U.PORTO

MASTER IN MODELING, DATA ANALYSIS AND DECISION SUPPORT SYSTEM IN OPTIMIZATION

APPLICATION OF PROCESS MINING TECNIQUES TO INVOICE PROCESS

Susana Isabel Do Nascimento Santos

M

2022

U.PORTO

FEP FACULDADE DE ECONOMIA
UNIVERSIDADE DO PORTO

APPLICATION OF PROCESS MINING TECNIQUES TO INVOICE PROCESS

Susana Isabel Do Nascimento Santos

Internship report

Master in Modeling, Data Analysis and Decision Support System in Optimization.

Supervised by
Prof. Maria Eduarda Silva

2022

Acknowledgments

Thank you to Susana Silva from kpitlean – Continuous Improvement Department from kpBuS Porto, who welcomed me fantastically into the team. Thanks for the patience, enthusiasm, and conditions offered to complete the telecommuting internship.
Thank you to Professora Maria Eduarda for her guidance and support in this project.
My gratitude to my family for their unconditional support. This report was a good company and distraction during the particular time of the pandemic.

Abstract

Process Mining is one of the newest and most sought-after fields of data mining in business process analysis, used to increase quality, reduce costs, and identify process risks.

Process mining allows us to find large and complicated process flows and analyze deviations, bottlenecks, and anomalies from historical data, called event logs. The techniques include process discovery (finding models from an event log), compliance checking (studying deviations), and enhancement (improving the model). Furthermore, it can be studied from three perspectives, control-flow (mapping), performance (timing), and organization (social).

This report addresses the analysis and process mining of a multinational packaging solutions company's shared service center's purchase-to-pay system. Four possible processed invoice flows were implemented using the available data. Moreover, some key performance indicators regarding duplicated and automated rates are defined and measured. For implementation, an R package - bupaR was used as a process mining software tool and Tableau as a business intelligence tool.

Keywords: Business Process Analysis, Process Mining, Accounts Payable, bupaR, Tableau, invoice process, purchase to pay, online analytical processing.

Table of Contents

List of Figures

Introduction

There is a massive amount of data generated at every moment. Consequently, it is said that we are in the era of big data, of incredible amounts of event data, data that is being recorded. This big data brings challenges of volume, velocity, variety, and veracity (the four Vs of big data). So, today's challenge is not to generate more data but to transform this data into real value (Aalst W. v., 2019).

Companies use information systems to handle multiple processes, often using one or more software. An example of this software is Enterprise Resource Planning (ERP), such as SAP and ORACLE, which is used in business transactions such as purchasing, paying, sales or manufacturing.

The recorded process data is often called a transaction log, audit trial, history, or event log. If the data have some associated activity or task, and even a person associated with the steps of each instant of the process, it is possible to apply process mining techniques to discover the actual process flow. It is possible to check the conformity of recorded data with expected behavior and to detect deviating cases. Performance metrics can be applied, unusually long cases can be detected, and bottlenecks identified. Process mining can improve elapsed time, prevent potential risks and fraud, and apply predictive techniques.

Process mining, or business process intelligence, is bridging the gap between classical process model analysis and data mining. Data mining is interested in isolated decisions or low-level patterns and does not look at the end-to-end process, and in process model analysis, data is ignored. Nevertheless, data mining and process mining tend to be combined to answer very advanced questions (Aalst W. v., Process Mining, Data Science in Action, 2016).

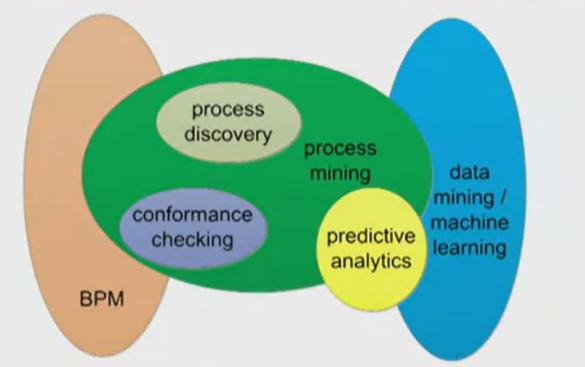


Figure 1: Diagram of the relationship between process mining, data mining, and business process management (BPM).

