. The result of his work was a scientific point of view of the workflow and the improvement of process and workers based in analysis and results. The first wave covered the Just in Time philosophy adoption by American companies. The increasing use of the computers, allowed the statistical software development and related data gathering techniques that measured, gathered, and interpreted results. During the second wave appeared more efficient data managing software. On the other hand, organizations turned from corporate mission focus and brainstorming into cross-functional teams, changing the “how” by “why”. The third wave began in the mid 1990s and continues in the present as the “coming of age” of process-focused business. Technology is shifting from being a process driver to a process enabler. The identity of the customer changed from markets to individuals with customized solutions. Just-in-time manufacturing of the first wave led to just-in-time supply chains of the third wave, with the accompanying need to understand processes across disparate enterprises.

2.5.State of the art of Business Process Management

Nowadays BPM has become the paradigm of organization’s management all over the world. Thus, there is a big world around this new management discipline, what surround from universities that are already providing formation programmes in BPM to big companies which are adopting the standards of this philosophy as well as many BPM’s associations and consultant companies which try to lead organizations during the adoption of BPM. Some facts are below exposed in order to support the previous ideas. Worldwide known names in fields of Leadership, Management, Human Development and Organizational Development, have recognized that BPM brings to the companies the resources for a quick adaptation to the needs of customers, adoption the new IT, and competitive advance.

Name	Career	Quote
Michael Martin Hammer	American engineer, management author, and a former professor of computer science at the Massachusetts Institute of Technology (MIT), known as one of the founders of the management theory of Business process reengineering (BPR).	"BPM is the improvement of products and services using structured services optimization based on systematic design and management of business process". [6]
Robert Samuel Kaplan	American accounting academic, and Emeritus Professor of Leadership Development at the Harvard Business School.	"Executives must lead by adapting to dynamic, highly competitive environments, communicating vision and strategy to employees, and inspiring employees to innovate to achieve organizational objectives". [7]

Peter Michael Senge	American systems scientist who is a senior lecturer at the MIT Sloan School of Management.	"If people do not share a common vision, and do not share common mental models about the business reality within which they operate, empowering people will only increase organizational stress and the burden of management to maintain coherence and direction". [8]
---------------------	--	--

Table 8 Some representative important characters who support BPM. Source: own elaboration based on: [6] [7] [8].

International organizations and universities try to develop and consolidate theories and offer formation programmes about BPM.

Name	Location	Services
Club BPM [9]	Latin America and Spain	Spreading and promoting the BPM as the most profitable discipline that Spanish and Latin American organization could adopt.
		Provide professional with its BPM: RAD (Rapid Analysis & Design) methodology in order to a faster implantation of BPM discipline.
		Consulting during the BPM implantation.
		Provide formation and certification in BPM discipline
		Continuous researching in BPM evolution.
ABPMP [10]	Global	Education and Certification in BPM competency model
		Training providers, organizations and Educational Institutions that ABPMP International Board has approved to offer training in BPM knowledge areas.
BPM Institute [11]	United States of America	Formation and Certification in 6 different areas related with BPM: Agile Business Analysis Business Architecture Business Process Management Digital Decisioning & Analytics Operational Excellence SOA
BPM Centre [12]	Eindhoven University of Technology TU/e (Netherlands), and Queensland University of Technology QUT (Australia)	It is a collaborative virtual research centre where it can be found many books, conferences, and research in BPM

OMG BPMN [13]	Global	OMG is the responsible of the standardization of BPMN (Business Process Model & Notation), and support another modeling languages, such as UML (Undefined Modeling Language).
---------------	--------	---

Table 9 Some BPM association around the worlds. Source: own elaboration based on [9] [10] [11] [12] [13].

IT Organizations are investing many resources for the development of BPM suites and supporting software.

Name	Product
IBM	IBM Business Process Manager [14]
SAP	SAP NetWeaver Business Process Management [15]
Oracle	Oracle BPM Suite [16]
Aura Portal	Aurora BPM Suite [17]
Camunda	Open source platform for BPM and workflow. [18] It is possible to create: <ul style="list-style-type: none"> • BPMN workflows. • CMMN cases. • DMN decisions.
Signavio	Signavio Process Manager [19]

Table 10 Technology Expertise Organizations which provides BPM solutions. Source: own elaboration based on [14] [15] [16] [17] [18] [19].

Most international companies are adopting this management discipline in order to solve different organizational problems. There are some representative examples in the following table.

Enterprise	Supplier	Problem	Results
Adidas [20]	Bizagi	To deliver process automation across various departments including supply chain, marketing, finance, retail and e-commerce.	Adidas connected 500 sales operations, 400 factories and multiple SAP and supplier systems to streamline over 5,000 purchase order changes per month, resulting in operational cost savings of 60%. [20]
Toyota Spain [21]	Aura Portal	Toyota wanted to stablish the Environmental Management	GEA-Toyota. It is a management system based in BPM which

		Systems standard ISO 14001 in every concessionaire	resulted in: 98% compliance the standard, improvement of 70% in wasted time on intern auditions, improvement of 80% on intern communication flow, improvement of 88% on Analysis Capability
Bayer Healthcare [22]	ARIS Process Performance Manager	Bayer needed to monitor SAP timely data entry and inventory accurately without having a lot of resources. The company was facing those challenges with its existing scorecard implementation based on SAP KPI.	After implantation, Bayer Healthcare achieved: Daily reporting of business process performance, process analysis capabilities in the hands of managers, elimination of manual creation of SAP Scorecard, scalable application can be used for many other processes, ability to analyse data source other than SAP

Table 11 Companies which have opted for implantation of BPM solutions. Source: own elaboration based on: [20]. [21] [22].

3. Business Process Models

After introducing business process management philosophy, it is moment to talk about the different bias and languages to translate those business processes into a formal and intuitive sequence of activities, such that all stakeholders can easily understand, in other words, create a business process model.

There are many different languages to translate business processes into models. In this chapter, the most commonly used ones will be mentioned as well as the modelling languages needed for good understanding of the process mining techniques which will be shown at a later stage.

3.1. Petri Nets

A Petri-net is a triplet $N = (P, T, F)$.

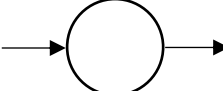


Name	Description	Graphical representation.
P	P is a finite set of places.	
T	T is a finite set of transitions.	
F	F is a set of directed arcs, called the flow relation.	

Table 12 Basic structures of Petri nets. Source: own elaboration.

As a result of the combination of the elements showed below, different types of splits and joins, such as AND split/join, XOR split/join...are configured and thus, the petri net is built.

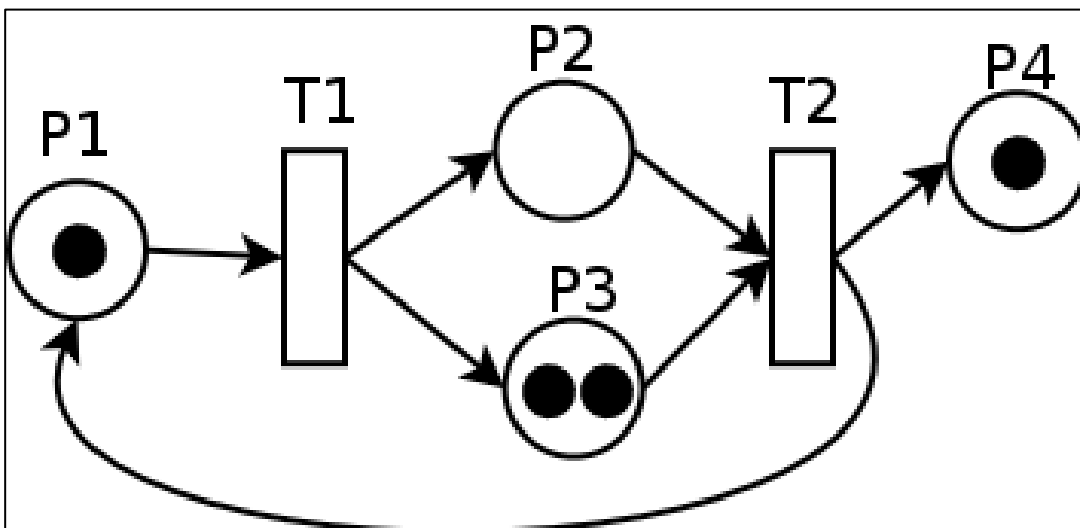


Figure 4. Petri Net. Source: [23]

Petri net is governed by firing rules, tokens, the black dots in figure above, can represent, people, items, documents, etc. The tokens can flow through the Petri net, and the state of a Petri net is determined by the distribution of their tokens. If a place contains at least one token that place is marked. In the figure above, P1, P3, and P4 are marked.

3.2. Workflow Nets

Workflow nets are a sub-class of Petri nets. The main difference is that Workflow nets are characterized by a place where the process starts and a place where the process ends.

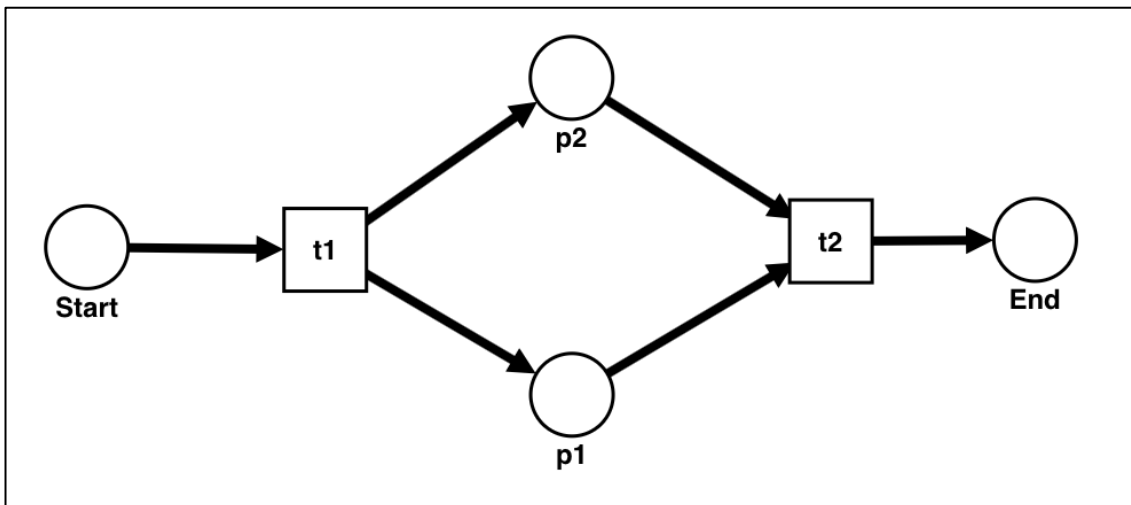


Figure 5. Workflow net. Source: own elaboration.

Workflow nets are particularly relevant in the context of BPM because it is easy to show the life-cycle of the instances in the process. Examples may be customer orders, job applications... here the process model is initiated once for each case.

3.3. Swim Lanes

Swim Lanes are not a modelling language itself, is rather a representative and useful notational addition, often incorporated into BPMN, EPC, UML, or simple flowcharting as a mean for defining the performer responsible for performing an activity.

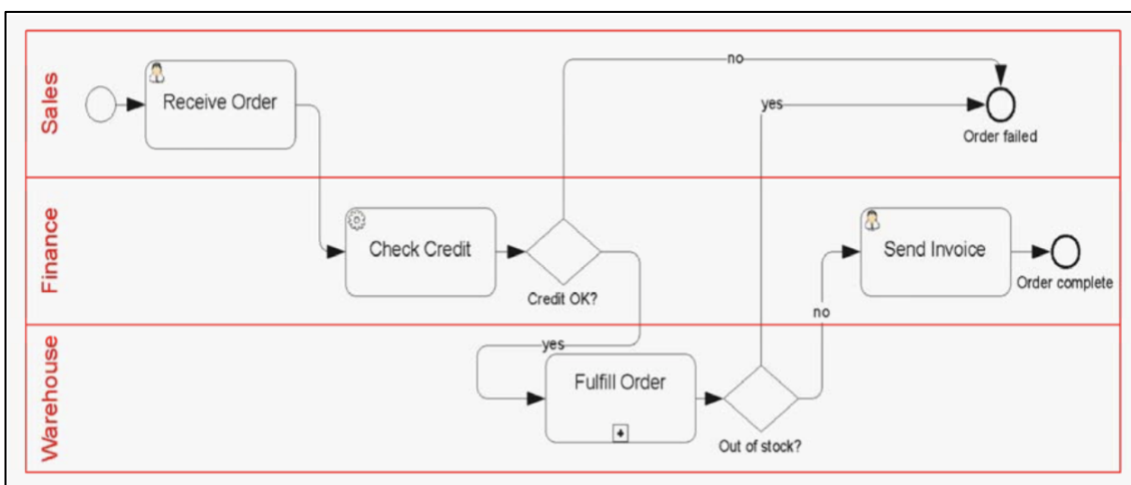


Figure 6. Swim Lanes express different responsible. Source: [24]

3.4. Unified Modeling Language UML

Unified Modeling Language is a standard visual language proposed for analysing, designing, and implementation of software-based systems. Some organization uses it for the design of business process, but it is a secondary use.

UML 2.5 (current version) is not complete and it is not completely visual. Given an UML diagram it is not possible to be sure to understand part of the process. Since some concepts in UML does not have a graphical representation at all.

There are two main big groups of diagrams:

- **Structure diagrams:** show static structure of the system and its parts on different abstraction and implementation levels and how those parts are related to each other. The elements in a structure diagram represent the meaningful concepts of a system, and may include abstract, real world and implementation concepts. [25]
- **Behaviour diagrams:** show the dynamic behavior of the objects in a system, which can be described as a series of changes to the system over time. [25]

Within those two big groups, there are many other types of diagram, each one for a different purpose. But it does not belong to this literature.

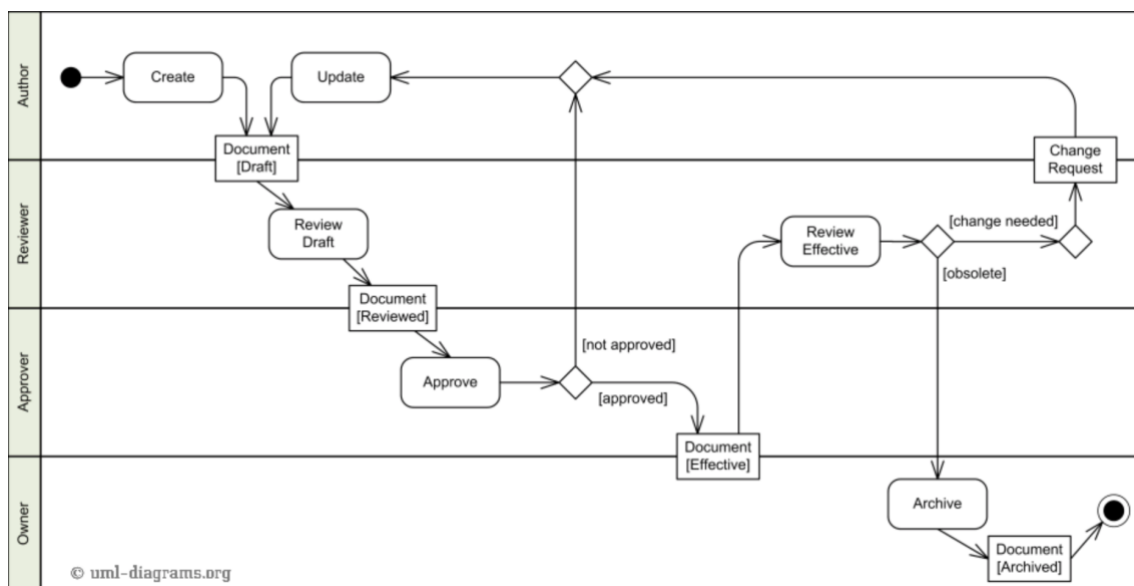


Figure 7. Example of Document management process using UML 2.5. Source: [26].

For further information:

- <http://www.uml.org>

3.5. IDEF

Integrated DEFinition is a family of modeling notation concepts. There are sixteen IDEF methodologies, nevertheless IDEF0, for modeling business functions or systems, and IDEF3, for modeling business processes are the most used in BPM context.

Developed by US Air Force, it was widely used and available in many modeling tools for many years and is now in the public domain.

The notation employs a very simple set of symbols consisting of process boxes with arrows showing inputs, outputs, controls, and mechanisms.

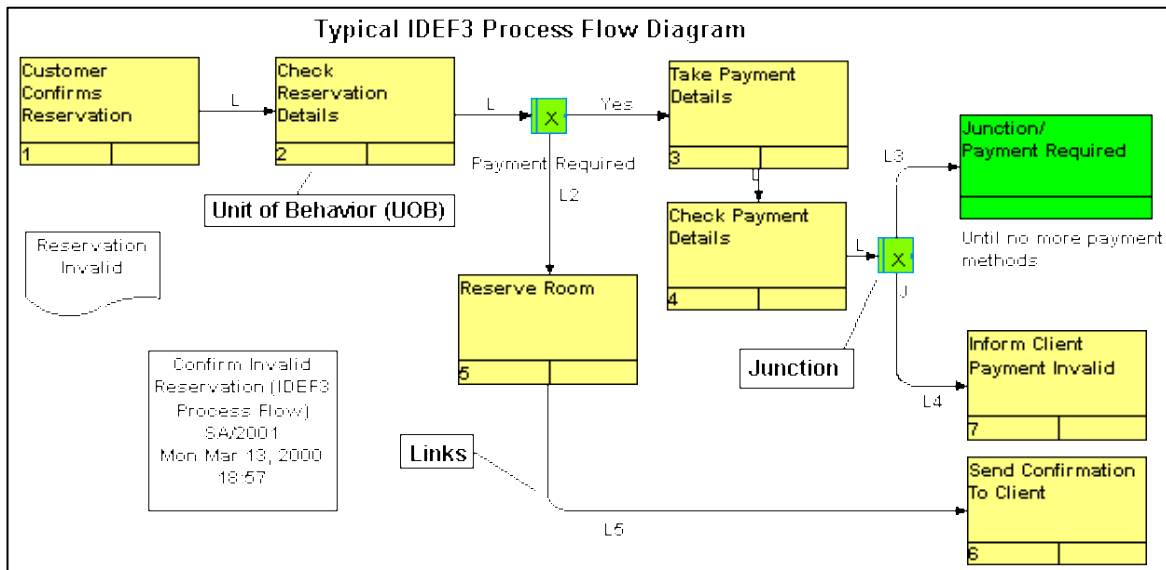


Figure 8 Example of IDEF3 process diagram. Source: [27]

For further information:


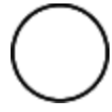


- Draft Federal Information Processing Standards Publication 183.

3.6. Business Process Model & Notation BPMN

Developed by Business Process Management Initiative (BPMI) and standardized by Object Management Group (OMG). Business Process Model and Notation is nowadays one of the most used languages to model business process. Since it is very intuitive and easy comprehensive for many audiences.

In BPMN 2.0 there are three basic types of notational elements:

- Task/Activities: task is the work unit, a work to do.
- Events: despite there are many types of event, a simple event express starting, ending, or middle happenings.
- Gateways: AND, OR, and XOR are the most used gateways, nonetheless, many other gateways based on different events are possible.

Element	Symbol
Task	
Event	  
	Star event Intermediate event End event




Gateway	 AND	 XOR	 OR
---------	--	--	---

Table 13. Main elements in a BPMN model. Source: own elaboration based on: [28]

The following picture, Figure 9, shows an illustrative example of BPMN process model.

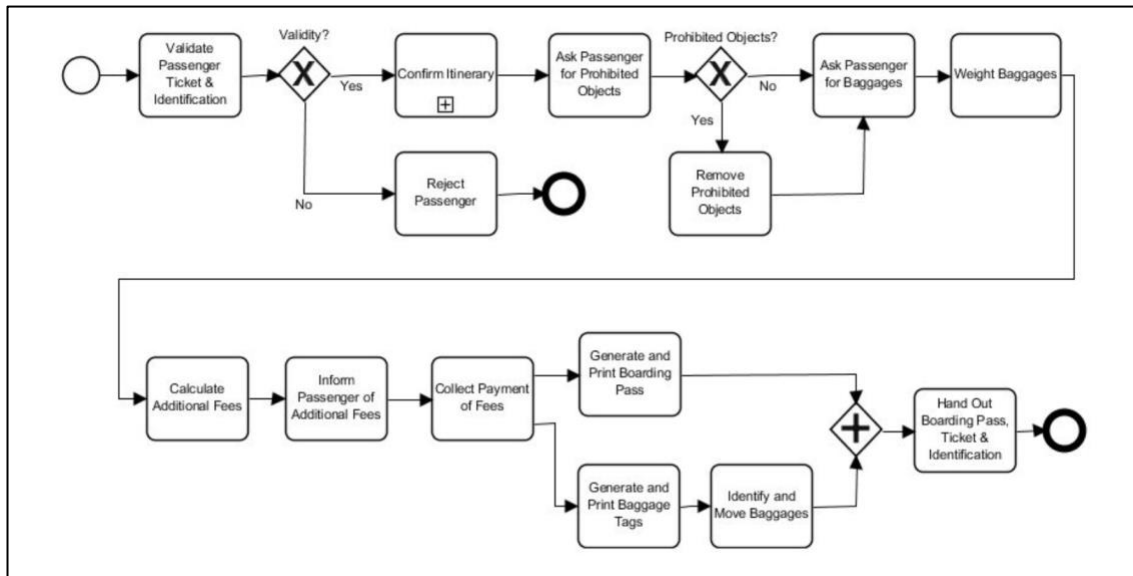


Figure 9 BPMN airline check-in, simple example. Source: [29]

Business Process Model and Notation is quite broader modelling language than here is explained, and contains many more elements.

For further information:

- <http://www.bpmn.org>
- <http://www.omg.org>
- http://www.bpmb.de/images/BPMN2_0_Poster_EN.pdf

3.7. Event Process Chain, EPC

Event Process Chain is a widely used in Germany and many other countries in Europe, especially in multinational enterprises. It ranges from a very simple to very complex models, nevertheless in this literature will be only mentioned the basics.

In an EPC diagram three main elements can be found, activities, events and logical operators.




Element	Description	Symbol
Activity	Describe an incidental task that typically consumes time and resources.	
Event	Describe an occurred condition that cause a sequence of activities.	
Logical Operator	in EPC AND, OR and XOR gateways are possible	 AND OR XOR

Table 14 Main elements in EPC. Source: own elaboration based in: [30].

Events can trigger functions, and functions are triggered by events. A simple model is built through a sequence of event-activity-event plus the logical operators.