Many other authors, such as (Grossman & Rinderle-Ma, 2015), place process mining in the business intelligence theme as one more tool for data analytics along with Data Mining and Online Analytical Processing (OLAP), Figure 2.

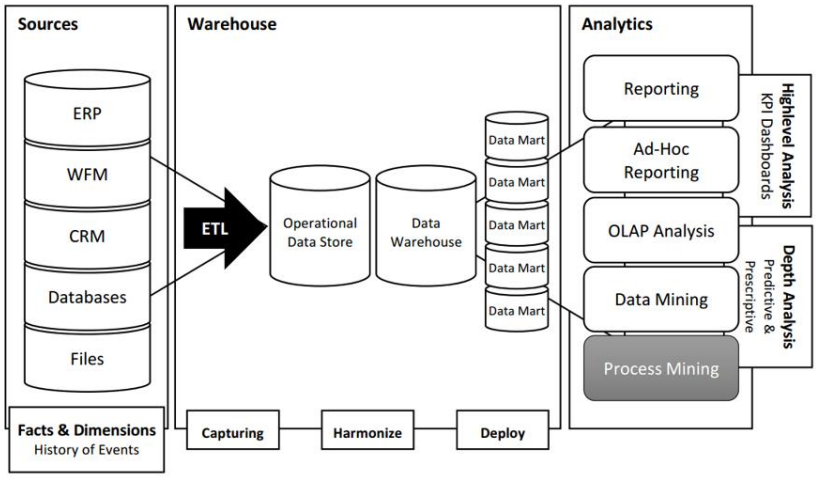


Figure 2: Business Intelligence Framework.

This report addresses the data extracted from an IT software application (Basware) regarding a purchase-to-pay (P2P) process used by the Accounts Payable (AP) team to ensure that orders for goods or services provided by suppliers to the company were paid on time. The company is Klöckner Pentaplast, a global leader in packaging solutions under the Finance kp Shared Service Center in Porto.

The P2P process is crucial to the organization's value chain because it can be subject to business risks, such as extended delivery times, decreasing efficiency of production or increasing costs, or the risk of potential fraud. Analyzing the process provides insight into improvement potentials and prevention of potential risks.

This report consists of two main parts, one with the essential process mining theory and some business process definitions under the AP department's responsibility; the second part is the application of the techniques to data, with the respective analyses and suggestions for the problems found in the process.

The first part is based on the book (Aalst W. v., 2016), and in the second part on the report (Kiarash Diba, 2019), the BPI Challenge winner, which is a challenge held every year and open to all to apply process mining techniques on real data.

Tableau Prep was chosen to extract, load, and transform the data (ELT software), Tableau Desktop as a Business Intelligence tool to visualize and analyze the data, and the bupaR package from R was used to apply the process mining techniques.

Part I – The Theory

1. Process Models

1.1. Introduction

It is essential to understand and have some notions of process models to understand some process mining techniques better. In this chapter, it is presented the essentials of the process model concepts. The book (Aalst W. v., 2016) describes several commonly used models: transition systems, Petri nets, Business Process Modeling and Notation (BPMN), C-nets, EPCs, YAWL, and process trees. In this report, only two are described: Petri nets, the oldest and best-researched process modeling language, and BPMN, which has become one of the most widely used languages for modeling business processes.

Process models help to manage complexity by providing information and documenting procedures. Some errors create a theoretical model, such as the model describing an idealized version of reality or the inability to capture human behavior adequately. An inadequate model can lead to wrong conclusions, and process mining allows fact-based design models, can show that all kinds of inefficiencies occur, and can also visualize the remarkable flexibility of some workers to deal with varying problems and workloads.

Figure 3 shows a process model expressed in a Petri net, and Figure 4 shows the same process in BPMN. The model describes the handling of a request for compensation within an airline. Customers may request compensation for various reasons, such as a delayed or canceled flight.

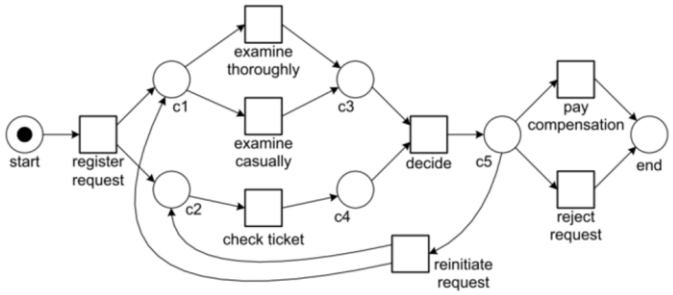


Figure 3: A Petri net modeling the handling of compensation requests.

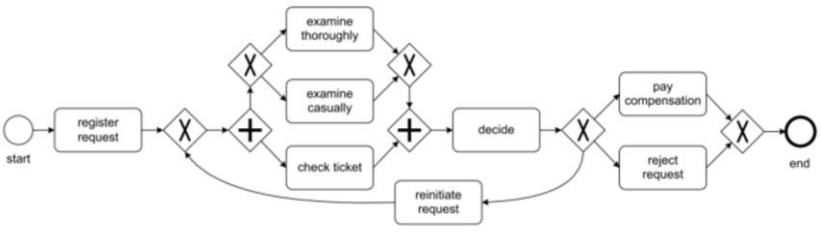


Figure 4: The same process modeled in terms of BPMN.

Many process mining algorithms use Petri nets during discovery or conformance checking. However, this does not imply that the end-user needs to see any Petri nets. Tools such as ProM can convert Petri nets to other notations such as BPMN and vice versa.

1.2. Petri Nets

Petri nets are the oldest process modeling language. It is simple and intuitive, allowing the application of various techniques and analyses. One of its best features is to allow the modeling of concurrency.

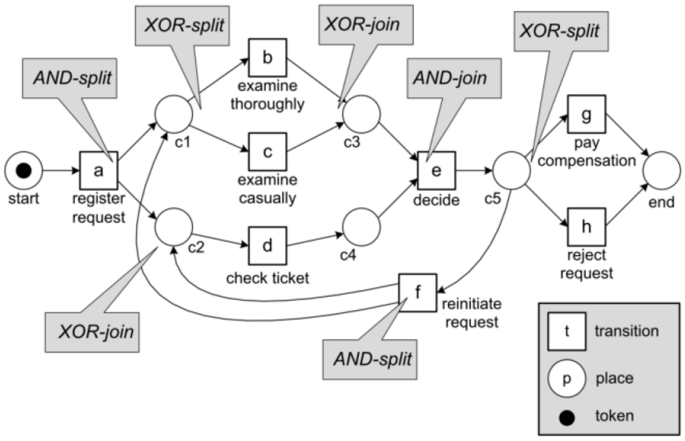


Figure 5: A marked Petri net.

Elements named tokens can flow through the structure, where their distribution across the places determines the state of a Petri net and is referred to as its marking. In the initial marking shown in Figure 5, there is only one token, and 'start' is the only marked place.

A Petri net is a static graph consisting of places and transitions, but tokens move according to rules:

- Places: are represented by circles, and every Petri net has one represented by the beginning and one represented by the end of the process. Token goes through the structure depending on whether there is an option or/and. Whether there is a parallel union or join, for the option 'and' and if it is a parallel split, one transition will be connected with multiple output places. The token will slip in the number of available transitions. If it is a parallel join, multiple input places are connected to a single transition, and every token of the multiple places will be in just one. For the option 'or,' whoever fires first consumes the token, disabling the other transition, and the join must be synchronizing.
- Transition: are represented by squares and only distribute tokens in the Petri net. The transitions move tokens from the input places to the output places connected and only fire if all the input places have a token to consume.

1.3. Business Process Modeling Notation (BPMN)

Recently, the Business Process Modeling Notation (BPMN) has become one of the most widely used languages to model business processes.