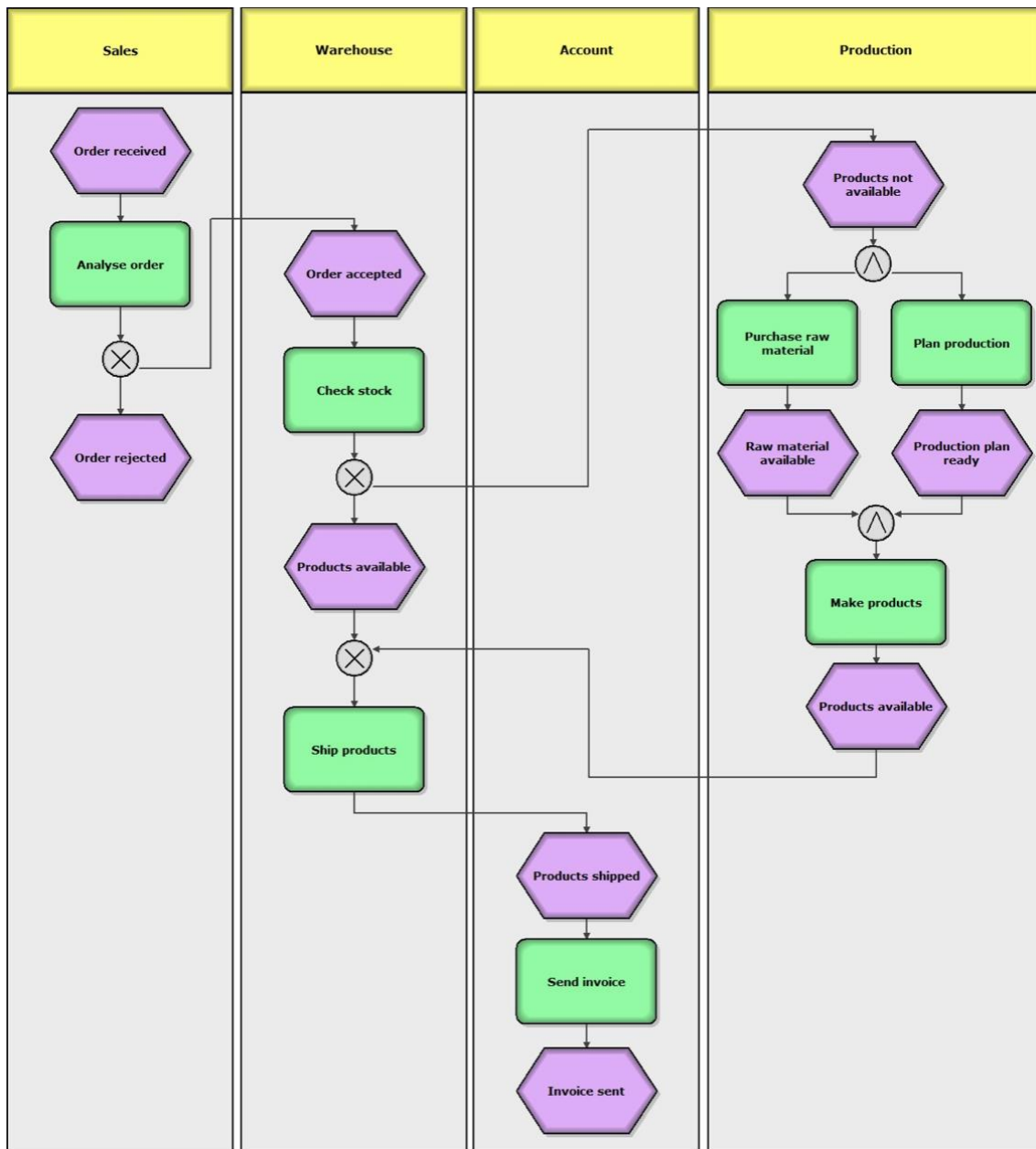


Figure 10. EPC example. Source: [31]

For further information:

- https://www.epc-standard.org/collaborate/Towards_EPC_standardization
- <http://www.ariscommunity.com/>

4. Process Mining

4.1. Why Process Mining is useful for BPM

In the current globalized world, day after day organizations are more and more focused in processes and how to manage them through the IT solutions. The IT solutions generate great amount of data about the processes continually. It seems obvious that

organizations want to take advantage of these data in order to find weak points and improve their own processes or generate advantage over the competence. This premise leads this literature to the next point, Process Mining.

The three types of process mining:

- Discovery: try to discover new business process through an event log
- Conformance: try to know whether the event log fits with the model proposed.
- Enhancement: propose and make improves in a business process model for a better explanation of the reality.

can be placed in the BPM lifecycle in the following way.

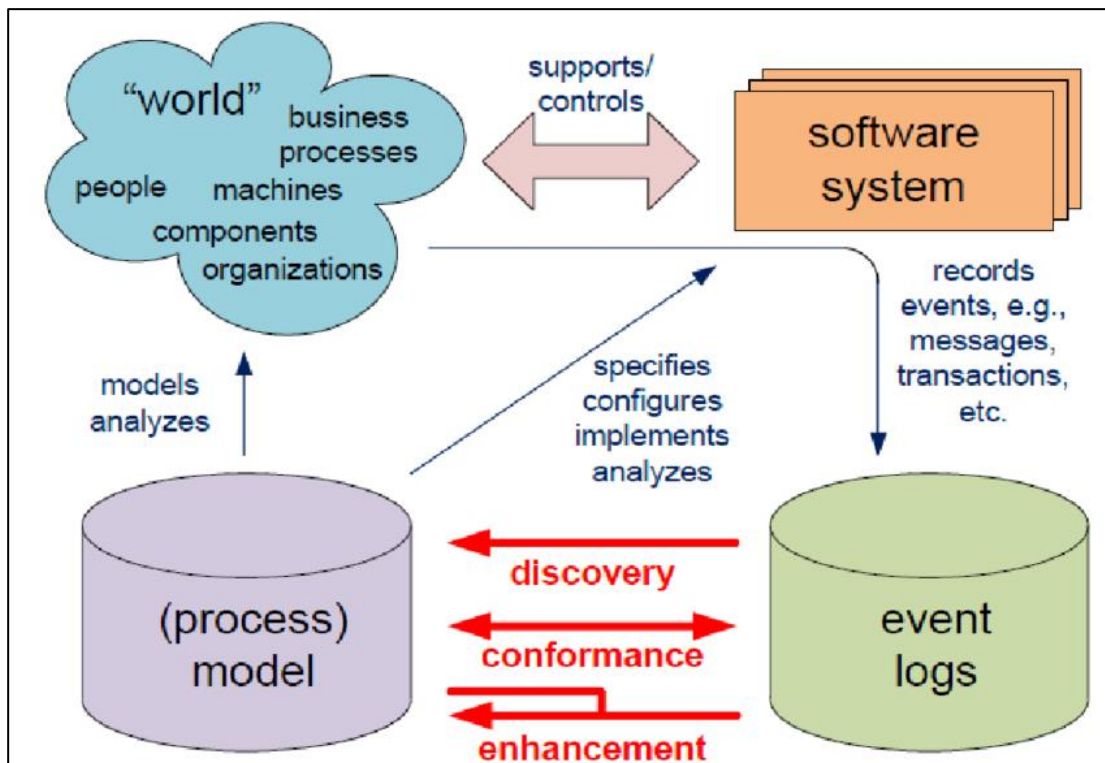


Figure 11. Discovery, conformance, and enhancement within the BPM lifecycle. Source: [32].

Process mining is becoming a vital support tool at the time to manage processes. Since now it makes very easy discover, check and improve business process.

Examples of how process mining may help organizations are, for example, the capacity of process mining for detecting bottlenecks and delays during the execution of the real processes.

On the other hand, process mining allows the application of the standards such as ISO 14000 (Environmental Management), ISO 9000 (Quality Management) and the analyse of the compliance, as well as a fast detection of deviations and misalignments in the real process.

4.2. What is process mining

Process mining is a set of techniques that allows to analyse the business process through an event log. The aim of process mining is to extract information from the process, stored in the log events, and turn it into knowledge. This knowledge implies to find the

route of the real process, including, for example, the resources that execute the process or the time that take to complete an activity.

It is important to distinguish the difference between process mining and process monitoring, since process mining is a discipline that tries to turn information, hosted in the data generated by the events, into useful knowledge in order to improve, discover or check the conformance of the business processes. Whereas process monitoring is a real time analysis of the performance of the different processes, generally manufacturing processes. Process monitoring is conducted using checklists and guidelines. Those checklists are developed jointly with project staff. The same checklists and guidelines are used by field staff while implementing project activities.

The main purpose of the Process Mining is fill the gap and act as the bridge between data science and process science [33],p 25. Therefore, is vital to know what data science is and what is process science.

4.3.Data Science and Process Science

Data science is a multidisciplinary discipline composed of many different sub-disciplines related with data analysis, and how to turn these data into value, in form of reports, diagnostics, predictions, and recommendations. The main disciplines that compose Data Science are presented on the table below.

Discipline	Description
Statistics	It is the origin of data science. It is typically split into descriptive and inferential statistics.
Algorithms	They are vital in any approach analysing data. The larger is a data set, the more important is the algorithm.
Data Mining	Is the analysis of data sets in order to find unsuspected relationships and to summarize the data in novel ways that are useful for the data owner.
Machine learning	Try to develop computer programs that automatically improve with the experience.
Process mining	It seeks the confrontation between event data and process model, for example, by discovering models from event data, or checking whether a model is aligning with what happen in the real-life process.
Predictive analytics	it consists in extracting information from data sets, in order to determine patterns and predict future outcomes and trends.
Databases	They are used to store data. Must serve to two purposes, structuring data easily and providing scalability and reliable performance.
Distributed systems	These provide the infrastructure to conduct analysis. A distributed system is composed of interacting components that coordinate their actions to achieve a common goal.

Visualization & visual analytics	The purpose of this discipline is to ease the interpretation of result after making an analyse.
Business models & marketing	Try to turn data into value, including business value.
Behavioural/ social science	Behavioural science is the analysis of social behaviour, while social science studies the social systems and the relationships among individuals within a society, both in order to influence people guiding the costumer to a product.
Privacy, security, laws and ethics	In order to ensure the confidentiality of the data and avoid the bad ethical conducts.

Table 15 Data science set of disciplines. Source: own elaboration based on [33],pp 12-17.

The term “process science” is used to group many disciplines which range from knowledge about information technology to knowledge about management science with the purpose of improving and running operational processes. Now it is showed the most important disciplines of this broad working field.

Discipline	Description
Stochastics	It provides a repertoire of techniques to analyse random processes. That is done in order to allow the model for analysis.
Optimization	Techniques which aim to provide the best alternative from al large or even infinite set of alternatives.
Operation management & research	It deals with the design, control and management of products, processes, services and supply chains.
Business process management	As it has been shown, it is the discipline that combines approaches for the design, execution, control, measurement, and optimization of business processes.
Process mining	Attending the quote in the introduction, it is also part from the process science. The focus is not in process modelling, but exploiting event data generated in the processes.
Business process improvement	It is an "umbrella term" for a set of different disciplines which aim at process improvement such as TQM, Six Sigma and Lean, BPR...
Process automation & workflow management	It focuses in the development of information system supporting operational business process from the routing to the distribution of work. By changing the configuration of the developed model, the real one is changed.
Formal methods & currency theory	based on theoretical computer science, in particular, logic calculi, formal languages automata theory, and program semantics.

Table 16 Process science set of disciplines. Source: own elaboration based on [33],pp 12-17.

In conclusion, data science (data-centred) disciplines tend to search for relatively simple patterns in large data sets, such as association rules, decision trees, cluster, etc. However, these patterns do not describe the end-to-end process.

On the other hand, process science approaches focus on end-to-end process but do not count with the event data. Process mining provides a bridge between both, data science and process science management allowing go one step further in the analysis of business process.

4.4. History of PM

The data collection is not an innovative activity. Therefore, human beings have been collecting data in order to turn it into value information and knowledge since they acquired the ability to write. In the ancient Babylonia 3000 B.C. They used mud tables where it was collected the data about the farming production or the loans and exchanges they carried out, with the purpose of knowing whether their farms were getting better, or someone debt too much letters of a credit. Another example are Egyptians, they had a very good structured data sets of population and income. [34] Statistics, what can be considered as the mother of data mining, is not either a new invent surged from the new technologies. The history of the statistic can be resumed in five big stages as it is showed in the following table.

Stage	Age	Description
1. Census	3000 (B.C.)- 15th century	It is the first stage of statistics, this is not the statistic as people know nowadays, but at the moment that a politic authority is built, the efforts for knowing the amount of people and the wealth of the country begin, and start the first administrative works.
2. Scientific method	15th C. - 17th C.	During these centuries figures such as Leonardo de Vinci, Nicolas Copernicus, Galileo, Neper, William Harvey, Sir Francis Bacon and René Descartes made important contributions to the scientific method, so when the international commerce appeared, there was already an applicable method for dealing with economic data.
3. Theory of probabilities	17th C. - 18th	Mathematics as Bernoulli, Frances Maseres, Lagrange and Laplace developed the theory of probabilities, but it was applied only on the games of chance. Until the 19th century it was not applied for solving Scientifics problems.

4.Theory of observation errors, least squares, and correlation.	19th C.	Between 1800 - 1820, Laplace and Gauss developed the theory of the observation errors and with the help of Legendre, they also developed the method of least square. In the last part of 19th century, Sir Francis Gaston introduced the method known as Correlation, in order to measure the relative influence of the factors over the variables.
5. Theoretical statistics	20th C.	Statistic is considered as an independent science and becomes what people know currently as statistics, appeared the Bayesian methods, multivariant analysis methods, The Theory of Games, and the probabilistic models.

Table 17 History of Statistic. Source: own elaboration based on: [35], [36].

With the emergence of computers, and afterwards, the internet, the amount of data is getting bigger following an exponential function, is at this point, when emerge the need of new techniques and methods for analysing and getting useful information from the data. The data mining is not a new term, since the 60's there were terms used by statisticians such as data fishing, data archaeology or even data mining with the purpose of finding correlations within the data without a previous hypothesis.

It is during the 80's when the term Data Mining acquire its importance as a part of a bigger process called KDD, Knowledge Discovery in Databases (1889). Data mining is an opportunity for the organizations for dealing with massive volume of data, and analyse it in order to find useful information for the marketing projects. From 2000 on, the internet became a tool used for most of the people, so data mining techniques and methods evolved for trying to cover all the types of data, therefore, appearing techniques as OLAP (On-Line Analytic Process) or EDA (Exploratory Data Analysis) which surround multivariate exploratory techniques and neural networks.

In 2010, data mining is just a part of a larger set of disciplines known as data science, and it is at this point, when the analysis of data is not only important down the point of view of finding trends and trying to influence consumer to buy a product or service, but the organizations, want to know more about their processes, whether they are working in the correct way, uncover new processes from the data sets, or improve their own processes. Data mining splits in a new discipline called Process Mining, which emerge in order to solve all these points.

The term process mining is a relatively new. It was introduced in detail in the "*Process Mining Manifesto*" [37], published in 2012(final version), written by 77 researchers and experts in process mining techniques all around the world known as IEEE *Task Force on Process Mining*. This manifesto was written in order to spread the topic of process mining, increasing the maturity of the process mining as a new tool for re-designing, control, and support the business process models. Since 2012 on, many IT companies have developed their own process mining tools.

Name	Company
ProM	Technique University of Eindhoven (Open-source software)
Disco	Fluxicon
Interstage Automated Process Discovery	Fujitsu
BPMone	Pallas Athena
ARIS Process Performance Manager	Software AG
QPR Process Analyzer	QPR
Celonis Proactive Insights	Celonis

Table 18 Some companies that offer their own Process Mining Suites. Source: own elaboration.

Nowadays process mining allows companies to take advance of that in order to improve and support their business processes in a quick changing and competitive environments.

4.5. Process Mining concepts

Process mining is an emerging research discipline sat between data mining on the one hand, and process modelling and analysis on the other. Process mining provides a set of computational techniques and software, which allow organizations to extract information and create knowledge from it, by analysing End-to-End business process, more specifically, the event data generated by them. Event logs are the starting point of process mining techniques, they are stored in some Information System, for example an ERP such as SAP. The next figure illustrates an example of event log.

	A	B	C	D	E	F
1	Case ID	Start Timestamp	Complete Timestamp	Activity	Resource	Role
2	1	2011/01/01 00:00	2011/01/01 00:37	Create Purchase Requisition	Kim Passa	Requester
3	2	2011/01/01 00:16	2011/01/01 00:29	Create Purchase Requisition	Immanuel Karagianni	Requester
4	3	2011/01/01 02:23	2011/01/01 03:03	Create Purchase Requisition	Kim Passa	Requester
5	1	2011/01/01 05:37	2011/01/01 05:45	Create Request for Quotation Requester	Kim Passa	Requester
6	1	2011/01/01 06:41	2011/01/01 06:55	Analyze Request for Quotation	Karel de Groot	Purchasing Agent
7	2	2011/01/01 08:16	2011/01/01 08:26	Create Request for Quotation Requester	Alberto Duport	Requester
8	4	2011/01/01 08:39	2011/01/01 09:00	Create Purchase Requisition	Fjodor Kowalski	Requester
9	2	2011/01/01 09:34	2011/01/01 09:38	Analyze Request for Quotation	Karel de Groot	Purchasing Agent
10	5	2011/01/01 09:49	2011/01/01 10:35	Create Purchase Requisition	Esmana Liubiata	Requester
11	2	2011/01/01 10:16	2011/01/01 10:21	Amend Request for Quotation Requester	Christian Francois	Requester Manager
12	2	2011/01/01 11:15	2011/01/01 11:48	Analyze Request for Quotation	Magdalena Predutta	Purchasing Agent
13	6	2011/01/01 11:20	2011/01/01 11:37	Create Purchase Requisition	Christian Francois	Requester
14	1	2011/01/01 11:43	2011/01/01 12:09	Send Request for Quotation to Supplier	Karel de Groot	Purchasing Agent
15	1	2011/01/01 12:32	2011/01/01 16:03	Create Quotation comparison Map	Magdalena Predutta	Purchasing Agent
16	2	2011/01/01 12:33	2011/01/01 12:39	Amend Request for Quotation Requester	Esmana Liubiata	Requester Manager
17	2	2011/01/01 13:28	2011/01/01 13:38	Analyze Request for Quotation	Karel de Groot	Purchasing Agent
18	7	2011/01/01 14:05	2011/01/01 15:00	Create Purchase Requisition	Esmana Liubiata	Requester
19	8	2011/01/01 14:27	2011/01/01 15:17	Create Purchase Requisition	Fjodor Kowalski	Requester
20	2	2011/01/01 15:18	2011/01/01 15:40	Send Request for Quotation to Supplier	Francois de Perrier	Purchasing Agent
21	2	2011/01/01 15:55	2011/01/01 16:43	Create Quotation comparison Map	Karel de Groot	Purchasing Agent
22	9	2011/01/01 16:17	2011/01/01 16:34	Create Purchase Requisition	Tesca Lobes	Requester
23	6	2011/01/01 17:32	2011/01/01 17:45	Create Request for Quotation Requester	Alberto Duport	Requester
24	8	2011/01/01 18:00	2011/01/01 18:07	Create Request for Quotation Requester	Tesca Lobes	Requester
25	6	2011/01/01 18:39	2011/01/01 18:55	Analyze Request for Quotation	Magdalena Predutta	Purchasing Agent

Figure 12. Example of event log. Source: [38].

Every row in the Figure 12, corresponds to an event, the set of events that belong to the same case ID, conform a trace, in Figure 12, it is possible to observe, for example the trace for the case2 σ_2 (Create purchase requisition, create request for quotation requester, analyse quotation...). An event has different properties (columns). There are three generic properties that must appear in every event log for being able to extract useful information from it:

Property	Description	Why is vital	Figure 3
Case ID	A case identifier, is necessary to distinguish different executions of the same process. What precisely the case ID is depends on the domain of the process.	Case ID is necessary to distinguish the different executions of the same process.	The case ID represents a purchase order.
Activity	Represent process steps or status changes that were performed in the process.	Data must be detailed enough. Therefore, to guarantee that one can access to the whole history of a case, it is vital to add some identifier that express what steps was executed in each case.	The activity name refers to a Purchase activity
Timestamp	Timestamp is a time indicator that points the moment at each activity started and ended.	At least one timestamp is needed, in order to bring the events in the right order. Timestamps are needed to identify delays between activities and bottlenecks.	There are two timestamps, one for the start of the activity, and one that indicates the end of that activity.

Table 19 The three vital properties in all event log. Source: own elaboration.

There can be additional data if it is considered necessary. For example, in Figure 12 has been added two columns, one showing the resource, who or what is implementing the activity, and the other one showing the role. In the example the resources are people, therefore, may be useful that the role of the people who is doing the activities appears. On the other hand, there are process models. A process model, as mentioned in chapter 2 is a virtual representation of the logical sequence of a process, that is happening or bound to happen in the reality, it can be represented in many ways, but in process mining are often used BPMN and Petri nets. An example of process model is again shown in the following figure.

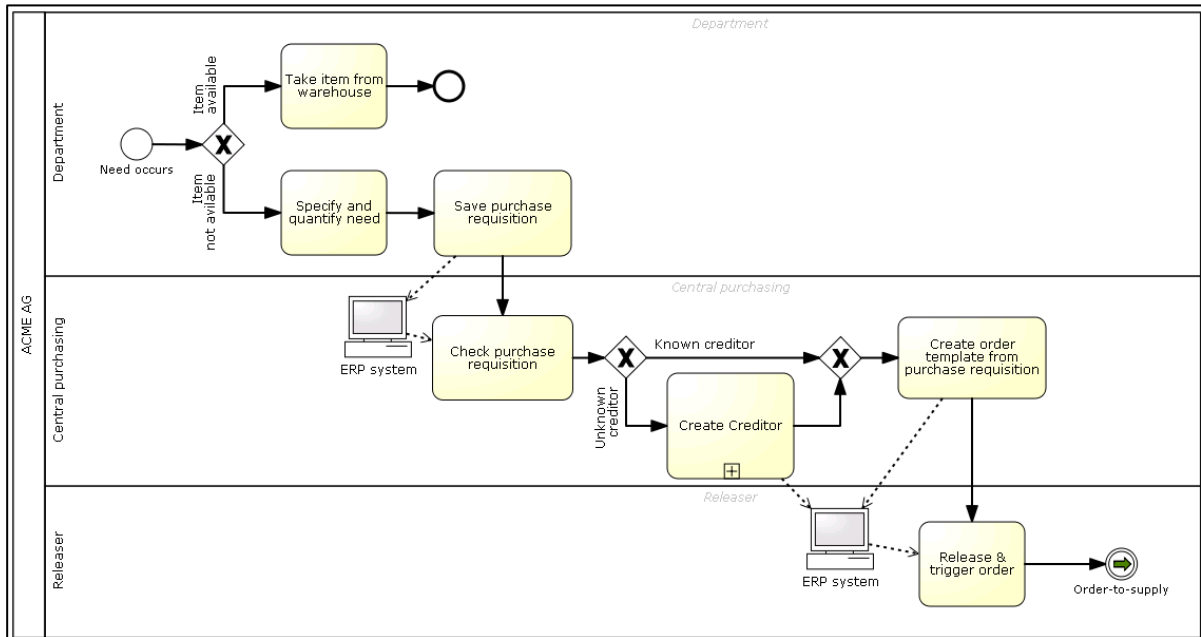


Figure 13. Example of business process model represented via BPMN2.0. Source: [39].

The business process in Figure 13 consists in the creation of an order for an item, it follows, in a logical sequence, how the activities are being executed, and how the process go through the different departments.

Process models should be balanced between four quality criteria: fitness, simplicity, precision, and generalization.

Property	Objective
Fitness	The ability to explain the observed behaviour.
Simplicity	Occam's razor, the model should be as simple as possible.
Precision	Avoiding underfitting. If the model is too general it can explain the observed behaviour, but it will not be possible extract useful information.
Generalization	Avoiding overfitting. If the model is too specific, may not fit with future events.

Table 20 Four properties of process models. Source: own elaboration.

One of the main points of process mining is to relate the event log, which contains the information about what happens in the reality, with the model, used to explain the behaviour of the process. The terms Play-in, Play-out, and Replay are used for representing this relationship.

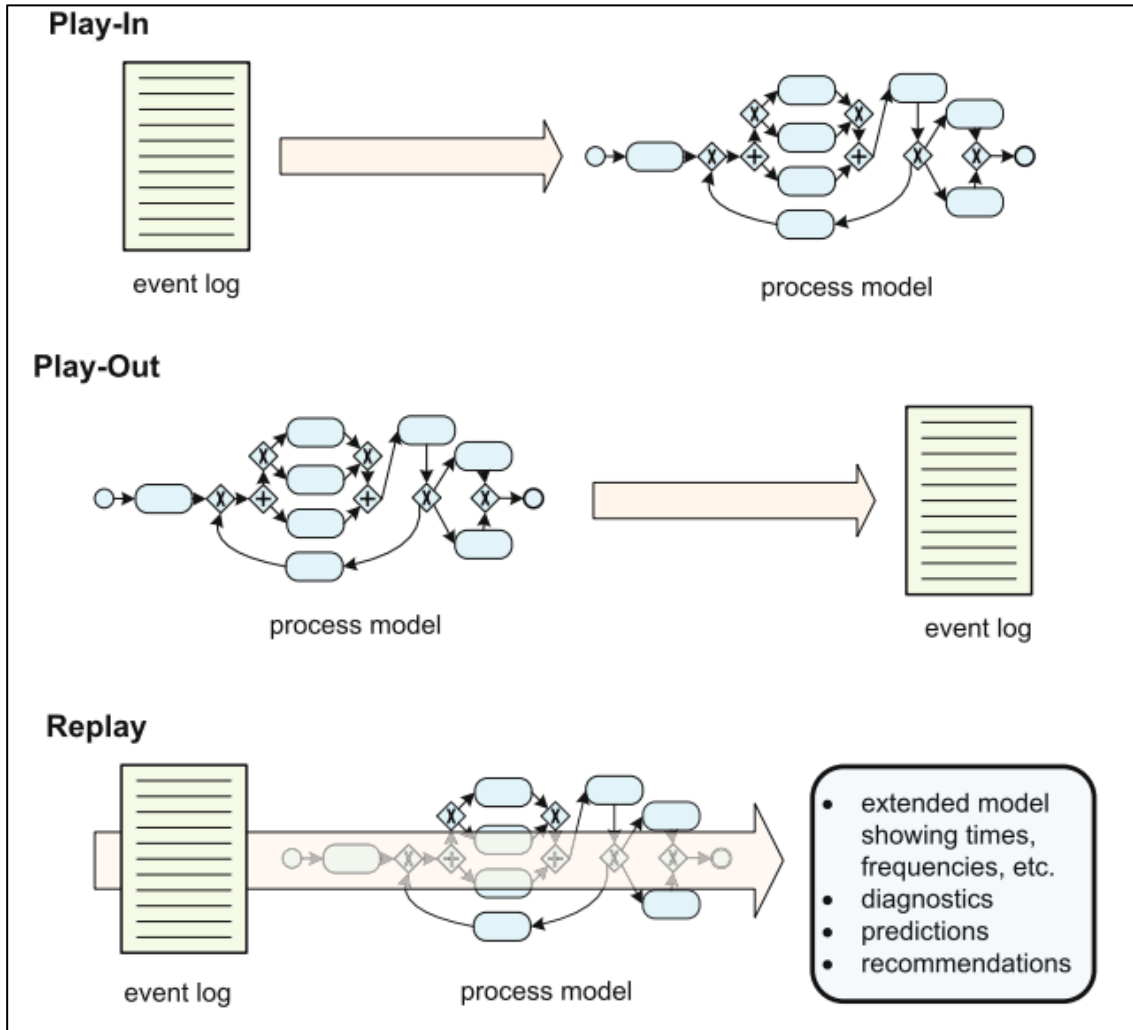


Figure 14. Three ways to relate the model with the reality. Source: [40]

Relationship	Input	Output	Description
Play-in	Event log	Process model	A process model is built in base on an event log. It is very useful for discovering process mining.
Play-out	Process model	Event log	It is referred to the classical use of a model. Having a model, it is used for simulating new behaviours and testing operational processes to find weak points all from a theoretical point of view.

Replay	Both process model and event log	Extended model, reports, diagnostics...	The event log is replayed and contrasted with the process model. The objectives can be: to extend the model adding frequencies in order to detect bottlenecks, to check the conformance between the model and the reality, and predict new behaviours between others.
--------	----------------------------------	---	---

Table 21 Play-in, Play-out, and Replay explanation. Source: own elaboration.

It is important to notice that relationships between the models and the logs are directly linked to the three different types of process mining, since depending on the type of process mining it is being used, a different relation will be used as well. However, play-in, play-out, and replay just refer the way in which a model or event log can relate each other. Discovery, conformance, and enhancement techniques, are which provide the means for making it possible.

4.6.Types of Process Mining

Process mining is used in order to support the take of decisions, implement improvements in the business processes or check whether the process is working or not, on the way it is supposed to do. There are three different types of process mining.

Type	Purpose
Discovery	Producing models from an event log, without using any other a priori information. There are many techniques to generate process model from raw event data, for example α algorithm, which is able to discover a Petri net by identifying basic process patterns in the event log. Process discovery is often used as starting point for other types of analysis.
Conformance	Comparing an existing process model with an event log of the same process. The comparison shows where the real process deviates from the modelled process. Conformance checking can be used to check whether reality, as recorded in the log, conforms to the model and vice versa.
Enhancement	Taking an event log and process model and extend or improve the model using the observed events. This third type of process mining aims at changing or extending the a priori model.

Table 22 Types Of process mining. Source: own elaboration based on [41].

4.7.Discovery techniques, α -algorithm

Once the Event log is obtained, discovery techniques try to discover a process, in the case of α -algorithm, represented via Petri net, which fits with the behaviour observed on the event log. These techniques are play-in type and the most famous of them is the α -algorithm.

4.7.1. α -algorithm

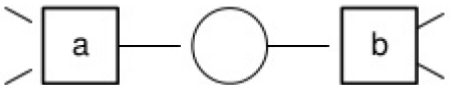
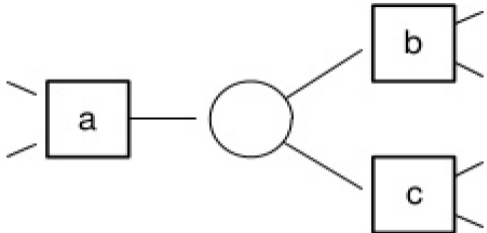
This algorithm, because of its simplicity, is a good point for introducing the topic of process discovery, however, this algorithm should not be seen as a very practical technique as an isolated algorithm, that is because it has some problems at time to deal with noise, short loops, and complex routes or event logs. Nevertheless, many of its ideas have been embeddedness in more complex and robust techniques.

The α -algorithm scans the event log looking for some particular patterns. There are four log-based ordering relations that aim to capture important patterns in the log.

Pattern	Characterisation	Description
Direct succession	$x > y$	if xy is observed in log traces "...xy...".
Casualty	$x \rightarrow y$	If in the traces "...xy..." is observed but not "...yx...".
Parallel	$x y$	If both, "...xy..." and "...yx..." are observed.
Unrelated	$x \# y$	if neither "...xy..." nor "...yx..." are observed.

Table 23 The four patterns in α -algorithm. Source: own elaboration.

The typical pattern representations that can be found are the following:

Type	Pattern	Figure
Sequence	$a \rightarrow b$	
XOR-split	$a \rightarrow b, a \rightarrow c, b \# c$	

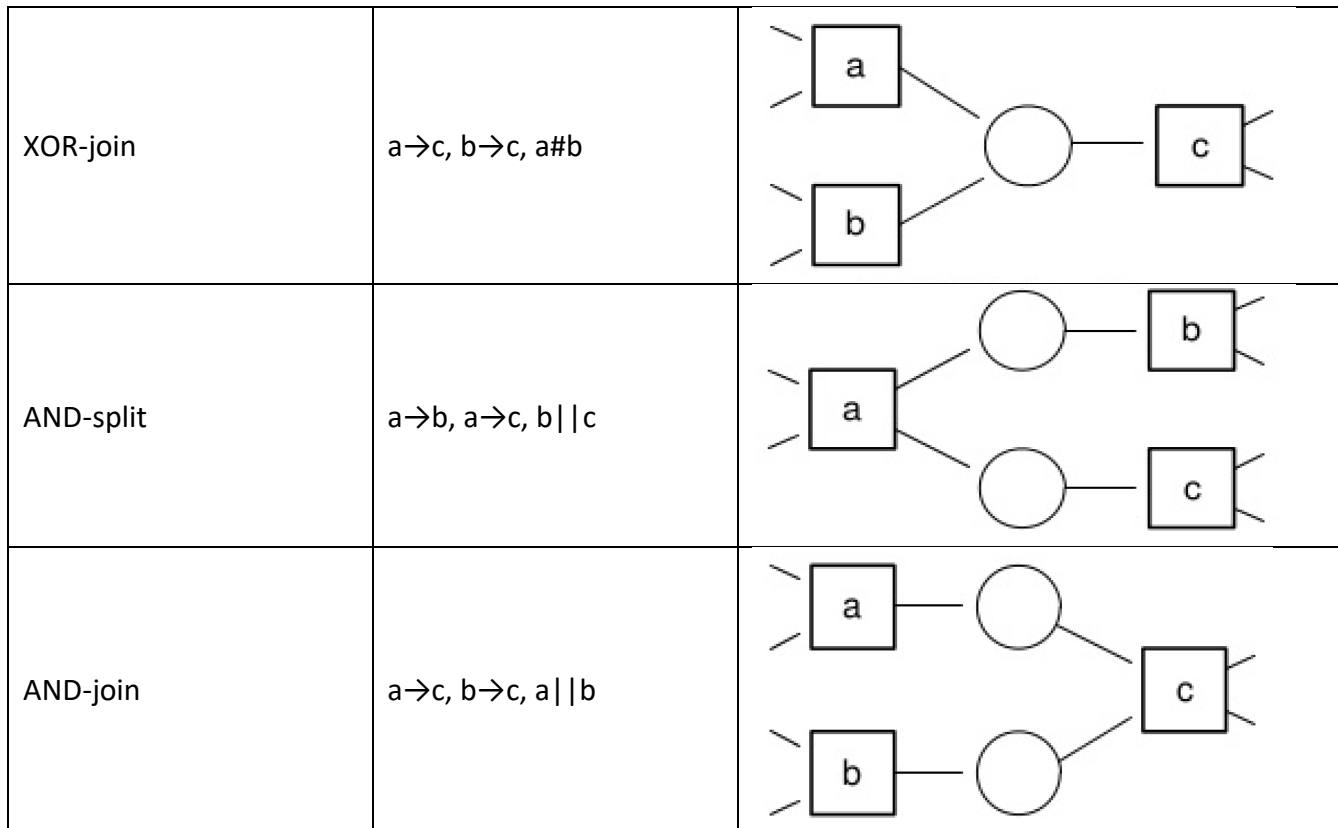


Figure 15. Basic process patterns. Source: [33], p 169.

There are two assumptions that should be consider.

- Perfect information, completeness.
- Absence of noise in the log.

In order to a better understanding it will be showed one simple example step by step. Considering the event log $L_1 = [abcd, acbd, aed]$, the log-based ordering relations are:

$> L1$	$[(a,b),(a,c),(a,e),(b,c),(b,d),(c,d),(c,b),(e,d)]$
$\rightarrow L1$	$[(a,b),(a,c),(a,e),(b,d),(c,d),(e,d)]$
$\parallel L1$	$[(b,c),(c,b)]$
$\# L1$	$[(a,a),(a,d),(b,b),(b,e),(c,c),(c,e),(d,a),(d,d),(e,b),(e,c),(e,e)]$

Table 24 Log based ordering relations. Source: own elaboration

Now the footprint can be built. The footprint is a matrix that shows the type of relation between the activities that compose the traces.

	a	b	c	d	e
a	#	\rightarrow	\rightarrow	#	\rightarrow
b	\leftarrow	#	\parallel	\rightarrow	#
c	\leftarrow	\parallel	#	\rightarrow	#
d	#	\leftarrow	\leftarrow	#	\leftarrow
e	\leftarrow	#	#	\rightarrow	#

Figure 16. Footprint. Source: own elaboration.

By analysing the footprint, a rough sketch is drawn following the next rules.

T_i , initial transition.	In the table, there is no incoming edges (\rightarrow) in the column if its transition. $T_i = a$
T_o , final transition.	in the table, there is no outgoing edges (\leftarrow) in the column if its transition. $T_o = d$
T_n , rest of transitions.	

Table 25 Initial transition T_i , final transition T_o , and T_n transitions. Source: own elaboration

In the table, by using the symbols \rightarrow , and $||$ the following graph can be constructed.

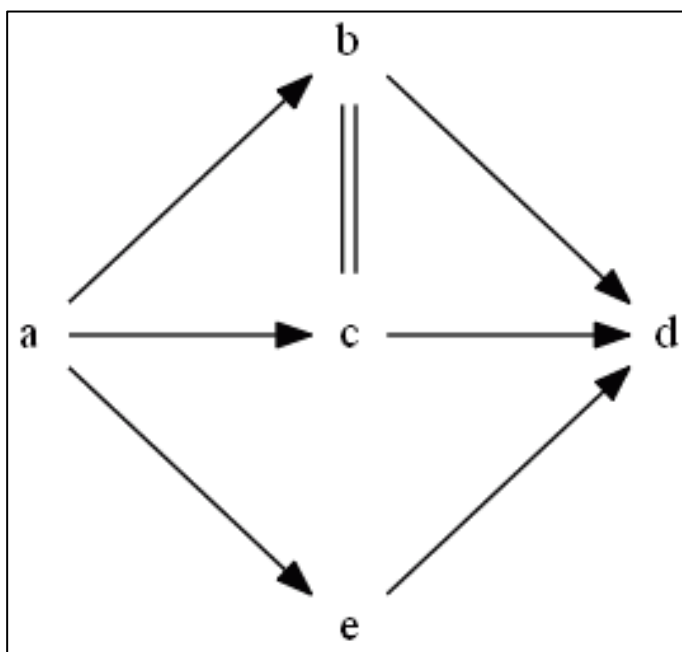


Figure 17. Rough sketch. Source: [42].

Now with this graph the following table is built. In it, it is represented the maximal set A, B. It means, the different places it must be added to the Petri net, as well as the pattern that represent each pair $\{A, B\}$. In order to configure the table, it is important to follow the next rules:

- $\forall a1, a2 \in A: a1 \neq a2$
- $\forall b1, b2 \in B: b1 \neq b2$
- $\forall a1 \in A, \forall b1 \in B: a1 \rightarrow b1$
- If there are parallel activities, these cannot belong to the same set.

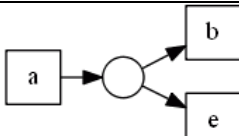
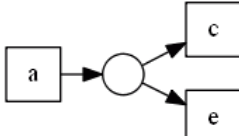
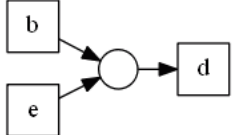
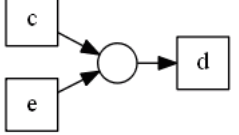
	A	B	
p1	{a}	{b,e}	
p2	{a}	{c,e}	
p3	{b,e}	{d}	
p4	{c,e}	{d}	

Table 26 Patterns for the Petri net. Source: [42].

The result of fix the different patterns is the following Petri net.

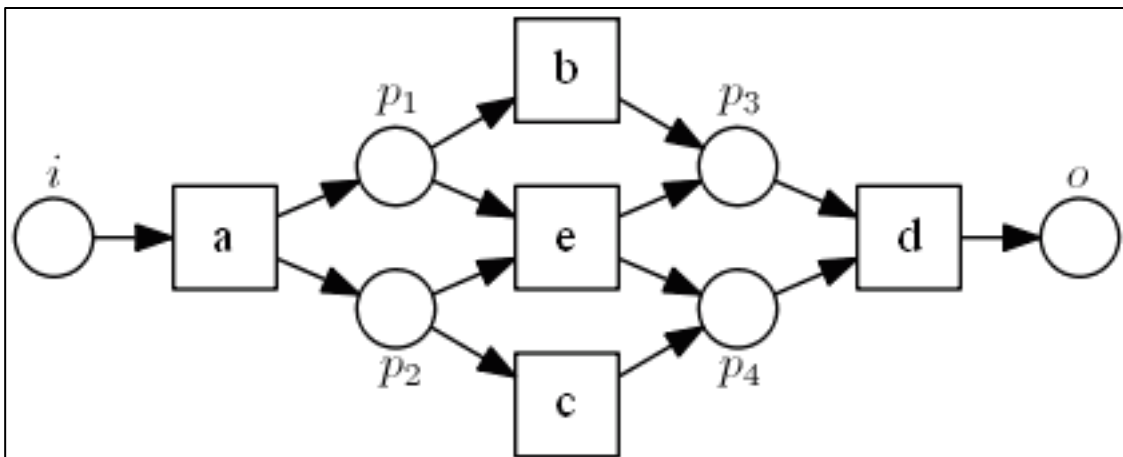


Figure 18. Final Petri net. Source: [42]

After showing a basic idea of how α -algorithm works, the whole α -algorithm is described. L is an event log over some set T of activities.

1. $T_L = \{t \in T \mid \exists \sigma \in L \ t \in \sigma\}$,
2. $T_I = \{t \in T \mid \exists \sigma \in L \ t = \text{first}(\sigma)\}$,
3. $T_O = \{t \in T \mid \exists \sigma \in L \ t = \text{last}(\sigma)\}$,
4. $X_L = \{(A, B) \mid A \subseteq T_L \wedge A \neq \emptyset \wedge B \subseteq T_L \wedge B \neq \emptyset \wedge \forall a \in A \ \forall b \in B \ a \rightarrow_L b \wedge \forall a_1, a_2 \in A \ a_1 \#_L a_2 \wedge \forall b_1, b_2 \in B \ b_1 \#_L b_2\}$,
5. $Y_L = \{(A, B) \in X_L \mid \forall (A', B') \in X_L \ A \subseteq A' \wedge B \subseteq B' \Rightarrow (A, B) = (A', B')\}$,
6. $P_L = \{p_{(A, B)} \mid (A, B) \in Y_L\} \cup \{i_L, o_L\}$,
7. $F_L = \{(a, p_{(A, B)}) \mid (A, B) \in Y_L \wedge a \in A\} \cup \{(p_{(A, B)}, b) \mid (A, B) \in Y_L \wedge b \in B\} \cup \{(i_L, t) \mid t \in T_I\} \cup \{(t, o_L) \mid t \in T_O\}$,
8. $\alpha(L) = (P_L, T_L, F_L)$.

Figure 19. α -algorithm. Source: [33], p 171.

Step	Explanation	Example: $L_1 = [abcd, acbd, aed]$
1.	It is checked which activities appear in the log T_L . These will correspond to the transitions of the generated Petri-net.	$T_L = \{a, b, c, d, e\}$
2.	T_I is the set of star activities.	$T_I = \{a\}$
3.	T_O is the set of end activities.	$T_O = \{d\}$
<p>Steps 4 and 5 form the core of the α-algorithm. It is aimed at constructing places named $p_{(A, B)}$, such that A is a set of input transitions ($\bullet p_{(A, B)} = A$) and B is the set of output transitions ($p_{(A, B)} \bullet = B$) of $p_{(A, B)}$.</p> <p>All elements of A should have causal dependencies with all elements of B, i.e., for all $(a, b) \in A \times B$: $a \rightarrow_L b$. Moreover, the elements of A should never follow one another, i.e., for all $a_1, a_2 \in A$: $a_1 \#_L a_2$. Similar requirements for B.</p>		
4.	X_L is the set of all such pairs that meet the requirements just mentioned above.	$X_L = \{(\{a\}, \{b\}), (\{a\}, \{c\}), (\{a\}, \{e\}),$ $(\{a\}, \{b, e\}), (\{a\}, \{c, e\}),$ $(\{b\}, \{d\}), (\{e\}, \{d\}), (\{c\}, \{d\}), (\{b, e\}, \{d\}), (\{c, e\}, \{d\})\}$
5.	Y_L contains all maximal pairs, non-maximal pairs are removed, otherwise there would be too many places.	$Y_L = \{(\{a\}, \{b, e\}), (\{a\}, \{c, e\}), (\{b, e\}, \{d\}),$ $(\{c, e\}, \{d\})\}$
6.	P_L contains every element $(A, B) \in Y_L$, these correspond to a place $p_{(A, B)}$ connecting transitions A to transitions B . P_L also contains a unique source place i_L and unique sink place o_L .	$P_L = \{p_{(\{a\}, \{b, e\})}, p_{(\{a\}, \{c, e\})}, p_{(\{b, e\}, \{d\})}, p_{(\{c, e\}, \{d\})},$ $i_L, o_L\}$
7.	Here, the arcs of the WF-net are generated. All start transitions in T_I have i_L as an input place and all end transitions T_O have o_L as output place. All places $p_{(A, B)}$ have A as input nodes and B as output nodes.	$F_L = \{(a, p_{(\{a\}, \{b, e\})}), (p_{(\{a\}, \{b, e\})}, b), (p_{(\{a\}, \{b, e\})}, e),$ $(a, p_{(\{a\}, \{c, e\})}), (p_{(\{a\}, \{c, e\})}, c), (p_{(\{a\}, \{c, e\})}, e), (b,$ $p_{(\{b, e\}, \{d\})}), (e, p_{(\{b, e\}, \{d\})}), (c, p_{(\{c, e\}, \{d\})}), (e,$ $p_{(\{c, e\}, \{d\})}), (p_{(\{b, e\}, \{d\})}, d), (p_{(\{c, e\}, \{d\})}, d), (i_L, a),$ $(d, o_L)\}$

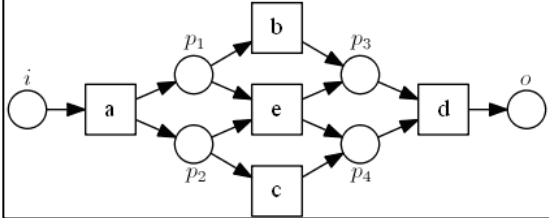
8.	The result is a Petri net $\mathbb{P}_L = (P_L, T_L, F_L)$ that describes the behaviour seen in event log L .	
----	---	--

Table 27 α -algorithm step by step and example. Source: own elaboration based on [33], pp 171-174.

The α -algorithm has some limitations that should be mentioned.

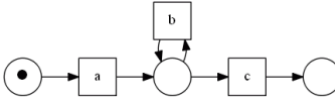
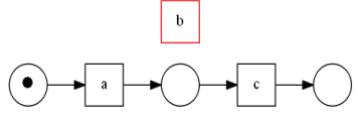
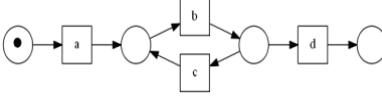
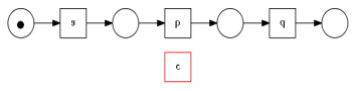
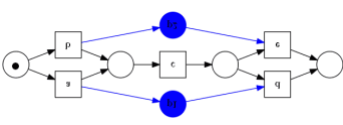
Limitation	Example log	α -algorithm should do	α -algorithm does
Loops of length 1.	[ac,abc,abbc,abbbc]		
Loops of length 2.	[abd,abcdbd,abcbcbd]		
Non-local dependencies.	[acd,bce]		

Table 28 Limitations of α -algorithm, examples. Source own elaboration based on [42].

α -algorithm was developed by Will van der Aalst et al in 2006, from then on, many variations has been developed, however, α -algorithm must not be seen as a practical algorithm, since it was used to explore the theoretical limits of process discovery [33], p 239..

4.7.2. Advanced techniques

α^+ -algorithm

α^+ -algorithm is the next step of the α -algorithm, and it can deal with short loops (lengths 1 and 2), by adding two new log-based ordering relations.

- $a \triangle b \Leftrightarrow a \triangle b \Leftrightarrow$ there is a subsequence ...aba.....aba... in the logs.
- $a \diamond b \Leftrightarrow a \diamond b \Leftrightarrow$ there are sequences ...aba.....aba... and ...bab.....bab...

and redefining the relations which caused the error.

- $a \rightarrow b \Leftrightarrow a > b \wedge (b \not> a \vee a \diamond b)$. This way we can correctly identify the follow relation when there's a loop of length 2.
- $a || b \Leftrightarrow a > b \wedge b > a \wedge a \not\diamond b$. By adding the last condition way, we don't misidentify the *parallel* relation.

It solves two of the three main problems of the α -algorithm. About non-local dependencies, it is not an isolated problem of the α -algorithm, since is a problem shared by many of discovery mining techniques nowadays.

There are many discovery mining techniques, but for practical use, it is vital that noise and incompleteness are well-handled, and few discovery algorithms can do that. That is why, for this literature, just three has been chosen. Each algorithm will be described briefly in order to have an idea of how it works.

Heuristic mining

Heuristic mining algorithms are characterized by taking frequencies of events and sequences into account when constructing a process model.

A good example of heuristic mining approach is:

1. A rough model of the process is created, for example, using the α -algorithm.
2. The model is refined first, by analysing the frequency of each trace that appears in the log, and after setting the minimum number of times that a trace must appear in the log to be considered determinant, the traces that do not appear enough are omitted, and if it does not compromise the rest of traces, that path from the model is erased.
3. The dependency between activities are quantified,

$$|a \rightarrow_L b| = \begin{cases} \frac{|a >_L b| - |b >_L a|}{|a >_L b| + |b >_L a| + 1} & \text{if } a \neq b \\ \frac{|a >_L a|}{|a >_L a| + 1} & \text{if } a = b \end{cases}$$

for example, through this equation, it is possible to uncover the fraction of times that one activity follows another and itself, what make possible to deal with loops of any length and add paths, that were not discovered by the α -algorithm, improving the model.

The basic idea is that infrequent paths, should not be incorporated into the model. That makes the approach to the solution more robust than most other approaches.

This approach is very generic but there are other powerful heuristic approaches such as Fuzzy mining.