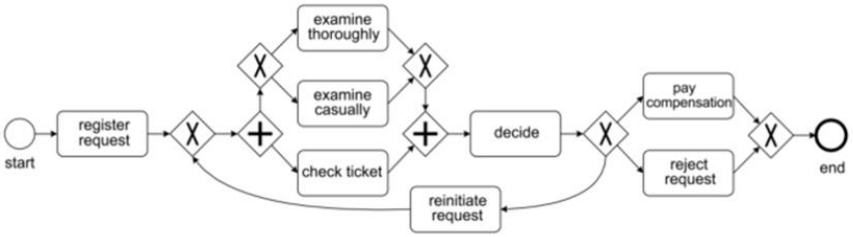


Figure 6: Process model using the BPMN notation.

Like the Petri net model, there is a beginning and an end. The activities can be carried out in parallel, and a rhombus with an X represents a choice, and only one of the activities can be chosen. Figure 6 shows that there are split and join gateways of different types: AND, XOR, OR. The splits are based on data conditions. The notation may become more complicated according to the process model, Figure 7. Process discovery algorithms use Petri nets, but BPMN is used to present the results to organizations. Both models describe the same behavior.

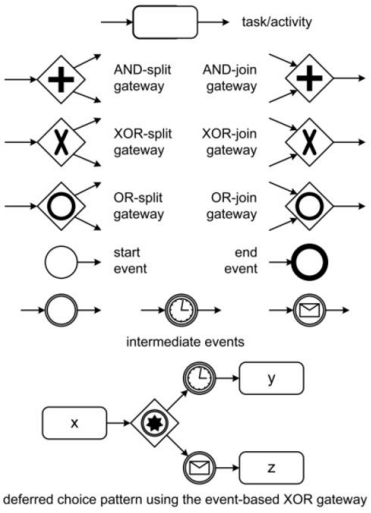


Figure 7: BPMN notation.

2. Process Mining

2.1. Introduction

The increase in data availability gave rise to process mining, which focuses on extracting insights about processes from event logs, that is, data with the historical record. However, it is not a trivial task since business processes are usually very complex, sometimes, event logs contain only a sample of all possible cases, or errors and inconsistencies create additional difficulties. (GertJanssenswillen, 2021)

Process mining has three types for conducting the event logs: process discovery, conformance checking, and enhancement. First, process discovery takes an event log and produces a model using an algorithm without any other information about the process. The second type is conformance, where the model discovered is compared to an existing process and is used to check if reality, as recorded in the log, is the same as the theoretical. Finally, enhancement is to improve an existing process model using information about the actual process recorded, modifying the model to reflect reality better.

Alongside, the relationship between a process model and 'reality' captured in the form of an event log can be reflected by the terms Play-in, Play-out, and Replay (Aalst W. v., 2019):

- Play-in, the goal is to construct a model, learning a process model without modeling from raw event data, which corresponds to process discovery.
- Play-out is about the classical use of process models not involving any event data, which can be used for analyzing business processes.
- Replay uses an event log and a process model as input. The event log is replayed on top of the process model to perform conformance checking, allowing to see discrepancies between the log and the model, investigate performance problems, and construct predictive models by replaying many cases and learning the expected time until an activity completion.

In addition, it is possible to study the model from several perspectives, including the control flow perspective ('How with case flows?'), the organizational perspective ('Who is involved?'), and the performance perspective ('When it occurs?').

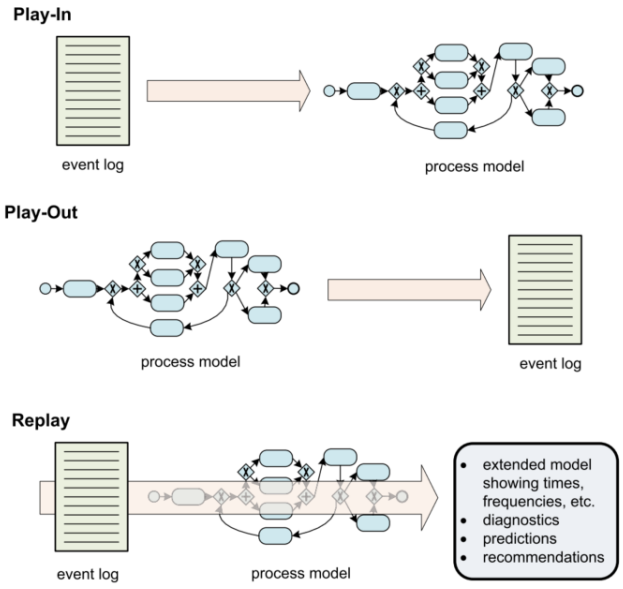


Figure 8: There are three ways of relating event logs and process models.