For further information:

- Heuristic mining: [33], pp 201-207.
- Fuzzy mining: [43]

Genetic Process Mining

Evolutionary approaches such as Genetic algorithms for process mining try to mimic the natural evolution. Such approaching is not deterministic and depend on randomization to find new solutions.

Like in any genetic algorithm there are four main steps: initialization, selection, reproduction, and termination.

Step	Description
Initialization	The initial population is created. The processes (individuals) generated may have little to do with the event log. While the activities are the same, the behaviour is completely different.
Selection	The fitness of each individual is computed. Fitness function determines the quality of each individual with the log. A simple example of fitness function can be the proportion of traces in the log that can be replayed in the model.
Reproduction	The selected parents, are used to create a new offspring. Here two genetic operators are used, crossover and mutation. First is applied the crossover, two parents (models) are taken and used for creating two new models. These models are added to the pool with the other child models. Then these child models are modified by different mutations. Again, fitness is computed and the best past to the next round (elitism) and produce new offspring.
Termination	The evolution process terminates when a satisfactory solution is found. A determined fitness can be an example of the criteria for terminating the process.

Table 29 Four steps for Genetic Process Mining. Source: own elaboration based on [33], pp 207-209.

For further information: [33], pp 207-209.

Inductive Mining

Inductive process discovery techniques are based on well-known data mining technique, Decision Trees. The translation of the decision tree into the inductive process discovery techniques is a Process Tree. Inductive mining techniques include members that can handle infrequent behaviours and deal huge models and logs while ensuring formal correctness criteria. On the other hand, the result returned by these techniques can be easily turned into Petri-nets or BPMN models. Inductive mining is currently one of the leading process discovery approaches due to its flexibility, formal guarantees and scalability.

For further information: [33], pp 222-236.

5. Conformance checking

After discussing briefly how does process discovery work in order to construct a model from an event log. It is time to suppose that both model and event log are available, on the one hand, the model might be made by hand or might be discovered. Moreover, the model may be normative, how perfectly rational should be the model's behaviour, or may be descriptive, which try to explain the real behaviour of the process model.

Conformance checking relates events in the log with the activities in the process model and compare both. It is relevant for business alignment and auditory, since the log can be replayed in order to find deviations suggesting fraud or inefficient. Besides, conformance checking can be used for measuring the performance of the discovery techniques, and repairing models that are not aligned with the reality.

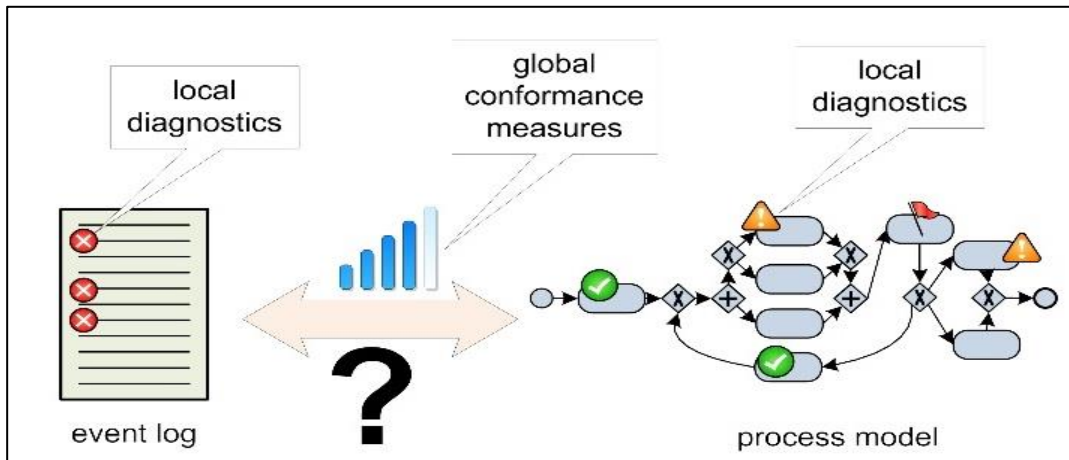


Figure 20. The main idea of Conformance Checking. Source: [33], p 244.

5.1. Main uses of conformance checking: Business alignment and auditory

There are many practical applications of conformance checking, nevertheless, the most known are: Business alignment and auditory.

Name	Goal	Motivation	Process mining as support
Business Alignment	The goal is to make sure that the information system and the real business process are well alignment.	Most organizations use product software, these softwares are not developed for a specific organization. A good example is SAP, which is based on so-called "best practices", i.e., typical scenarios and processes are implemented. Despite of such process are configurable, the particular needs of an organization may be not the same from what the software offers. On the other hand, processes may change faster that the information system, due to external influences.	Process mining can assist in improving the alignment of information systems, business processes, and the organization. By analysing the real processes and diagnosing deviations and discrepancies.

<p>Audit</p>	<p>Evaluating the organizations and their processes.</p>	<p>Ensure that the processes are executed respecting the boundaries set by managers, governments, and other stakeholders. Specific rules may be enforced by law or companies' policies. And the auditor should check whether these rules are followed or not. It is made by analysing a small set of samples of different processes.</p>	<p>Process mining allows to evaluate all events in a business process. And It can be done while the process is still running, providing new ways of auditing.</p>
--------------	--	--	---

Table 30 Business Alignment and Auditory. Source: own elaboration based on: [33], pp 243-246.

Part of audits, are bound to check GRC, “*Governance, Risk management and Compliance*”. These terms surround the organizations’ capability to reliably achieve objectives, address uncertainty and act with integrity. It is important to highlight that process mining, in particular conformance checking, is a powerful tool for supervising the compliance of the standards and regulations both external and internal. Since conformance checking can be used to reveal deviations, defects, even near incidents. However, as mentioned, there are other uses for process mining, such as repairing models, which consist in, after applying conformance checking techniques, it is possible to make a diagnosis about the model and paths that are never taken can be removed from the model or, to the contrary, activities those often appear and are not in the model, can be added to the model.

After an introduction of the conformance checking and how it can help the companies to align their business process and improve the quality of the audits. Now it is moment to explain the main techniques to reach these targets.

5.2. Techniques

The conformance checking techniques are mainly focused in one of the four quality criteria. Among fitness, precision, generalization, and simplicity, fitness is what is most related with conformance, that is why it will be introduced how fitness can be quantified.

5.2.1. Comparing Footprints

The first conformance checking approach is the so-called footprint comparison, based on [33], pp 263-267. The concept of footprint matrix showing causal dependencies in an event log was introduced in Figure 16. From the point of view of models, it may be also made a footprint matrix by playing-out the model and recording the execution sequences. It is very simple to calculate the fitness for this approach. After building both footprint matrices, the model’s matrix and the log’s matrix, are overlapped. The number of cells that do not match are computed, and the value of the fitness is reached through the expression:

$$fitness(L, N) = 1 - \frac{N_{dev}}{N}$$

Note that, N_{dev} , is the number of cells that do not match, whereas N is the total number of cells on the matrix.

The following event log, L_{full} , and models $N1$ and $N2$, will illustrate how does this approaching work.

Event log L_{full} :

Frequency	Reference	Trace
455	σ_1	(a,c,d,e,h)
191	σ_2	(a,b,d,e,g)
177	σ_3	(a,d,c,e,h)
144	σ_4	(a,b,d,e,h)
111	σ_5	(a,c,d,e,g)
82	σ_6	(a,d,c,e,g)
56	σ_7	(a,d,b,e,h)
47	σ_8	(a,c,d,e,f,d,b,e,h)
38	σ_9	(a,d,b,e,g)
33	σ_{10}	(a,c,d,e,f,b,d,e,h)
14	σ_{11}	(a,c,d,e,f,b,d,e,g)
11	σ_{12}	(a,c,d,e,f,d,b,e,g)
9	σ_{13}	(a,d,c,e,f,c,d,e,h)
8	σ_{14}	(a,d,c,e,f,d,b,e,h)
5	σ_{15}	(a,d,c,e,f,b,d,e,g)
3	σ_{16}	(a,c,d,e,f,b,d,e,f,d,b,e,g)
2	σ_{17}	(a,d,c,e,f,d,b,e,g)
2	σ_{18}	(a,d,c,e,f,b,d,e,f,b,d,e,g)
1	σ_{19}	(a,d,c,e,f,d,b,e,f,b,d,e,h)
1	σ_{20}	(a,d,b,e,f,b,d,e,f,d,b,e,g)
1	σ_{21}	(a,d,c,e,f,b,d,e,f,c,d,e,f,d,b,e,g)

Table 31 Event log L : a= register request, b= examine thoroughly, c= examine casually, d= check ticket, e= decide, f= reinitiate request, g= pay compensation, and h= reject request. Source: [33], p 247.

Models N_1 and N_2 :

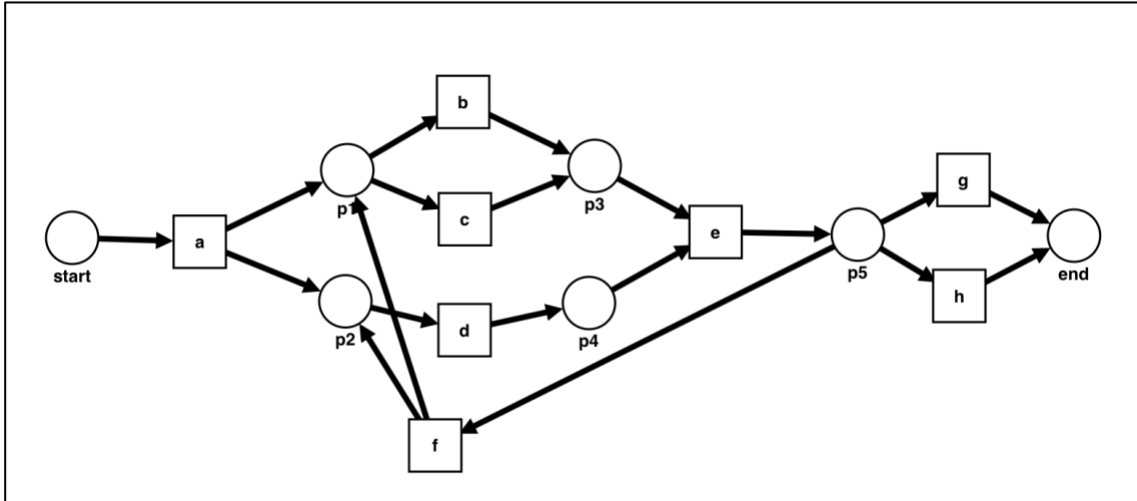


Figure 21. Model N_1 . Source: [33], p 248.

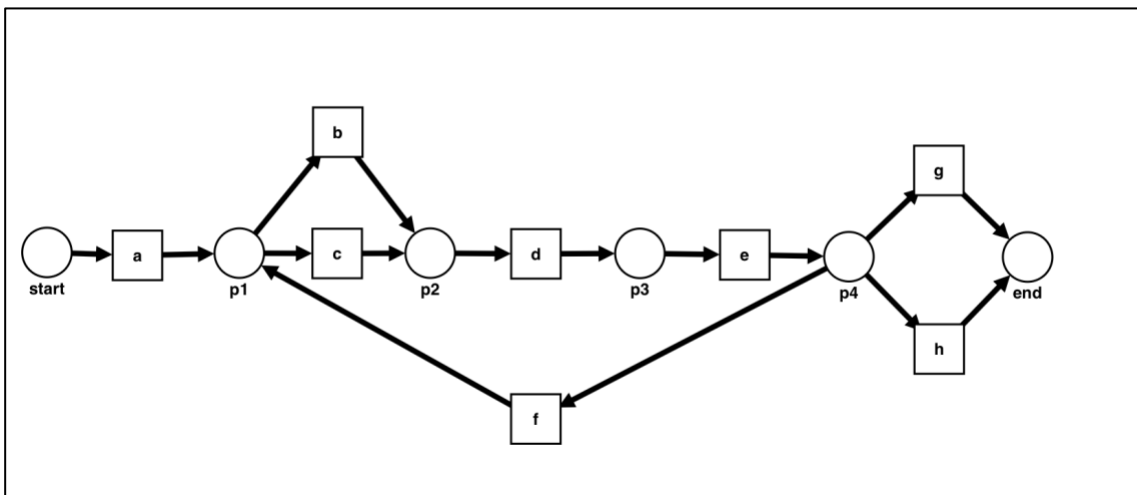


Figure 22. Model N_2 . Source [33], p 248.

Therefore, the footprint matrixes for L_{full} and N_1 are:

Footprint of L_{full} :

	a	b	c	d	e	f	g	h
a	#	→	→	→	#	#	#	#
b	←	#	#		→	←	#	#
c	←	#	#		→	←	#	#
d	←			#	→	←	#	#
e	#	←	←	←	#	→	→	→
f	#	→	→	→	←	#	#	#
g	#	#	#	#	←	#	#	#
h	#	#	#	#	←	#	#	#

Figure 23. Footprint of L_{full} . Source: [33], p 264.

Footprint of N_1 :

	a	b	c	d	e	f	g	h
a	#	→	→	→	#	#	#	#
b	←	#	#		→	←	#	#
c	←	#	#		→	←	#	#
d	←			#	→	←	#	#
e	#	←	←	←	#	→	→	→
f	#	→	→	→	←	#	#	#
g	#	#	#	#	←	#	#	#
h	#	#	#	#	←	#	#	#

Figure 24. Footprint of N_1 . Source: [33], p 264.

The fitness for this example is 1, due to the footprint of both, L_{full} and N_1 are the same. Now it is the turn of N_2 :

Footprint of N_2 :

	a	b	c	d	e	f	g	h
a	#	→	→	#	#	#	#	#
b	←	#	#	→	#	←	#	#
c	←	#	#	→	#	←	#	#
d	#	←	←	#	→	#	#	#
e	#	#	#	←	#	→	→	→
f	#	→	→	#	←	#	#	#
g	#	#	#	#	←	#	#	#
h	#	#	#	#	←	#	#	#

Figure 25. Footprint of N_2 . Source [33], p 264.

Therefore, there are some cells that do not match:

	a	b	c	d	e	f	g	h
a				→:#				
b				:→	→:#			
c				:→	→:#			
d	←:#	:←	:←			←:#		
e		←:#	←:#					
f				→:#				
g								
h								

Figure 26. Differences between L_{full} and N_2 . Source: [33], p 264.

The cells in red in the Figure 26, are the ones that do not match, note that for example, according to the event log, activity “a” can be followed by “d”, however, in the model

this movement is not possible. Now, it is easy to compute the fitness by applying the previous equation. Thus, the conformance based on the footprints is $1 - \frac{12}{64} = 0.8125$. Conformance analysis based on footprints is only useful if the log is complete, it means that all activities that can follow one another do so at least once in the log. Because of usually the log is not complete this approaching won't be chosen for the experimental part of this literature.

5.2.2. Token replay

Token replay approach, based on [33], pp 246-267, consists in, having an event log and a Petri-net model, the event log is replayed trace by trace. By doing it, it is possible to uncover deviations between the model and the reality. The basic idea of token replay is the following one. Over the model a token is generated, it will be consumed and again generated by executing activity by activity of each trace moving through the places of the model. When one activity is not possible to be executed in the order that the trace indicates, or that activity does not exist in the model, one fictitious token will be created in order to finish the succession of activities in that trace, whereas the real token will stay remaining. After doing it with the whole log, the number of tokens that have been created, consumed, generated, and remaining, are quantified, making possible to calculate the fitness.

Finally, there are four accumulative parameters that must be taken into account in order to calculate the fitness.

Name	Abbreviation	Explanation
Produced tokens	p	When the environment produces tokens in the places, then a token is "produced".
Consumed tokens	c	When the environment consumes tokens in the transitions, then a token is "consumed".
Missing tokens	m	When one of the transitions cannot be triggered in the model and it is necessary to add a token in some place for completing the replaying of the trace. Then a missing token is added.
Remaining tokens	r	When a token remains in a place and is not consumed after replaying the trace. Then a token is remaining.

Table 32. Token replay fitness parameters. Source: own elaboration.

After replaying a trace, fitness of the trace σ replayed in the model N , follows the equation:

$$fitness(\sigma, N) = \frac{1}{2} \left(1 - \frac{m}{c} \right) + \frac{1}{2} \left(1 - \frac{r}{p} \right)$$

Therefore, they will be shown a couple of example, in order to illustrate how does token replay approach work.

Supposing the Event log L_{full} introduced before and the models N1 also introduced above alongside the new model N3:

Model N₃:

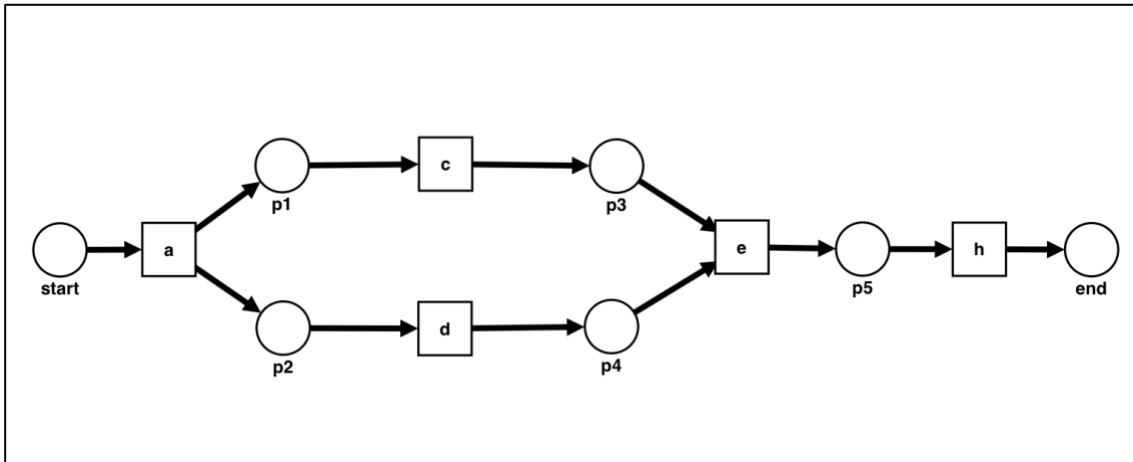
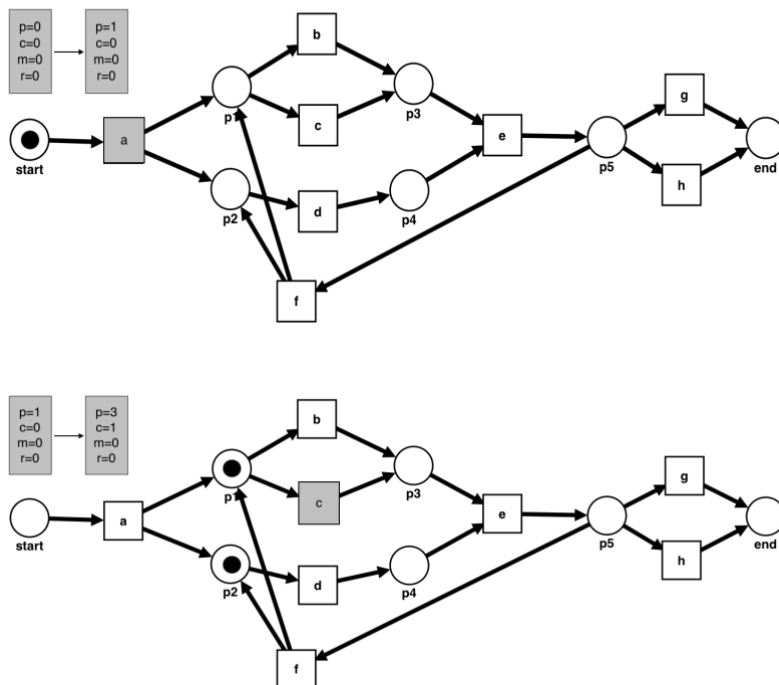
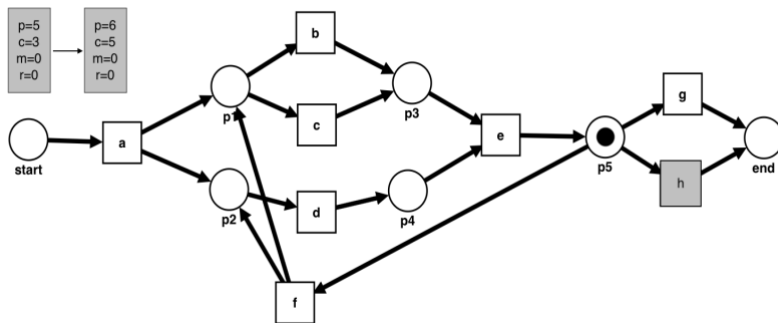
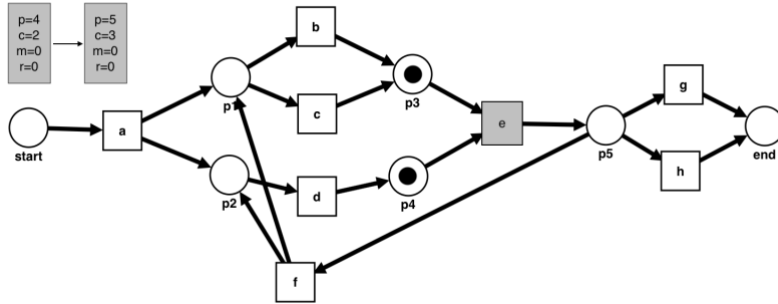
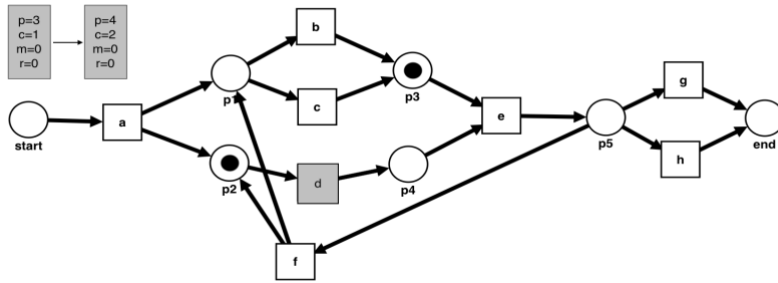


Figure 27. N₃. Source: own elaboration based: [33], p 248.

In the following figure, Figure 28, It will be replayed $\sigma_1: (a,c,d,e,h)$ over N1, showing how accumulative parameters (p, c, m, r) evolve.





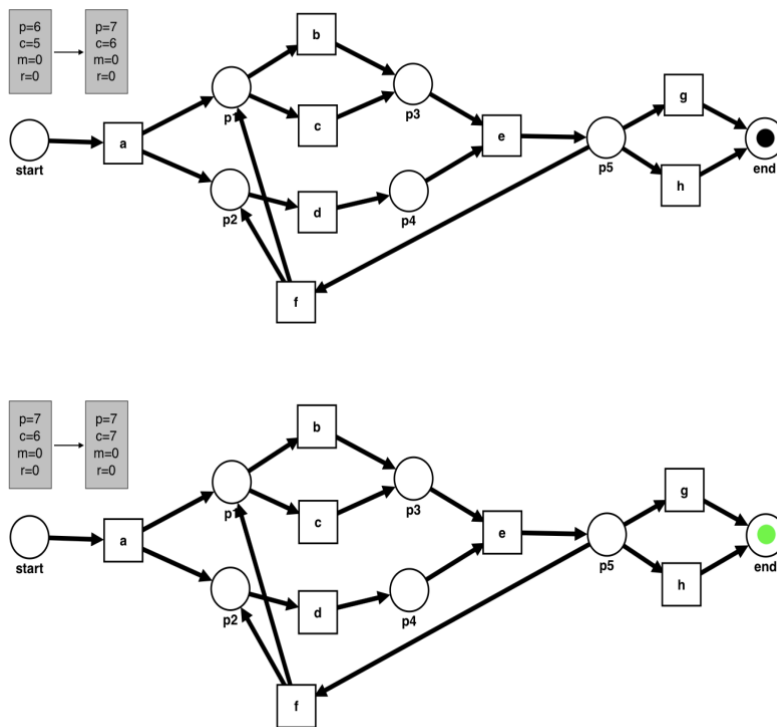
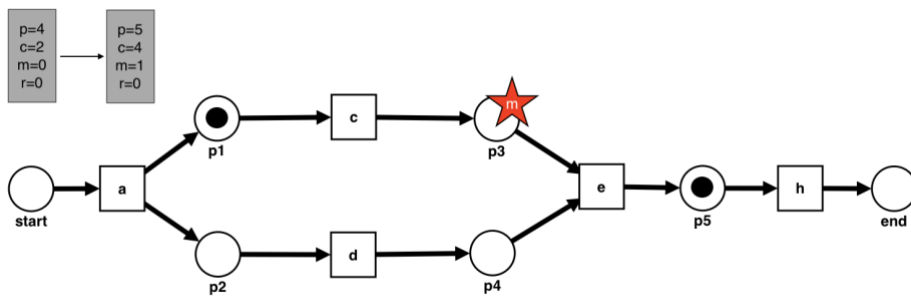
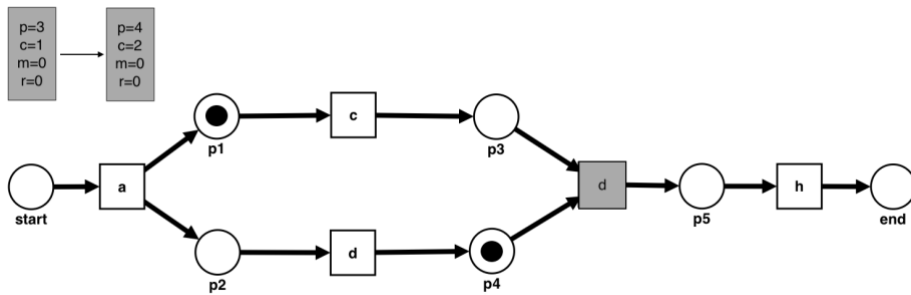
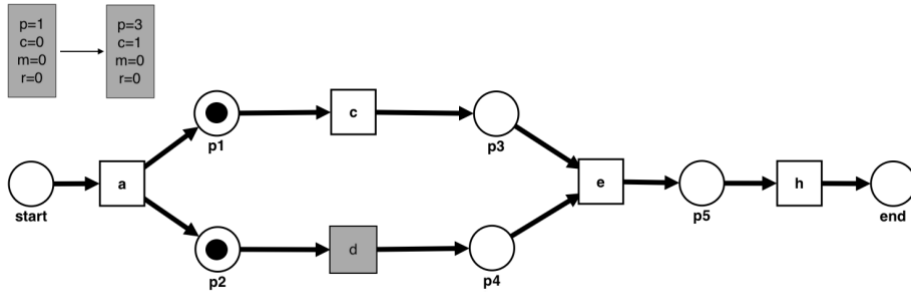
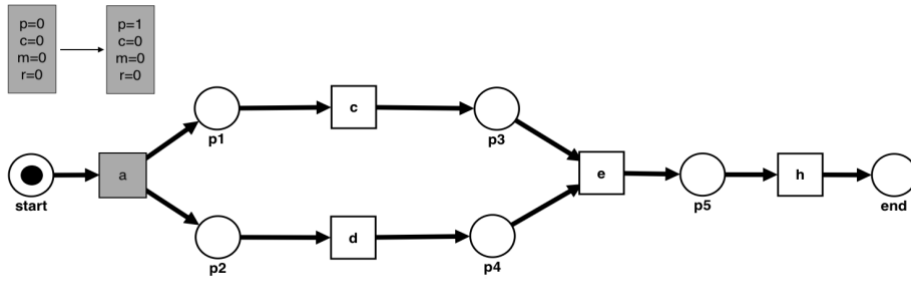


Figure 28. Token replay of trace σ_1 over model N_1 . Source: Own elaboration.

After replay σ_1 within the model N_1 , it is possible to get the fitness parameter by applying the equation.

$$fitness(\sigma_1, N_1) = \frac{1}{2} \left(1 - \frac{0}{7} \right) + \frac{1}{2} \left(1 - \frac{0}{7} \right) = 1$$

In conclusion, the fitness of the trace σ_1 and the model match perfect one to another. The next example, trace σ_2 (a,b,d,e,g) replayed over model N_3 . First, there are some points to take in mind. First of all, some activities of the trace do not correspond to any activity in the model. Hence, σ_2 is replaced by $\sigma'_2 = (a,d,e)$. The following figure shows the process of replaying σ'_2 in the model N_3 .



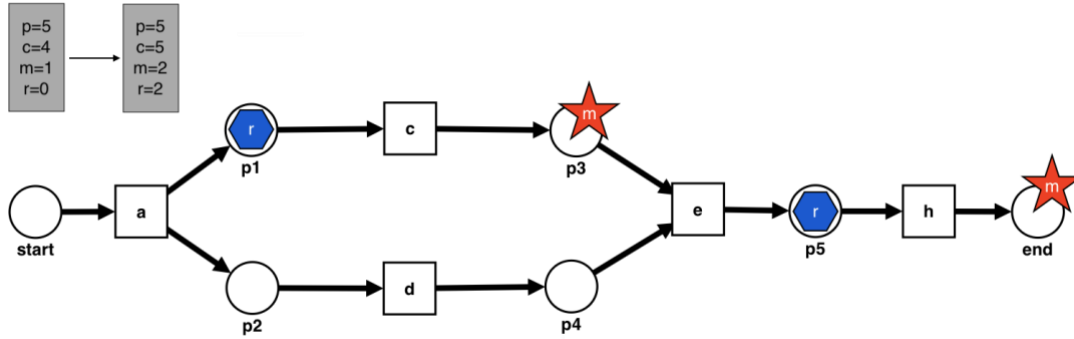


Figure 29. To replay σ_2 (a,b,d,e,g), first is necessary to remove activities that not correspond with the model. Replaying $\sigma'_2 = (a,d,e)$ within the model N_3 shows that there are either tokens remaining or missing. Source: own elaboration.

The first problem appears when replaying e. It is because c does not appear in the trace, however, in the model is necessary to trigger e, so a missing-tag is added in p3. After replaying σ'_2 , there are two marks, p1 and p5, since activity h does not appear in the trace. Now the model needs to consume one token from place end, nevertheless, end is not marked. Thus, a new missing-tag is added in place end. Moreover, two tokens are remaining. These are the final values for the parameters $p=5$, $c=5$, $m=2$, $r=2$. Therefore, fitness for this trace in the model is the following.

$$fitness(\sigma_2, N_2) = \frac{1}{2} \left(1 - \frac{2}{5} \right) + \frac{1}{2} \left(1 - \frac{2}{5} \right) = 0.6$$

It is simple to deduce why the conformance is so poor:

- c should happen according to the model but did not do it.
- e happened but it was impossible according to the model.
- h was supposed to happen, but it did not happen.

These two examples show how to analyse the fitness of a single trace. Now the approach for the whole Event log is nearly the same:

$$fitness(L, N) = \frac{1}{2} \left(1 - \frac{\sum_{\sigma \in L} L(\sigma) \times m_{N,\sigma}}{\sum_{\sigma \in L} L(\sigma) \times c_{N,\sigma}} \right) + \frac{1}{2} \left(1 - \frac{\sum_{\sigma \in L} L(\sigma) \times r_{N,\sigma}}{\sum_{\sigma \in L} L(\sigma) \times p_{N,\sigma}} \right)$$

Note that $\sum_{\sigma \in L} L(\sigma) \times m_{N,\sigma}$ is total number of missing tokens while replaying the entire log, because $L(\sigma)$ is the frequency of trace σ and $m_{N,\sigma}$ is the number of missing tokens for a single instance of σ . The closer is fitness to 1, the better is the fitness between the event log and the model. The rest variables are defined in the same way.

By replaying the entire log L_{full} in both example model, the fitness can now be computed.

$$fitness(L_{full}, N_1) = 1$$

$$fitness(L_{full}, N_2) = 0.8797$$

Focusing on the model N_3 the diagnosis suggests the following problems:

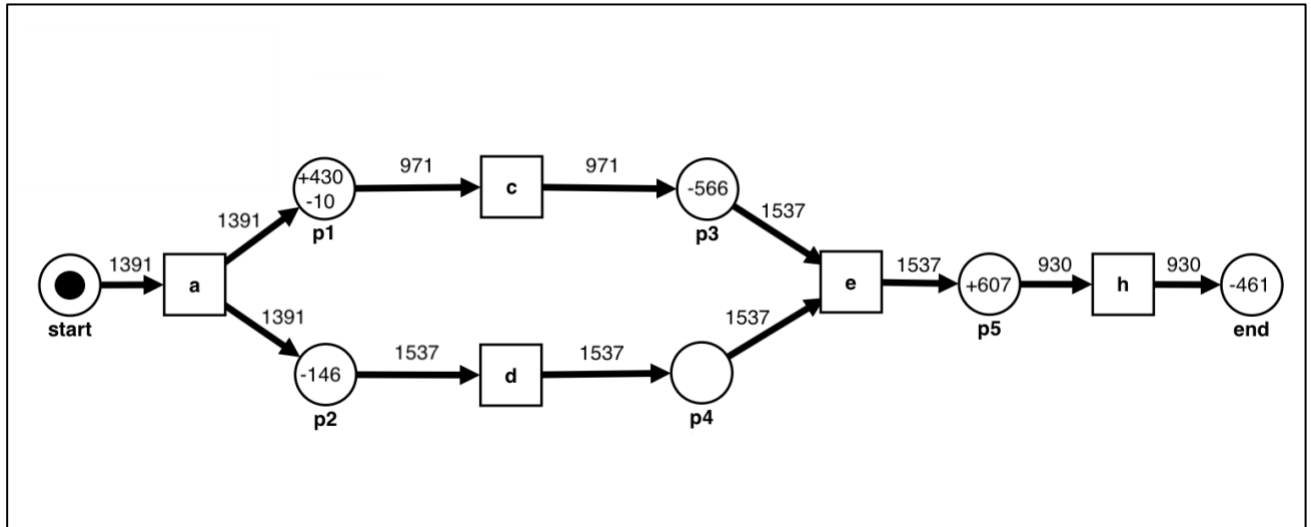


Figure 30. Result of replaying L_{full} in model N3. Numbers on the rows point the number of times an activity happened, the numbers in the places indicate the missing tokens (- symbol) and the remaining tokens (+ symbol).
 Source: own elaboration based on: [33], p 255.

Place	Problem detected	Interpretation
p1	430 tokens remain in p1, because c did not happen while the model expected c to happen.	430 cases should be examined casually (activity c), but they were not.
p1	10 tokens were missing in p1, because c happened while it was not possible according to the model.	10 cases were examined casually (activity c) without registering any request (activity a)
p2	146 tokens were missing in p2, because d happened while it was not possible according to the model.	146 times the ticket was checked (activity d) without registering any request (activity a)
p3	566 tokens were missing in p3, because e happened while it was not possible according to the model.	566 times a decision was made (activity e) without being examined casualty (activity c)
p5	607 tokens remain in p5, because h did not happen while the model expected h to happen.	607 cases should be rejected (activity h), but they were not.
end	461 tokens of 1931 did not reach place end.	461 cases did not reach the place end because the request was not rejected (activity h)

Table 33. Problems diagnosed applying token replay approaching. Source: own elaboration.

Once the diagnosis is made, it is moment to go further. Many options are now available. The log can be split into two sublogs, one containing all the fitting cases, and another one with the non-fitting cases. Each sublog can be used for further analysis, for example

it is possible to use the non-fitting cases and apply discovery techniques, in order to find new process models. Moreover, it is interesting to know which resources handled the deviating cases and if the deviation provoked an increase in cost and time or, on the contrary, it saved money or time.

On the other hand, if fraud or policy violations are suspected, it is possible now to create a social network for analysing the relationships between the resources and find the core of the deviation, but it is another field of work, enhance techniques, therefore, it is not going to be discussed in this literature.

In conclusion, using token replay it can be differentiate between fitting and non-fitting cases. Moreover, it is a very intuitive and easy to apply approach. Nevertheless, it has some drawbacks.

- This approach is Petri-net specific.
- If a case does not fit, the approach does not create the corresponding path through the model, and the Petri-net is flooded by remaining and missing tokens.
- Token replay becomes complicated when duplicate and silent activities appear in the model.

Alignments approaching is introduced to overcome those problems.

5.2.3. Alignments

In an alignment γ , there are two rows, the top row corresponds to any case σ and the bottom row correspond to a path from the initial marking to the final marking of a model N .

The basic idea is that every trace is replayed over the model, facing that trace against a possible path of the model. The number of times that, the trace cannot make a movement because the activity cannot be executed in the model, or the model, cannot make a movement, because of, for example, the activity which appear in the trace does not exist in the model, are quantified and minimized. Note that alignments are used not only in process mining, but in many other disciplines such as bioinformatics. Therefore, there are many algorithms that can solve the optimization problem in order to find the best alignment possible. Afterwards the fitness is calculated through one equation that will be shown later on.

In order to explain in more depth the notions of alignments, based on [33], pp 256-263, it will be considered the trace $\sigma_3 = \langle a, d, b, e, h \rangle$, and both models N_1 Figure 31 and N_3 Figure 32:

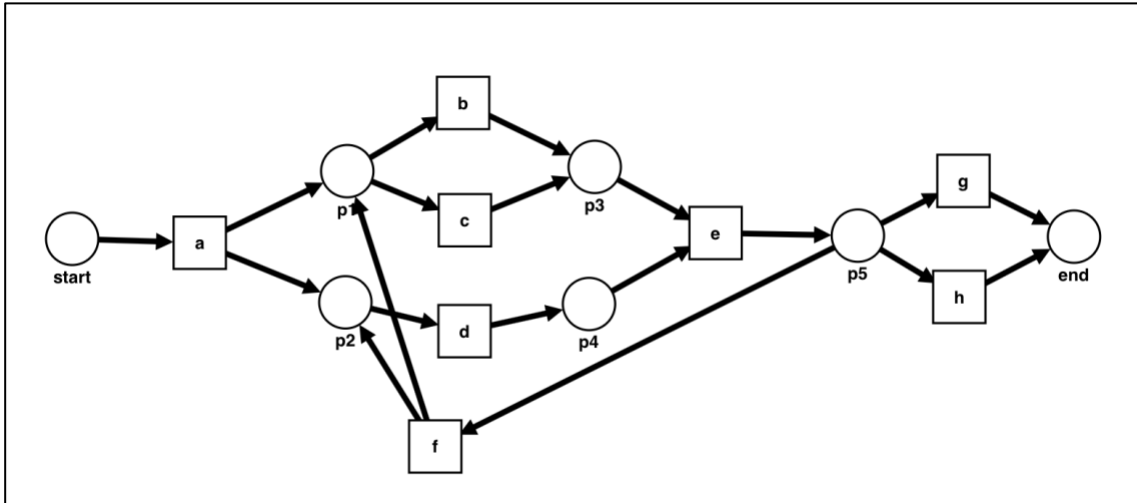


Figure 31. Model N_1 . Source: [33], p 248.

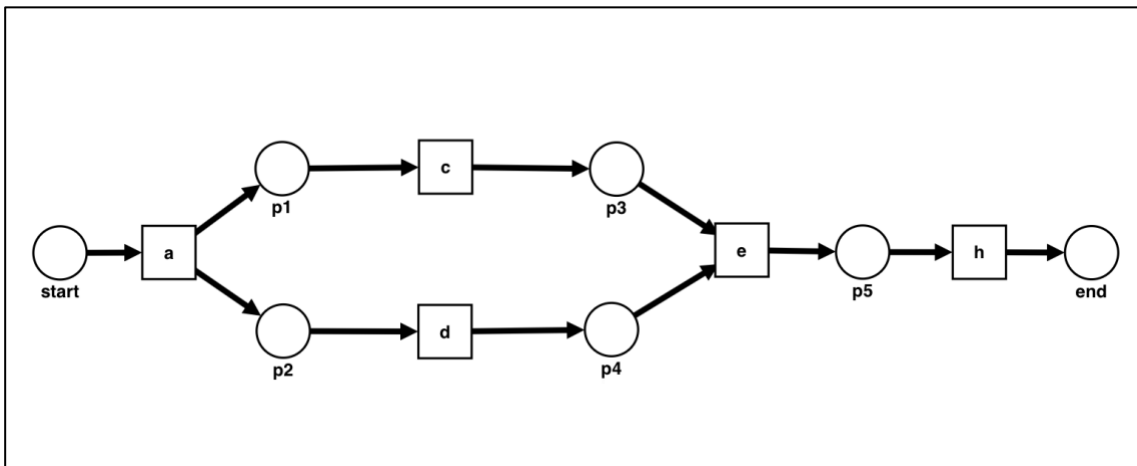


Figure 32. N_3 . Source: own elaboration based: [33], p 248.

The alignment between σ_3 and N_1 fits perfectly,

$$\gamma_1 = \begin{array}{|c|c|c|c|c|} \hline a & d & b & e & h \\ \hline a & d & b & e & h \\ \hline \end{array}$$

It corresponds to an optimal alignment. A so-called optimal alignment is the best match given a trace and a model, however, it must not be necessary only one perfect match as it is shown in the case of σ_3 and N_3 ,

$$\gamma_{3a} = \begin{array}{|c|c|c|c|c|} \hline a & \gg & d & b & e & h \\ \hline a & c & d & \gg & e & h \\ \hline \end{array}$$

$$\gamma_{3b} = \begin{array}{|c|c|c|c|c|} \hline a & d & \gg & b & e & h \\ \hline a & d & c & \gg & e & h \\ \hline \end{array}$$

$$\gamma_{3c} = \begin{array}{|c|c|c|c|c|} \hline a & d & b & \gg & e & h \\ \hline a & d & \gg & c & e & h \\ \hline \end{array}$$

There are three optimal alignments. The symbol “ \gg ” indicates misalignments. In the case γ_{3a} , the model makes a “c move” before d may occur. Afterwards the log makes a “b move”, what is impossible to occur in the model. The same interpretation follows the other two examples γ_{3b} and γ_{3c} .

It is possible to extract useful information from the alignments, such as what activities are often skipped or the ones that occur when it is not supposed to happen.

Alignments also can be applied in any representation language, it is not Petri-net specific. Moreover, token-based approach has many problems to deal with duplicate and silent activities, whereas alignments approaching has not problems dealing with them. Note that silent activities are those do not have a specific purpose or role in the process, however are needed in the model for structuring purposes. The following example tries to illustrate that.

The following model N_4 , Figure 33, has a duplicated and a silent activity, and it will be very useful to illustrate it. When a model has duplicate activities, it is useful to add, in the bottom row of the alignment (model moves), the transition that it is being fired. Now it is possible to define a move as a pair $(x, (y, t))$ where the first element refers to the log and the second element refers to the model.

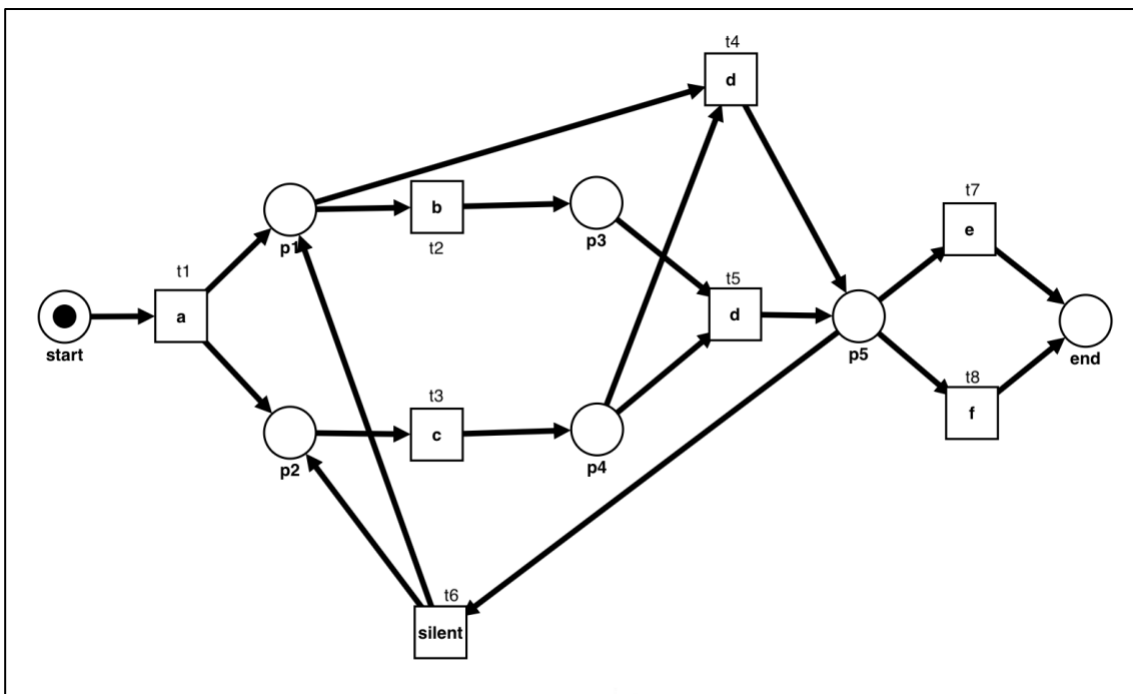


Figure 33. Model N_4 with duplicate and silent activities. Source [33], p 258.

Given $\sigma_4 = \langle a, c, d, b, c, d, c, d, c, b, d, f \rangle$ and the model N_4 , the optimal alignment would be,

$$\gamma_{4,4} = \begin{array}{|c|c|c|c|c|c|c|c|c|c|c|c|c|c|c|} \hline a & c & d & \gg & b & c & d & \gg & c & d & \gg & c & b & d & f \\ \hline a & c & \mathbf{d} & S & b & c & \mathbf{d} & S & c & \mathbf{d} & S & c & b & \mathbf{d} & f \\ \hline t1 & t3 & \mathbf{t4} & t6 & t2 & t3 & \mathbf{t5} & t6 & t3 & \mathbf{t4} & t6 & t3 & t2 & \mathbf{t5} & t8 \\ \hline \end{array}$$

Thanks to the alignment, it is possible to see how the model makes movements firing t4 or t5, depending on whether the trace follows one path or another, but the activity is the same.

On the other hand, all “»” refer to the silent activity, a silent activity is not considered as a misalignment and it is only indicating that the trace loops back three times.

Apart from the optimal alignments, there are many other (sometimes infinitely) combinations of alignments. For the example for model N4 and given $\sigma_5 = \langle a, c, d, f \rangle$ above an option might be,

$$\gamma_{4,5} = \begin{array}{|c|c|c|c|c|c|c|c|c|c|} \hline & a & c & d & f & \gg & \gg & \gg & \gg & \gg \\ \hline \gg & \gg & \gg & \gg & \gg & a & b & c & d & f \\ \hline & & & & & t1 & t2 & t3 & t4 & t8 \\ \hline \end{array}$$

In order to choose the most appropriate alignment, it is added a cost function $\delta(x, (y, t))$ to penalise undesirable moves.

$\delta(x, (y, t))$	Type of move	Cost
$\delta(x, (y, t)) / x=y$	Synchronous moves	0
$\delta(\gg, (S, t)) / S= \text{silent}$	Invisible model moves	0
$\delta(\gg, (y, t))$	visible model moves	> 0
$\delta(x, (\gg))$	log move	> 0

Table 34. Values of the different moves. Source: own elaboration.

Cost function may vary depending on the nature of the activity and. For simplicity, it is going to be assumed a binary cost function where value may be 1 or 0 depending of the move in the alignment.

An alignment is optimal if there are not any other alignment with lower costs. Once the cost function is introduced, it is possible to turn cost values into a fitness function,

$$fitness(\sigma, N) = 1 - \frac{\delta(\lambda_{opt}^N(\sigma))}{\delta(\lambda_{worst}^N(\sigma))}$$

where:

- $\lambda_{opt}^N(\sigma)$ is the optimal alignment of the trace σ .
- $\lambda_{worst}^N(\sigma)$ is the worst scenario alignment of the trace σ .
- $\delta(\lambda_{opt}^N(\sigma))$ is the cost of the optimal alignment.
- $\delta(\lambda_{worst}^N(\sigma))$ is the cost of the worst scenario alignment. The worst scenario is that there are no synchronous moves and only “moves in model only” and “moves in log only”.

Assuming that $\gamma_{4,5}$ is worst scenario and the optimal alignment fits perfect (cost =0), $fitness(\sigma_5, N_4) = 1 - \frac{0}{9} = 1$. Fitness can be extended to the whole log as:

$$fitness(L, N) = 1 - \frac{\sum_{\sigma \in L} L(\sigma) \times \delta(\lambda_{opt}^N(\sigma))}{\sum_{\sigma \in L} L(\sigma) \times \delta(\lambda_{worst}^N(\sigma))}$$

Where $\sum_{\sigma \in L} L(\sigma) \times \delta(\lambda_{opt}^N(\sigma))$ is the sum of all costs when replaying the entire event log with optimal alignment. The same happens for $\sum_{\sigma \in L} L(\sigma) \times \delta(\lambda_{worst}^N(\sigma))$, but in this case with the worst scenario.

In order to show an example, it will be taken L_{full} and the model N_3 . Several activities appearing in this log are not represented in the model, it causes that there will be several “log moves”. Generating a particular collection of optimal alignments and the worst-case scenario, they are obtained the following results.

Type	Number	Cost
Synchronous move	6064	0
Log move	1475	1475
Model move	891	891
Worst-case scenario	14494	14494

Table 35. Results of applying alignment approaching to L_{full} and N_3 . Source: own elaboration based on: [33], p 263.

Therefore,

$$fitness(L_{full}, N_3) = 1 - \frac{2366}{14494} = 0.8368$$

The diagnosis suggests the following problems:

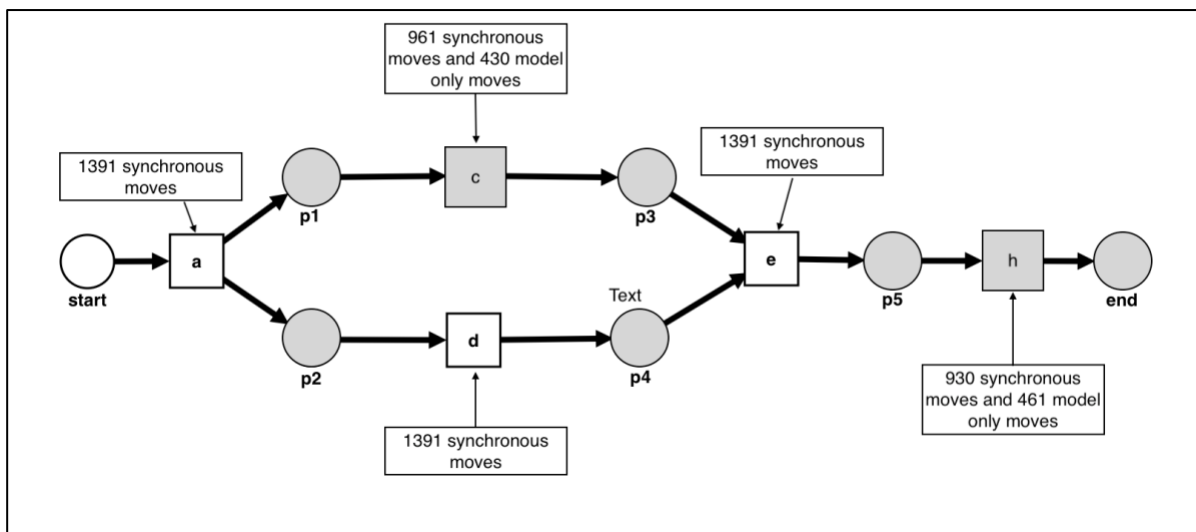


Figure 34. Result of alignment approaching over the model N_3 and the event log L_{full} . Source: [33], p 263.

- The shaded places indicate that activities were executed despite it was not possible according to the model. There were 1475 log only moves.
- Activity c was skipped 430 times.

- Activity h was skipped 461 times.

After analysing both approaches, it is possible to set some advantages [33], pp 262-263:

- Alignments provides more accurate diagnostics than token-based replay because the last one may provide misleading due to remaining tokens.
- Alignments are configurable through the cost function.
- Alignments can be used to map each case onto a feasible path in model. This is important for projecting information, such as bottlenecks, on models. Token-based replay also relates observed and modelled behaviour, but does not create the corresponding end-to-end execution sequence in the model.
- Alignments are independent of the representation bias used for the model.

In conclusion, after introducing the main techniques for conformance checking it has been decided that the following experimental chapter, where it will be applied conformance checking in order to compare a SAP business process with a model, will be made following alignment approach.

6. Application of Process Mining techniques on a data log given

In this chapter, a real data log from the Hochschule Albstadt-Sigmaringen SAP server will be analysed, pre-processed, the model will be discovered, and finally the conformance will try to be checked, in order to show how a real process mining project might be managed.

ProM Lite 1.2, will be the software used for developing this chapter, ProM Lite 1.2 is an open source process mining tool developed by Technische Universiteit Eindhoven (TU/e).

6.1. Presentation of the problem and data

The problems presented in this chapter are:

- Pre-Processing: how to turn a raw data file into a suitable event log.
- Discovering: analyse that event log for finding the real business process behind the data.
- Checking the conformance: in order to know, how good the process model explains the real behaviour.

The data file it is going to work with is called Saleslog_data_C.xlsx. It is an excel data log extracted from SAP, and turned into an .xlsx file. The data log contains the changes made in different sale orders, due to, for example, a too high price. By the analysis of this file it will try to deduce a business process. The file has been made by several pupils from the WIN bachelor's degree at Hochschule Albstadt-Sigmaringen.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
	Change Document	Object Value	Document Number	Table Name	Table Key	Field Name	Change Indicator	Text flag	Unit	Unit	CUKY	CUKY	New value	Old value	Data Filter Value for Data Aging
1															
2	VERKBELEG	000000009	377585	VBAP	2020000000009000010	ABGRU	U	1					03		
3	VERKBELEG	000000009	377585	VBAP	2020000000009000020	ABGRU	U	1					03		
4	VERKBELEG	000000009	377585	VBAP	2020000000009000030	ABGRU	U	1					03		
5	VERKBELEG	000000009	377585	VBUK	2020000000009	GBSTK	U	1				C		A	
6	VERKBELEG	000000010	377586	VBAP	2020000000010000010	ABGRU	U	1					03		
7	VERKBELEG	000000010	377586	VBAP	2020000000010000020	ABGRU	U	1					03		
8	VERKBELEG	000000010	377586	VBAP	2020000000010000030	ABGRU	U	1					03		
9	VERKBELEG	000000010	377586	VBUK	2020000000010	GBSTK	U	1				C		A	
10	VERKBELEG	000000011	377587	VBAP	2020000000011000010	ABGRU	U	1					03		
11	VERKBELEG	000000011	377587	VBAP	2020000000011000020	ABGRU	U	1					03		
12	VERKBELEG	000000011	377587	VBAP	2020000000011000030	ABGRU	U	1					03		
13	VERKBELEG	000000011	377587	VBUK	2020000000011	GBSTK	U	1				C		A	
14	VERKBELEG	000000013	377588	VBAP	2020000000013000010	ABGRU	U	1					03		
15	VERKBELEG	000000013	377588	VBAP	2020000000013000020	ABGRU	U	1					03		
16	VERKBELEG	000000013	377588	VBAP	2020000000013000030	ABGRU	U	1					03		
17	VERKBELEG	000000013	377588	VBUK	2020000000013	GBSTK	U	1				C		A	
18	VERKBELEG	000000014	377589	VBAP	2020000000014000010	ABGRU	U	1					03		
19	VERKBELEG	000000014	377589	VBAP	2020000000014000020	ABGRU	U	1					03		
20	VERKBELEG	000000014	377589	VBAP	2020000000014000030	ABGRU	U	1					03		
21	VERKBELEG	000000014	377589	VBUK	2020000000014	GBSTK	U	1				C		A	
22	VERKBELEG	000000015	377590	VBAP	2020000000015000010	ABGRU	U	1					03		
23	VERKBELEG	000000015	377590	VBAP	2020000000015000020	ABGRU	U	1					03		
24	VERKBELEG	000000015	377590	VBAP	2020000000015000030	ABGRU	U	1					03		
25	VERKBELEG	000000015	377590	VBUK	2020000000015	GBSTK	U	1				C		A	
26	VERKBELEG	000000016	377591	VBAP	2020000000016000010	ABGRU	U	1					03		
27	VERKBELEG	000000016	377591	VBAP	2020000000016000020	ABGRU	U	1					03		
28	VERKBELEG	000000016	377591	VBAP	2020000000016000030	ABGRU	U	1					03		
29	VERKBELEG	000000016	377591	VBUK	2020000000016	GBSTK	U	1				C		A	
30	VERKBELEG	000000017	377592	VBAP	2020000000017000010	ABGRU	U	1					03		
31	VERKBELEG	000000017	377592	VBAP	2020000000017000020	ABGRU	U	1					03		
32	VERKBELEG	000000017	377592	VBAP	2020000000017000030	ABGRU	U	1					03		
33	VERKBELEG	000000017	377592	VBUK	2020000000017	GBSTK	U	1				C		A	
34	VERKBELEG	000000018	377593	VBAP	2020000000018000010	ABGRU	U	1					03		
35	VERKBELEG	000000018	377593	VBAP	2020000000018000020	ABGRU	U	1					03		
36	VERKBELEG	000000018	377593	VBAP	2020000000018000030	ABGRU	U	1					03		
37	VERKBELEG	000000018	377593	VBUK	2020000000018	GBSTK	U	1				C		A	
38	VERKBELEG	000000019	377594	VBAP	2020000000019000010	ABGRU	U	1					03		
39	VERKBELEG	000000019	377594	VBAP	2020000000019000020	ABGRU	U	1					03		
40	VERKBELEG	000000019	377594	VBAP	2020000000019000030	ABGRU	U	1					03		
41	VERKBELEG	000000019	377594	VBUK	2020000000019	GBSTK	U	1				C		A	

Figure 35. Small part of Saleslog_data_c.xls. Source: own elaboration.

6.2. Data pre-treatment

This section shows how Saleslog_data_C.xlsx turns into a suitable event log. The first step to keep in mind is that not the whole raw data is needed.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
	Change Document	Object Value	Document Number	Table Name	Table Key	Field Name	Change Indicator	Text flag	Unit	Unit	CUKY	CUKY	New value	Old value	Data Filter Value for Data Aging
1															
2	VERKBELEG	000000009	377585	VBAP	2020000000009000010	ABGRU	U	1					03		
3	VERKBELEG	000000009	377585	VBAP	2020000000009000020	ABGRU	U	1					03		
4	VERKBELEG	000000009	377585	VBAP	2020000000009000030	ABGRU	U	1					03		
5	VERKBELEG	000000009	377585	VBUK	2020000000009	GBSTK	U	1				C		A	
6	VERKBELEG	000000010	377586	VBAP	2020000000010000010	ABGRU	U	1					03		
7	VERKBELEG	000000010	377586	VBAP	2020000000010000020	ABGRU	U	1					03		
8	VERKBELEG	000000010	377586	VBAP	2020000000010000030	ABGRU	U	1					03		
9	VERKBELEG	000000010	377586	VBUK	2020000000010	GBSTK	U	1				C		A	
10	VERKBELEG	000000011	377587	VBAP	2020000000011000010	ABGRU	U	1					03		
11	VERKBELEG	000000011	377587	VBAP	2020000000011000020	ABGRU	U	1					03		
12	VERKBELEG	000000011	377587	VBAP	2020000000011000030	ABGRU	U	1					03		
13	VERKBELEG	000000011	377587	VBUK	2020000000011	GBSTK	U	1				C		A	
14	VERKBELEG	000000013	377588	VBAP	2020000000013000010	ABGRU	U	1					03		
15	VERKBELEG	000000013	377588	VBAP	2020000000013000020	ABGRU	U	1					03		
16	VERKBELEG	000000013	377588	VBAP	2020000000013000030	ABGRU	U	1					03		
17	VERKBELEG	000000013	377588	VBUK	2020000000013	GBSTK	U	1				C		A	
18	VERKBELEG	000000014	377589	VBAP	2020000000014000010	ABGRU	U	1					03		
19	VERKBELEG	000000014	377589	VBAP	2020000000014000020	ABGRU	U	1					03		
20	VERKBELEG	000000014	377589	VBAP	2020000000014000030	ABGRU	U	1					03		
21	VERKBELEG	000000014	377589	VBUK	2020000000014	GBSTK	U	1				C		A	
22	VERKBELEG	000000015	377590	VBAP	2020000000015000010	ABGRU	U	1					03		
23	VERKBELEG	000000015	377590	VBAP	2020000000015000020	ABGRU	U	1					03		
24	VERKBELEG	000000015	377590	VBAP	2020000000015000030	ABGRU	U	1					03		
25	VERKBELEG	000000015	377590	VBUK	2020000000015	GBSTK	U	1				C		A	
26	VERKBELEG	000000016	377591	VBAP	2020000000016000010	ABGRU	U	1					03		
27	VERKBELEG	000000016	377591	VBAP	2020000000016000020	ABGRU	U	1					03		
28	VERKBELEG	000000016	377591	VBAP	2020000000016000030	ABGRU	U	1					03		
29	VERKBELEG	000000016	377591	VBUK	2020000000016	GBSTK	U	1				C		A	
30	VERKBELEG	000000017	377592	VBAP	2020000000017000010	ABGRU	U	1					03		
31	VERKBELEG	000000017	377592	VBAP	2020000000017000020	ABGRU	U	1					03		
32	VERKBELEG	000000017	377592	VBAP	2020000000017000030	ABGRU	U	1					03		
33	VERKBELEG	000000017	377592	VBUK	2020000000017	GBSTK	U	1				C		A	
34	VERKBELEG	000000018	377593	VBAP	2020000000018000010	ABGRU	U	1					03		
35	VERKBELEG	000000018	377593	VBAP	2020000000018000020	ABGRU	U	1					03		
36	VERKBELEG	000000018	377593	VBAP	2020000000018000030	ABGRU	U	1					03		
37	VERKBELEG	000000018	377593	VBUK	2020000000018	GBSTK	U	1				C		A	
38	VERKBELEG	000000019	377594	VBAP	2020000000019000010	ABGRU	U	1					03		
39	VERKBELEG	000000019	377594	VBAP	2020000000019000020	ABGRU	U	1					03		
40	VERKBELEG	000000019	377594	VBAP	2020000000019000030	ABGRU	U	1					03		
41	VERKBELEG	000000019	377594	VBUK	2020000000019	GBSTK	U	1				C		A	

Figure 36. Small part of Saleslog_data_c.xls. Source: own elaboration.

Figure 36 shows a small part of the raw data log.

First step in pre-processing is to select the data which is important, and identify what can be skipped. In this case, only four columns are needed:

- **Object Value:** is needed because it will be the Case ID, since it shows the code of the sale order that it is being changed.
- **Document Number:** it represents the succession of changes that Case ID has been affected. Therefore, it will represent the Timestamp. However, it is not a traditional timestamp, hence, it will be necessary to make some changes in the data.
- **Table Name and Field Name:** represent which table and field are being changed, by combining these two columns it is possible to obtain the activity.

Thus, the filtered data log corresponds to the following figure, Figure 37. It is important to note that this figure, Figure 37, do not represent the final data log it is going to work with. It is just the a filtering of the fields needed.

	A	B	C	D
132	0000000080	367664	KONVC	KBETR
133	0000000080	367664	KONVC	KPEIN
134	0000000080	367664	KONVC	WAERS
135	0000000080	367664	KONVC	KBETR
136	0000000080	367664	KONVC	KPEIN
137	0000000080	367664	KONVC	WAERS
138	0000000080	367664	KONVC	KBETR
139	0000000080	367664	KONVC	KPEIN
140	0000000080	367664	KONVC	WAERS
141	0000000080	367664	VBAK	KEY
142	0000000080	367664	VBAP	KEY
143	0000000080	367664	VBEP	KEY
144	0000000080	367664	VBEP	KEY
145	0000000080	367664	VBPA	KEY
146	0000000080	367664	VBPA	KEY
147	0000000080	367664	VBPA	KEY
148	0000000080	367664	VBPA	KEY
149	0000000112	368024	VBAP	LGORT
150	0000000112	368024	VBKD	BSTKD
151	0000000129	368665	VBAP	LGORT
152	0000000129	368665	VBKD	INCO1
153	0000000129	368665	VBKD	INCO2
154	0000000129	368665	VBPA	KEY
155	0000000129	368665	VBUK	UVALL
156	0000000129	368665	VBUK	UVVLK
157	0000000129	368762	VBAP	LGORT
158	0000000129	368763	VBPA	ABLAD
159	0000000130	368770	VBKD	BSTKD
160	0000000131	368771	VBAP	LPRIO
161	0000000131	368771	VBAP	LPRIO
162	0000000131	368777	KONVC	KBETR
163	0000000131	368777	KONVC	KPEIN
164	0000000131	368777	KONVC	WAERS
165	0000000131	368777	KONVC	KBETR
166	0000000131	368777	KONVC	KPEIN
167	0000000131	368777	KONVC	WAERS
168	0000000131	368777	KONVC	KBETR
169	0000000131	368777	KONVC	KPEIN

Figure 37. Small part of the first filtering applied over Saleslog_data_c.xls. Source: own elaboration.

Now, the useful data is clearer and it is possible to take it to the next step. As it was mentioned, the field in Saleslog_data_C.xlsx called Document Number, represents the Timestamp, one of the three vital properties that a log should contain. Nevertheless, Document Number is showing just a sequence of numbers, it is possible

to turn these number into a logical sequence of events by creating a fictitious Timestamp.

	A	B	C	D
178	0000000131	9-4-34 0:00	KONVC	KPEIN
179	0000000131	10-4-34 0:00	KONVC	WAERS
180	0000000131	11-4-34 0:00	VBAK	KEY
181	0000000131	12-4-34 0:00	VBAP	KEY
182	0000000131	13-4-34 0:00	VBAP	KEY
183	0000000131	14-4-34 0:00	VBEP	KEY
184	0000000131	15-4-34 0:00	VBEP	KEY
185	0000000131	16-4-34 0:00	VBPA	KEY
186	0000000131	17-4-34 0:00	VBPA	KEY
187	0000000131	18-4-34 0:00	VBPA	KEY
188	0000000131	19-4-34 0:00	VBPA	KEY
189	0000000141	20-4-34 0:00	VBAP	LGORT
190	0000000146	21-4-34 0:00	VBAK	VDATU
191	0000000146	22-4-34 0:00	VBKD	FKDAT
192	0000000150	23-4-34 0:00	VBEP	LDDAT
193	0000000150	24-4-34 0:00	VBEP	MBDAT
194	0000000150	25-4-34 0:00	VBEP	TDDAT
195	0000000150	26-4-34 0:00	VBEP	WADAT
196	0000000150	27-4-34 0:00	VBEP	WMENG
197	0000000150	28-4-34 0:00	VBEP	KEY
198	0000000150	29-4-34 0:00	VBUK	GBSTK
199	0000000168	30-4-34 0:00	VBEP	LDDAT
200	0000000168	1-5-34 0:00	VBEP	MBDAT
201	0000000168	2-5-34 0:00	VBEP	TDDAT
202	0000000168	3-5-34 0:00	VBEP	WADAT
203	0000000168	4-5-34 0:00	VBEP	KEY
204	0000000168	5-5-34 0:00	VBAK	VDATU
205	0000000168	6-5-34 0:00	VBKD	FKDAT
206	0000000175	7-5-34 0:00	VBEP	BMENG
207	0000000175	8-5-34 0:00	VBEP	KEY
208	0000000176	9-5-34 0:00	VBAP	KEY
209	0000000176	10-5-34 0:00	VBAP	KEY
210	0000000176	11-5-34 0:00	VBEP	KEY

Figure 38. Selected Timestamp. Source: own elaboration.

The selected Timestamp is a simple succession of days, the main purpose of doing it, is to add the feeling of logical sequence of events to the log.

Next the data log is turned into a .csv file since ProM Lite1.2 use a .csv file as a input to obtain a .xes, in other words, an Event Log.

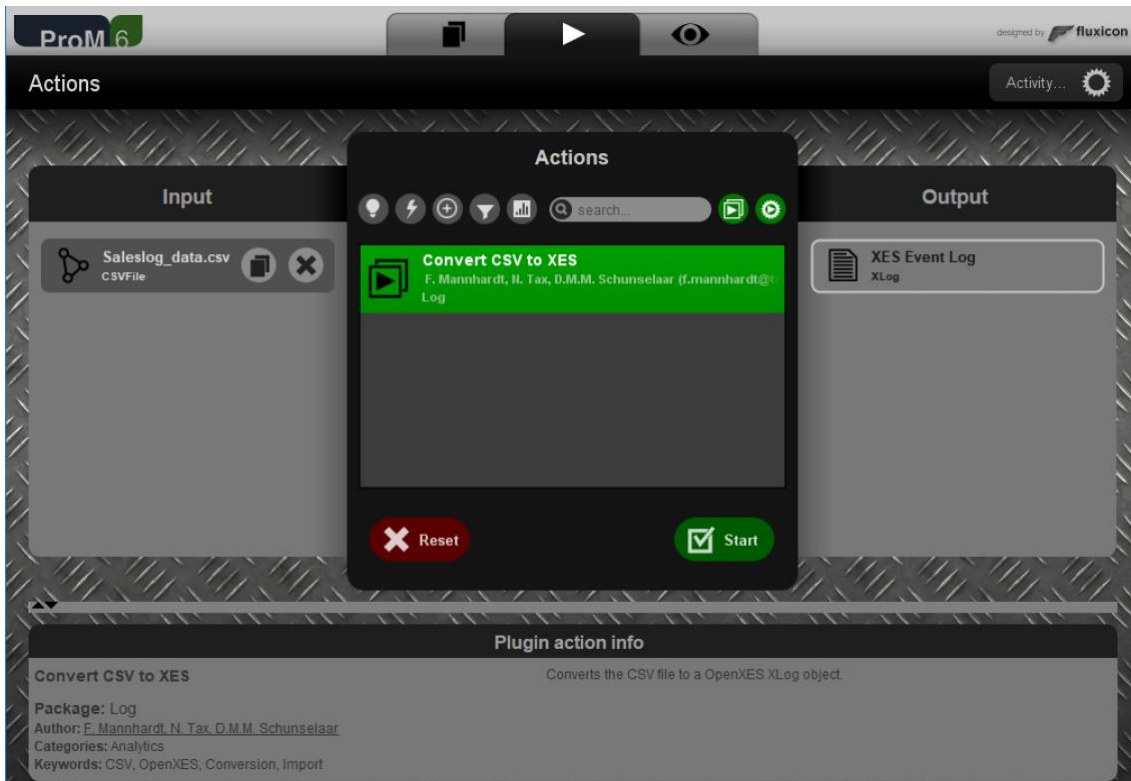


Figure 39. ProM Lite 1.2 Plug-in for turning *Saleslog_data_C.csv* into *Saleslog_data_C.xes*.

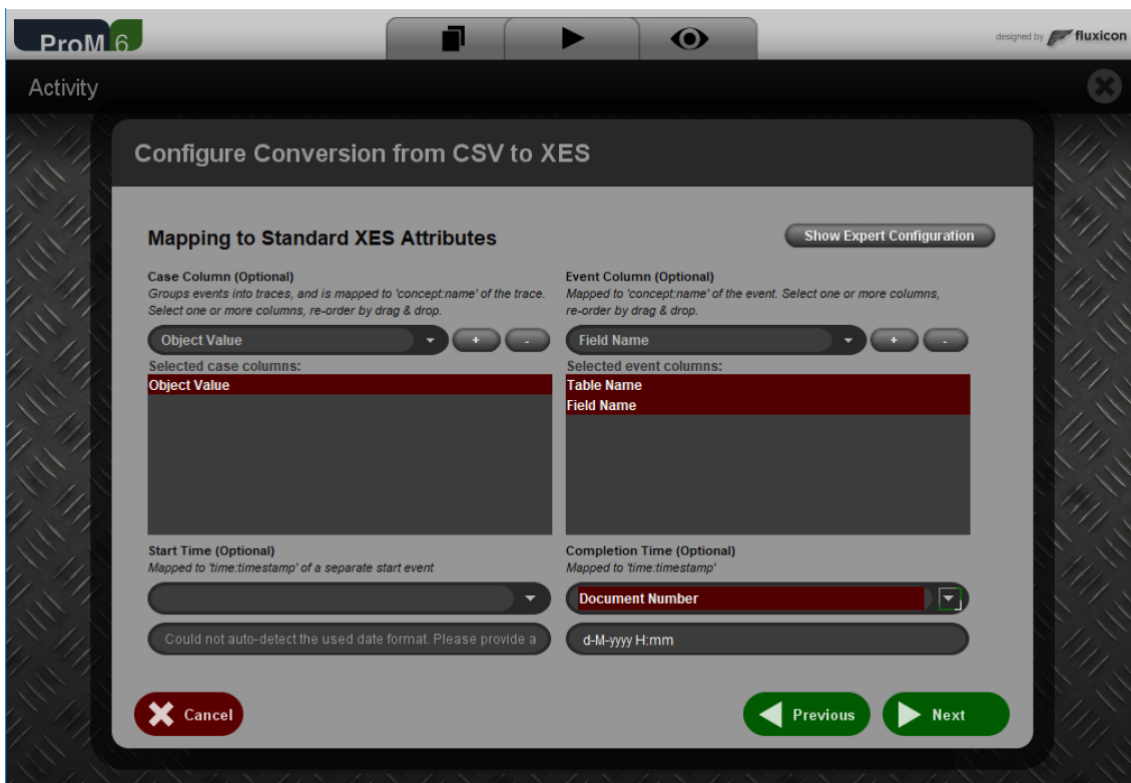


Figure 40. Display of configuration of the *Saleslog_data_C* Event Log parameters.

On Figure 40. It is shown the configuration of the three main parameters:

- **Case ID:** Object Value.

- **Activities/ Events:** Table Name & Field Name.
- **Timestamp:** Document Number, format d-M-yyyy H:mm.



Figure 41. Saleslog_data_C Event Log made by Prom Lite1.2

Figure 41, shows the main information of the Event Log.

- Events: 1654 events have happened in the Event Log.
- Event classes: It represents the different types of activities that there are in the Event Log, in this case, 64 different activities.

Class	Occurrences (absolute)	Occurrences (relative)
KONVC KBETR	185	11,19%
VBPA KEY	177	10,70%
KONVC KPEIN	172	10,40%
KONVC WAERS	172	10,40%
VBEP KEY	122	7,38%
VBAP ABGRU	96	5,80%
VBUK GBSTK	68	4,11%
VBAP KEY	64	3,87%
VBEP WADAT	44	2,66%
VBEP LDDAT	44	2,66%
VBEP TDDAT	44	2,66%
VBEP MBDAT	44	2,66%
VBAK KEY	38	2,30%
VBEP BMENG	26	1,57%

VBAP CMPRE_FLT	26	1,57%
VBAP CMPRE	25	1,51%
VBAP NETPR	23	1,39%
KONVC KEY	23	1,39%
VBAP LGORT	23	1,39%
VBEP EDATU	21	1,27%
VBKD FKDAT	19	1,15%
VBAK VDATU	18	1,09%
VBEP WMENG	18	1,09%
VBAP MPROK	17	1,03%
VBKD BSTDK	15	0,91%
VBAK BSTDK	14	0,85%
VBUK UVALL	10	0,61%
VBKD INCO1	9	0,54%
VBKD BSTKD	9	0,54%
VBUK UVVLS	9	0,54%
VBKD INCO2	6	0,36%
VBAP LPRI0	6	0,36%
VBUK UVVLK	5	0,30%
VBAP ARKTX	5	0,30%
VBAP KPEIN	4	0,24%
VBAP MATNR	4	0,24%
VBKD PRSDT	4	0,24%
VBAK KUNNR	3	0,18%
VBAP TAXM1	3	0,18%
VBAP ROUTE	3	0,18%
VBKD ZTERM	3	0,18%
VBKD KONDA	2	0,12%
VBAP VSTEL	2	0,12%
VBKD KURSK_DAT	2	0,12%
VBKD KURRF_DAT	2	0,12%
VBKD BZIRK	2	0,12%
VBAP BEDAE	2	0,12%
VBAP MTVFP	2	0,12%
VBKD KTGRD	2	0,12%
VBAK AUGRU	2	0,12%
VBKD FBUDA	2	0,12%
VBAP GSBER	1	0,06%
VBAK AUDAT	1	0,06%
VBAP VRKME	1	0,06%
VBAP KALNR	1	0,06%
VBKD BSTDK_E	1	0,06%
VBAP UMVKN	1	0,06%
VBAP UMVKZ	1	0,06%
VBAP WERKS	1	0,06%
VBAK AUTLF	1	0,06%
VBAP KMEIN	1	0,06%

VBAP PSTYV	1	0,06%
VBAK BNDDT	1	0,06%
VBPA ABLAD	1	0,06%

Table 36. Set of activities of the Event Log. Source: own elaboration.

- Cases: It represents the different traces, groups of events, founded in the Event Log.

Next figure, Figure 42, shows better the different traces, as well as the useful information.

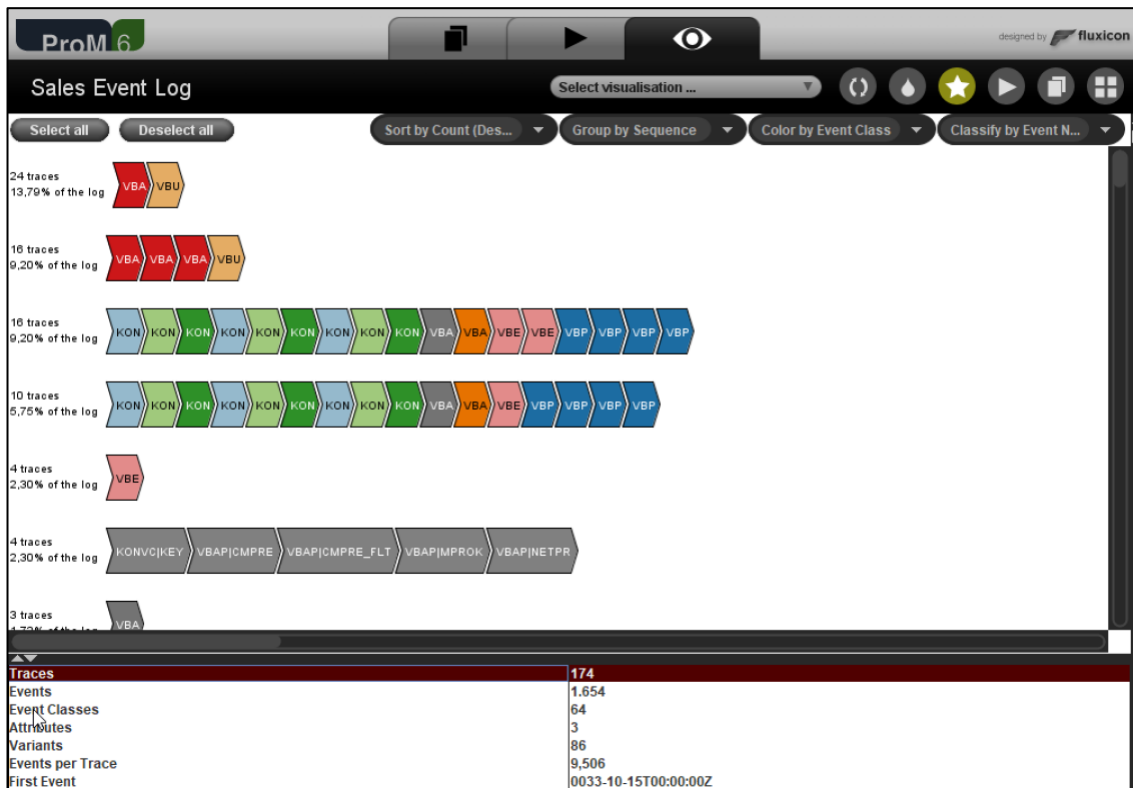


Figure 42. Part of the events grouped by traces.

After showing the information provided by the Event Log. It is possible to start with the process discovery section.

6.3. Application of Process Mining techniques.

In this section, it keeps working with the Event Log produced by Prom Lite 1.2, and it will be tried to discover a business process.

First of all the Sales Event Log is processed through the Alpha miner Plug-in, in order to show what would be the process discovered by it.

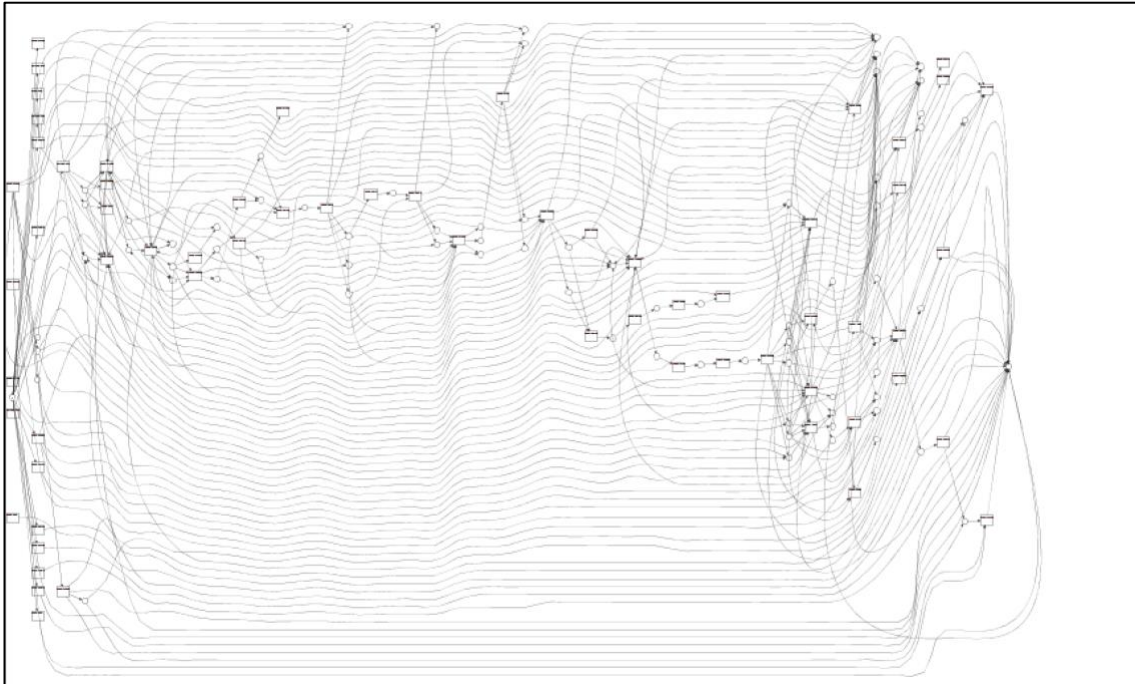


Figure 43. Saleslog_data_C Process model discovered by alpha algorithm.

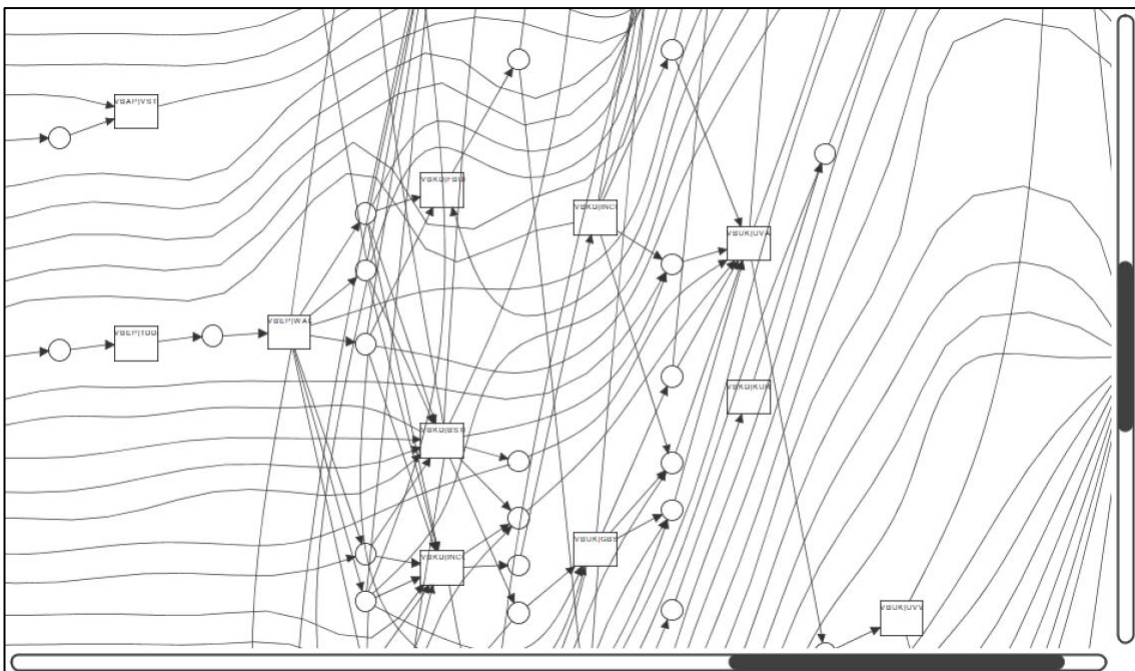


Figure 44. Saleslog_data_C Process model zoom in.

It is easy to see, that the model discovered by alpha algorithm is completely incomprehensible, it correspond to the so called “spaghetti process”. A “spaghetti process” is the name for the unstructured processes. Given that data set, it is not surprising that this kind of process has been discovered. The 1654 instances did not form a homogeneous group since it included many different activities and traces.

Now, it is going to apply more advanced techniques, and the results will be shown:

- Alpha+ algorithm:

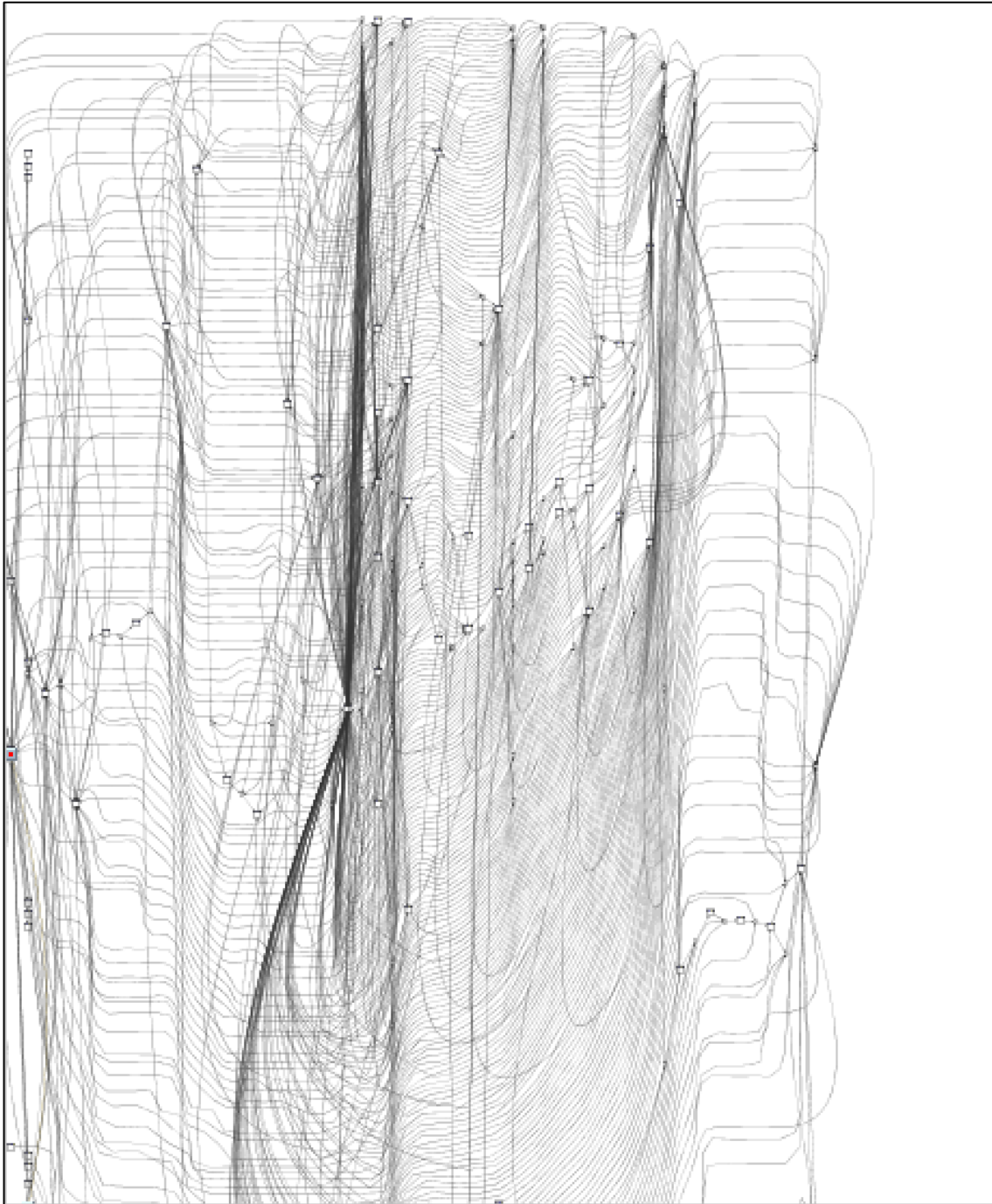


Figure 45. Alpha+ algorithm Saleslog_data_C process discovered.

Despite it is more advance discovery algorithm, Alpha+ algorithm discovers even more unclear process.

- Inductive miner:



Figure 46. Inductive miner Saleslog_data_C process discovered.

Note that the black box correspond to silent activities, introduced in Alignments approaching section.

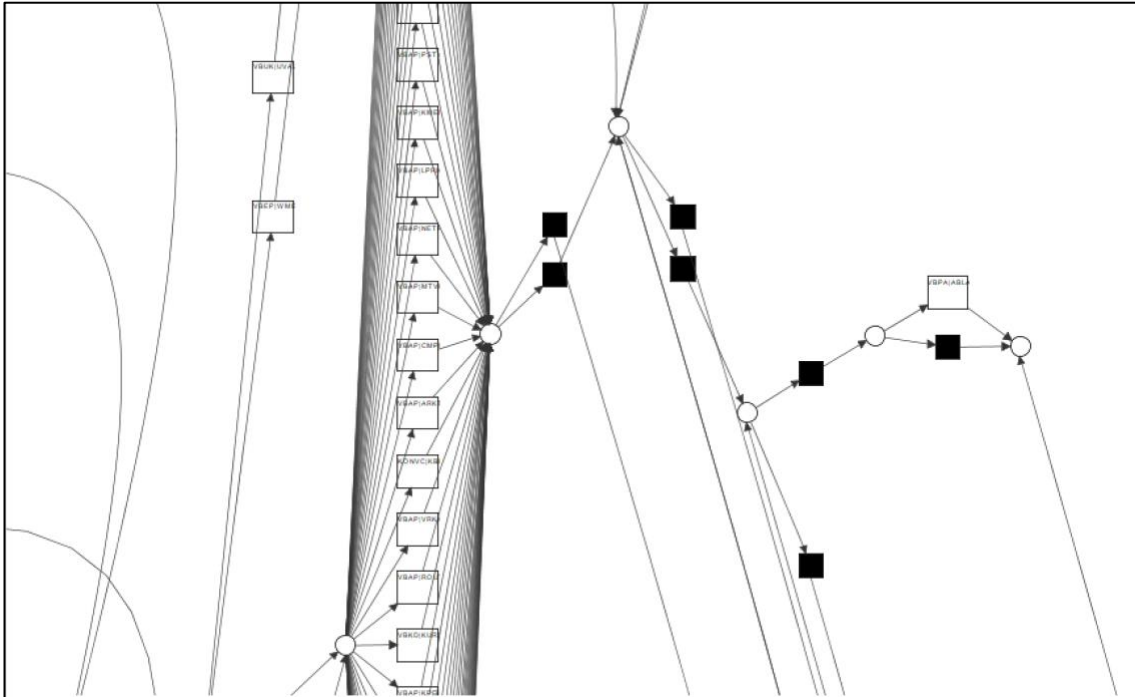


Figure 47. Inductive miner Saleslog_data_C process discovered, zoom in.

This process is quite clearer than the discovered by alpha and alpha+ algorithms. Moreover, it is more interesting from the point of view of uncovering a process. The process might be split in three different parts:

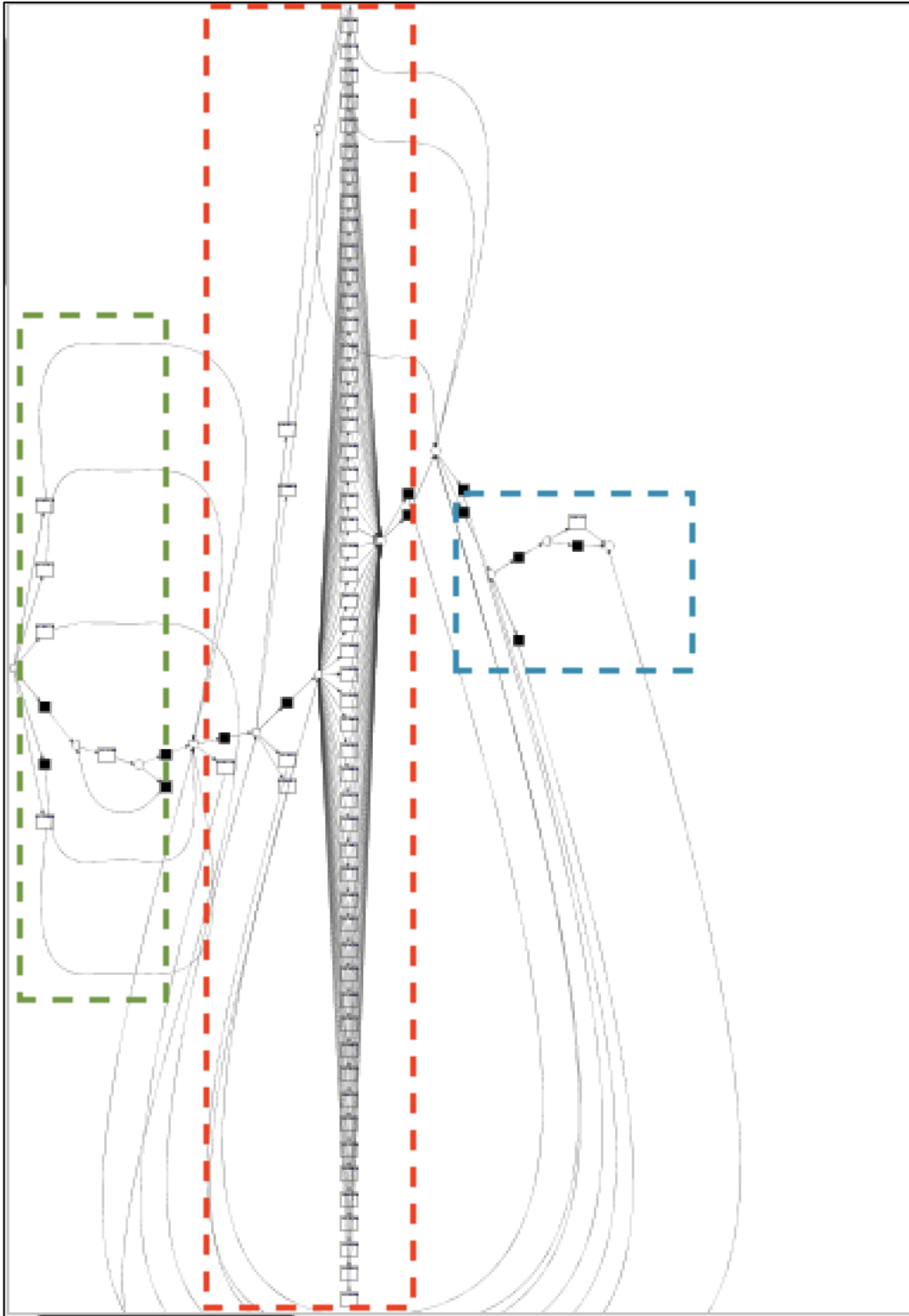


Figure 48. The three parts of the process: Green: Starting. Red: middle activities. Blue: Ending. Source: own elaboration.

Figure 48 shows the three parts of the model:

- Starting boxed in green.
- Middle activities in red.
- Ending activities in blue.

Whether the starting and ending are well defined, the middle activities suppose a complex set, which do not provide any type of information. Furthermore, by analysing deeper the Event log, many activities, among which starting activities are, just appear once or twice. If the activities with a frequency below 5 times (0,24% over the whole log), are considered as outliers, It is possible to filter the log erasing those activities. The purpose of doing it, it is to find an Event Log which will not be affected by noise, making easier the discovery of a possible feasible process.

Set of activities will be erased:

Class	Occurrences (absolute)	Occurrences (relative)
VBAP KPEIN	4	0,24%
VBAP MATNR	4	0,24%
VBKD PRSDT	4	0,24%
VBAK KUNNR	3	0,18%
VBAP TAXM1	3	0,18%
VBAP ROUTE	3	0,18%
VBKD ZTERM	3	0,18%
VBKD KONDA	2	0,12%
VBAP VSTEL	2	0,12%
VBKD KURSK_DAT	2	0,12%
VBKD KURRF_DAT	2	0,12%
VBKD BZIRK	2	0,12%
VBAP BEDAE	2	0,12%
VBAP MTVFP	2	0,12%
VBKD KTGRD	2	0,12%
VBAK AUGRU	2	0,12%
VBKD FBUDA	2	0,12%
VBAP GSBER	1	0,06%
VBAK AUDAT	1	0,06%
VBAP VRKME	1	0,06%
VBAP KALNR	1	0,06%
VBKD BSTDK_E	1	0,06%
VBAP UMVKN	1	0,06%
VBAP UMVKZ	1	0,06%
VBAP WERKS	1	0,06%
VBAK AUTLF	1	0,06%
VBAP KMEIN	1	0,06%
VBAP PSTYV	1	0,06%
VBAK BNDT	1	0,06%
VBPA ABLAD	1	0,06%

Table 37. Set of activities delayed.

Filtering activities that appear below 5 times in the Event Log:



Figure 49. Filtered Saleslog_data_C Event Log

The activities, by erasing the supposed outliers, have been decreased in almost 50% , it represents an interesting fact since now it is much more probable to get a feasible process model.

Process obtained by applying Inductive miner:

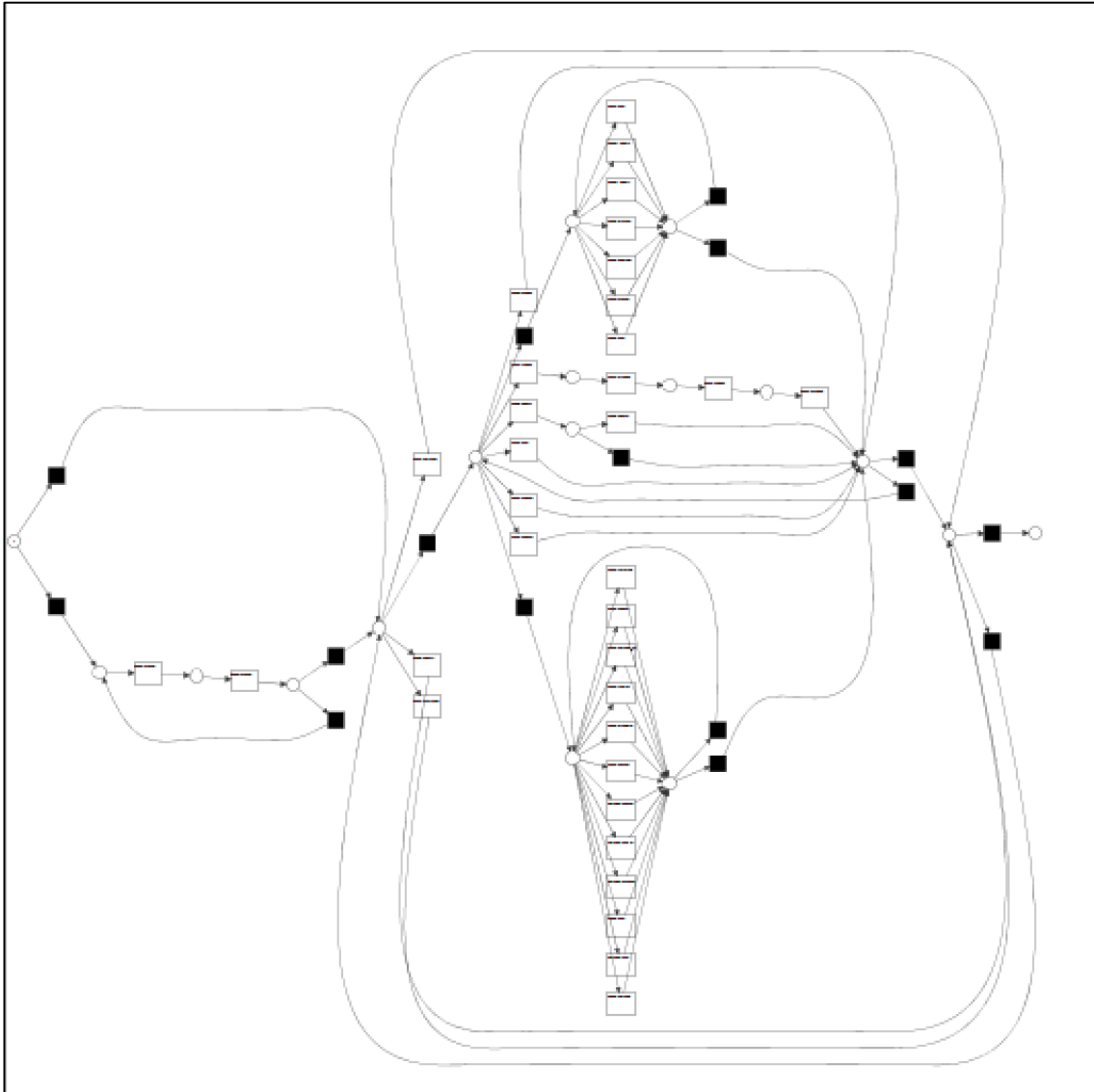


Figure 50 Inductive miner Saleslog_data_C filtered discovered process.

This process is much clearer than the one before applying the filter, and it can be again, split into three different types of activities:

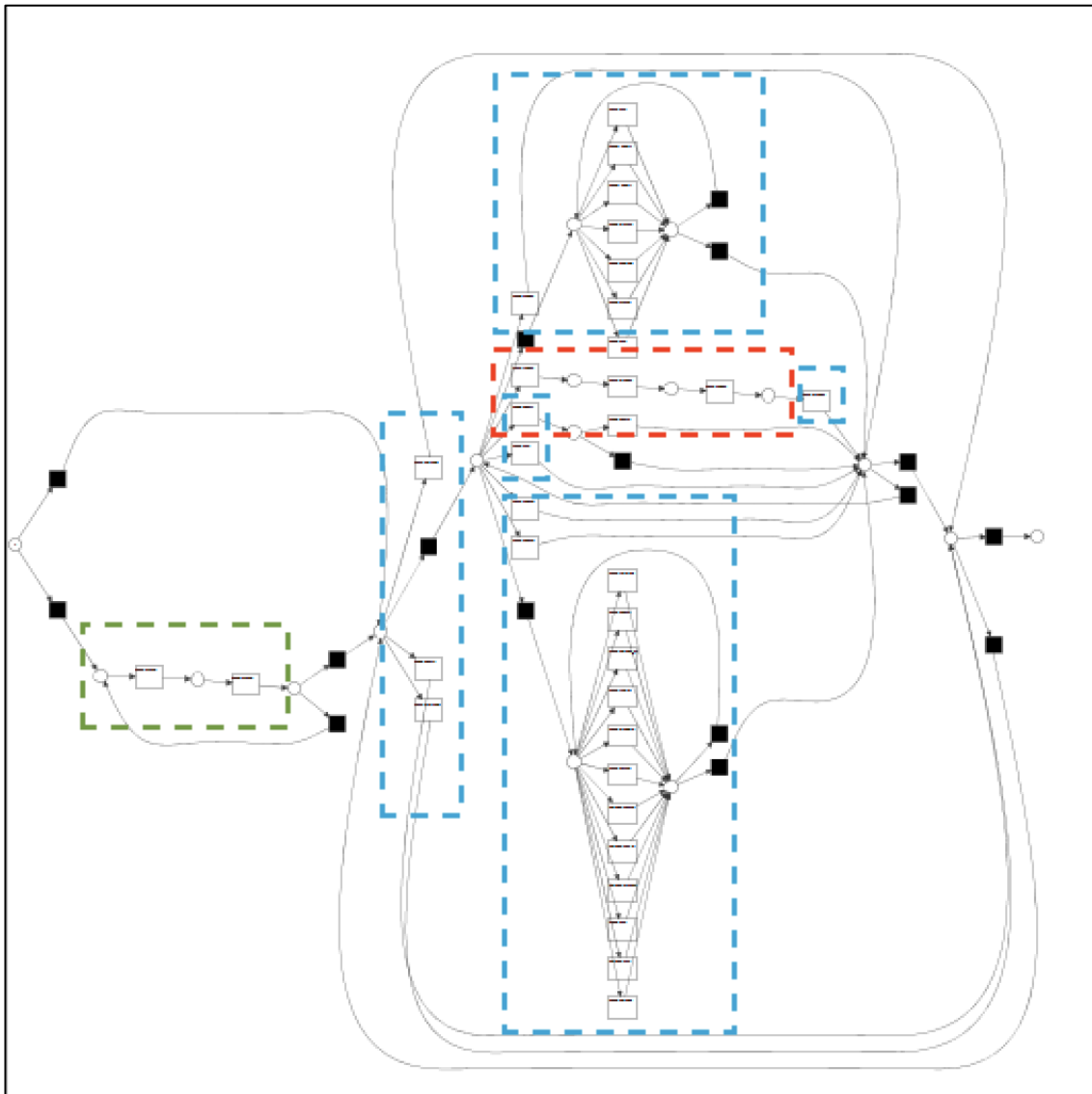


Figure 51. Different types of activities.

Again, the green box contains the starting activities, the red one the middle activities, and the blue one represents the ending ones.

This process model is much clearer, but it still keeps a high rate of complexity. Moreover, whether it is true that there are many less activities, attending to the four quality criteria is not balanced since it is not as simple as possible, and it is over fitted. Nevertheless, this model is the best candidate for testing the fitness between the real behaviour and the modelled one.

After discovering a process that it is supposed to be able to explain the behaviour of the real process, it is moment to check if it is really true.

As was decided on Conformance Checking section, Alignments will be the technique used for checking the conformance. Therefore the first step is select the model and the Event log that will be checked:



Figure 52. Previous step to apply alignments.

The process will be the one after filtering, Figure 50. The Event log checked will be whole Event Log. Therefore, next figure, Figure 52 shows the results of applying alignments:

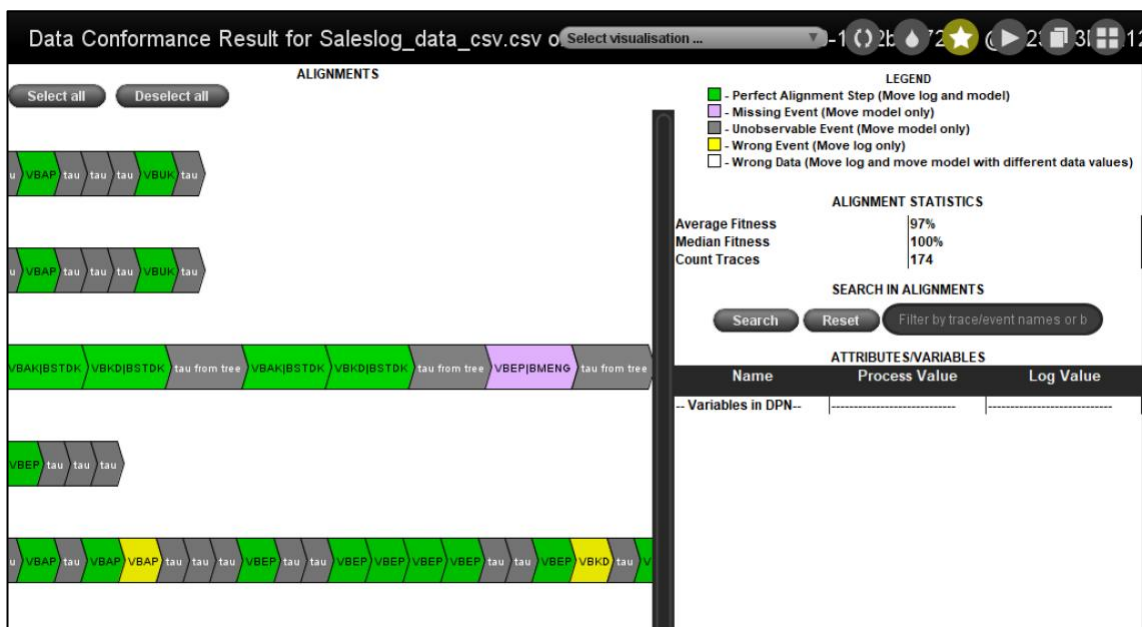


Figure 53. Result of applying alignments ProM Lite 1.2 Plug-in.

The alignments results are quite good, 97% of fitness. However, it is easy to realize that the process uncovered, despite it can explain the behaviour of the log, it just is able to do that with this log. Note that the activities coloured in yellow in Figure 53, are the ones which were erased from the log when the filter was applied, and at the same time are the one which provoke the 3% of unfitnes, since there is just one Missing event, coloured in purple.

6.4. Discussion of the results and recommendations.

The result of the experiment, were not the expected one from the point of view of discovering a process.

It is true, that Inductive Miner Plug-in has been able to order those activities in a process model, however, the discovered process, did not add useful information and knowledge about the data given.

From the result obtained is possible to deduce the following facts:

1. The log has been made in a random way, even though one of the main purposes of Process Mining is to discover hidden processes of event data, it is based on the premise that the data recorded in the Event Log belongs to some type of process. Nevertheless, in this case, it is possible to ensure that the Event Log, hence, the process discovered, they do not follow a logical sequence of events that correspond to any useful process. This fact is based on:
 - a. The log, at the beginning, even had not any timestamp.
 - b. The process selected as feasible, Figure 51, shows that, although there is a structure in the process, most of the activities belong to ending activities (blue box), thus, the algorithm just repeats these activities again and again until the log ends.
2. After checking the conformance, the process discovered fits at 97% it can look like good result, and it would be if the process was balanced in the four process mining quality criteria. Due to the data log structure, the process discovered is a succession of events bounded one another by silent activities, in order to make the transition of events possible. Therefore, the process probably, will not be meaningful for dealing with other event logs or traces.

Recommendations:

- It is recommended to revise the activities in data log, Saleslog_data_C, in order to detect outliers that are affecting to the process discovery.
- It is recommended to revise the whole data log, Saleslog_data_C, with people involved in its implementation, with the purpose of knowing which rules were followed at time to collect and record the data.

In conclusion, the data log, Saleslog_data_C, is useful as good example of how to manage a process mining project, pre-processing, data analysis, discovery, and conformance checking. Nonetheless, the results are not the expected from the point of view of the quality of the process discovered, its structure is complex and hard to understand. The process, is not a valid sales business process and should be revised, in order to detect possible problems in the data given.

7. Conclusion

This thesis, gather and explain the newest concepts in management and data science fields, Business Process Management and Process Mining.

At the beginning of this thesis, three questions were purposed:

- What is Business Process Management?
- What is Process Mining?
- How a Process Mining project is managed?

They have been successfully answered as follows:

Through the exhaustive study of different sources, examples, and points of view, it learns that Business Process Management is not just a set of suites or informatics tools, but it is a complex discipline which, after ten years of development it has become able to manage and improve all levels within the organizations:

- Operational Processes: Applying PDCA cycle in operational processes the continuous improvement is ensured.
- Support Processes: Ensuring the correct performance of operational processes through the definition of responsibilities, and arrangement of proper information to the people who must receive it.
- Management Processes: By the adoption of new roles dedicated to the analysis of the environment with specialized technologies in order to a fast adaptation to the changes and reach the goals of the organization.

If those concepts are fundamental, it is vital as well, the way in which the information is presented in order to ensure the proper understanding of the business process by all the involved stakeholders. That is why it has been introduced the main types of representation of a business process.

Secondly, it is the turn of Process Mining. It is established that Process Mining is a young discipline belonging to Data Science that, however, is the bridge between process management and data science. Since Process Mining allows to discover new business processes, check the conformance between the real behaviour and the process discovered, and enhance the business processes within the companies by reducing cost, improving and ensuring the quality and security through new audit concepts, and in sum, creating a competitive advantage for the organizations that apply this new discipline.

In this thesis are introduced the three different types of process mining:

- Discovery.
- Conformance.
- Enhancement.

Paying special attention to discovery and conformance, which are explained in depth, as well as their main approaches.

Thirdly, it has been conducted an experiment about analysing a real data log from the Hochschule Albstadt-Sigmaringen SAP server, with the purpose of showing how the different stages of a process mining project are managed.

The results, derived from the experiment, were illustrative from the point of view of explaining the different steps in the project. Nevertheless, they did not meet the expectations about discovering a useful business process, anyways, every step on the experiment was explained in detail, as well as the conclusions about it.

Finally, it is proposed, as a future line of research, the explanation and experimentation of the techniques, approaches, and different uses of Enhancement, the last type of Process Mining.

8. Bibliography

- [1] pruebacoaching, "fiverr.com," [Online]. Available: <https://www.fiverr.com/pruebacoaching/make-a-swimelane-flowchart-ie-cross-functional>. [Accessed 25 January 2018].
- [2] E. M. Margaret Rouse, "TechTarget SearcCio," June 2016. [Online]. Available: <http://searchcio.techtarget.com/definition/business-process>. [Accessed 25 January 2018].
- [3] Association of Business Process Management Professionals., BPM CBOOK, 1st ed., ABPMP, 2013, pp. 43-81.
- [4] D. B. Hitpass, BPM: Business Process Management: Fundamentos y Conceptos de Implementación, 4th Edition ed., Santiago de Chile: BHH Ltda. - Santiago de Chile, 2017.
- [5] S. P. A. S. Sandra Lusk, "Evolution of BPM as a Professional Discipline," www.bptrends.com, 2005.
- [6] P. M. Senge, The Fifth Discipline: The Art & Practice of The Learning Organization, 2nd ed., Currency, 2006.
- [7] C. W. A. G. K. N. D. Tim Weilkiens, OCEB 2 Certification Guide: Business Process Management - Fundamental Level, Morgan Kaufmann, 2016, p. 58.
- [8] R. S. Kaplan, "www.processexcellencenetwork.com," 04 06 2010. [Online]. Available: <https://www.processexcellencenetwork.com/lean-six-sigma-business-transformation/articles/lead-and-manage-using-the-balanced-scorecard>. [Accessed 15 11 2017].
- [9] Club-BPM, "www.club-bpm.com," Club-BPM, [Online]. Available: www.club-bpm.com. [Accessed 17 11 2017].
- [10] ABPMP International, "ABPMP International," [Online]. Available: <http://www.abmp.org>. [Accessed 18 11 2017].
- [11] BPM Institute , "BPM Institute," [Online]. Available: www.bpminstitute.org. [Accessed 18 11 2017].
- [12] BPM Center, "BPM Center," [Online]. Available: <http://bpmcenter.org>. [Accessed 18 11 2017].
- [13] "OMG Object Management Group," [Online]. Available: <http://www.omg.org/bpmn/index.htm>. [Accessed 25 January 2018].
- [14] IBM, "http://www-03.ibm.com," [Online]. Available: <http://www-03.ibm.com/software/products/es/business-process-manager-family>. [Accessed 18 11 2017].
- [15] T. Volmering, "https://blogs.sap.com," 6 May 2006. [Online]. Available: <https://blogs.sap.com/2008/05/06/introducing-sap-netweaver-business-process-management-bpm/>. [Accessed 18 November 2017].
- [16] ORACLE, "www.oracle.com," [Online]. Available: <http://www.oracle.com/us/technologies/bpm/suite/overview/index.html>. [Accessed 18 November 2017].

- [17] Aura Portal, "auraportal.com," [Online]. Available: <https://www.auraportal.com/product/>. [Accessed 18 November 2017].
- [18] "Camunda," [Online]. Available: <https://camunda.org>. [Accessed 25 January 2018].
- [19] "Signavio," [Online]. Available: <https://www.signavio.com/products/process-manager/>. [Accessed 25 January 2018].
- [20] Bizagi, "bizagi.com," [Online]. Available: <https://www.bizagi.com/uk/customers/case-studies/manufacturing-retail-adidas>. [Accessed 18 November 2017].
- [21] Aura Portal, "auraportal.com," [Online]. Available: <https://www.auraportal.com/case-studies/case-study-toyota/>. [Accessed 18 November 2017].
- [22] Aris Community by Software AG, "http://www.ariscommunity.com," [Online]. Available: <http://www.ariscommunity.com/users/jork/2008-11-16-review-process-intelligence-roadshow-proactive-process-monitoring-bayer-healthcare>. [Accessed 19 November 2017].
- [23] Drat, "commons.wikimedia.org," [Online]. Available: https://commons.wikimedia.org/wiki/File:Detailed_petri_net.png. [Accessed 7 December 2017].
- [24] ABPMP, BPM CBOOK, 2013.
- [25] uml-diagrams.org, "uml-diagrams.org," [Online]. Available: <https://www.uml-diagrams.org>. [Accessed 27 January 2018].
- [26] uml-diagrams.org, "uml-diagrams.org," [Online]. Available: <https://www.uml-diagrams.org/document-management-uml-activity-diagram-example.html>. [Accessed 27 January 2018].
- [27] IBM, "IBM Knowledge Centre," [Online]. Available: https://www.ibm.com/support/knowledgecenter/en/SS6RBX_11.4.2/com.ibm.s.a.bpr.doc/topics/tovmdlprocidef3.html. [Accessed 28 January 2018].
- [28] LucidChart, "LucidChart," [Online]. Available: <https://www.lucidchart.com/pages/es/s%C3%ADmbolos-bpmn-explicados>. [Accessed 29 January 2018].
- [29] Object Management Group Business Process Model and Notation, "OMG BPMN," [Online]. Available: <http://www.bpmn.org>. [Accessed 29 January 2008].
- [30] R. Baureis, "Aris Community by Software AG," 22 March 2010. [Online]. Available: <http://www.ariscommunity.com/users/rbaureis/2010-03-22-basic-rules-epc-modelling>. [Accessed 26 January 2018].
- [31] M. Tay, "blog.maxconsilium.com," 18 September 2013. [Online]. Available: <http://blog.maxconsilium.com/2013/09/process-notation-p1.html>. [Accessed 26 January 2018].
- [32] W. M. P. v. d. Aalst, "researchgate.com," [Online]. Available: https://www.researchgate.net/figure/304550918_fig1_Figure-1-Positioning-of-the-Three-Main-Types-of-Process-Mining-Wil-M-P-van-der-Aalst. [Accessed 3 December 2017].
- [33] W. v. d. Aalst, Process Mining Data Science in Action, Berlin: Springer, 2016.

- [34] M. Miyazaki, "FlyData," 3 November 2015. [Online]. Available: <https://www.flydata.com/blog/a-brief-history-of-data-analysis/>. [Accessed 29 January 2018].
- [35] D. R. Muñoz, Manual de estadística, pp. 4-8.
- [36] S. G. Cabria, "Origen y desarrollo de la estadística en los S. XVII y S. XVIII," in *Estadística Española*, 1982, pp. 7-28.
- [37] W. v. d. A. e. al, "win.tue.nl," September 28 2011. [Online]. Available: <http://www.win.tue.nl/ieeetfpm/downloads/Process%20Mining%20Manifesto.pdf>. [Accessed 29 January 2018].
- [38] process-mining.biz, "http://www.process-mining.biz," [Online]. Available: http://www.process-mining.biz/?page_id=90. [Accessed 30 November 2017].
- [39] SIGNAVIO, "docs.signavio.com," [Online]. Available: https://docs.signavio.com/userguide/editor/en/modeling_and_notations/bpmn/responsibility_assignment_raci.html. [Accessed 1 December 2017].
- [40] W. v. d. Aalst, "www.processmining.org," Eindhoven university of technology , [Online]. Available: http://www.processmining.org/_media/processminingbook/process_mining_chapter_07_conformance_checking.pdf. [Accessed 3 December 2017].
- [41] A. A. a. B. v. D. Wil van der Aalst, "WIREs Data Mining Knowl Discov," pp. 182-183, February 2012.
- [42] "ML wiki," [Online]. Available: http://mlwiki.org/index.php/Alpha_Algorithm. [Accessed 8 December 2017].
- [43] C. W. G. a. W. M. v. d. Aalst, "Fuzzy Mining – Adaptive Process Simplification Based on Multi-Perspective Metrics," Springer, Berlin, 2007.
- [44] W. M. P. v. d. Aalst, Artist, [Art]. 2011.