Van der Aalst (Aalst W. v., Process Mining, Data Science in Action, 2016), a professor with extensive research in this area, presents an approach of five steps for an approach to applying process mining techniques, Figure 9:

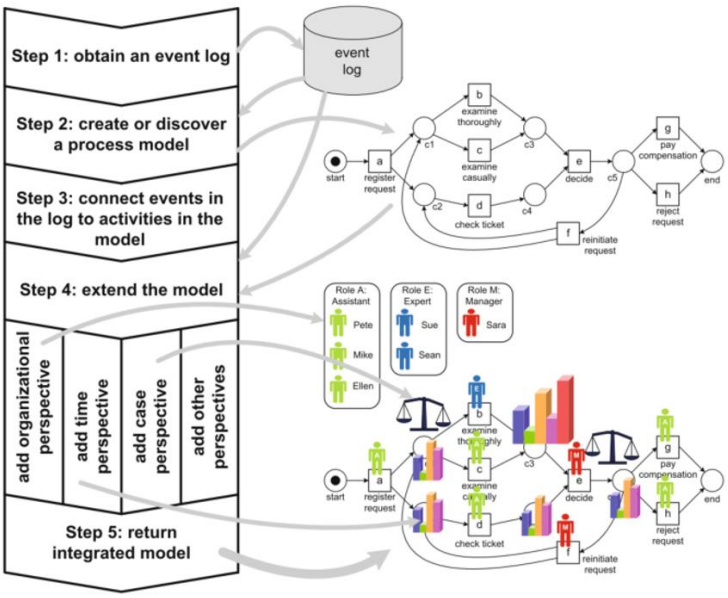


Figure 9: Approach to cover the organizational, time, and case perspectives.

- Step 1: Obtain a record of events by extracting data from various systems and exploring and filtering the process.
- Step 2: discover a process model using process mining techniques.
- Step 3: link events in the log to activities in the model.
- Step 4: extend the model, add the organizational perspective (identify events in the log entities that link activities to groups of resources), add the time perspective (timestamps and frequencies can be used to learn probability distributions), add the case perspective and add other perspectives (information about risks and costs can be added to the model).
- Step 5: return the integrated model.

There are several processes where process mining can be applied, and it is possible to find them in any organization, whether a company, a school, a hospital, or a finance department. The processes can be seen as 'maps' describing the operation of the organization and can be applied alongside new approaches, for example:

- Online process - can be used to explore processes in real-time using visualization and forecasting;
- Big data - helping with conventional approaches to handling data that is too large and complex;
- Cubes and OLAP (Online Analytical Processing) – analyzing organized events using different dimensions, for example, case types, regions, sub-processes, departments, and time windows.

2.2. Event log

Data can be extracted from various data sources, such as databases, flat files, message logs, transaction logs, and document management systems. The data is probably unstructured and scattered across different data sources, and it often takes quite a lot of effort to collect the relevant information. It helps that extraction is driven by questions rather than exploring all data availability.

The mining process results likely trigger new questions, which may lead to the exploration of new data sources and more detailed data extractions. Often several iterations of extraction and filtering are required. However, there is a minimum amount of information contained in the data to be able to apply process mining.

An event log is a set of historical, or time-stamped, data with uniquely identified activities and cases. It is necessary to clean and transform the data so it can be analyzed as event records, but it must have a temporal record, and it is crucial to understand the 'cases' or 'activities'. An event log must have at least six different pieces of the required information (bupaR, 2019):

- A timestamp, a time date
- A case identifier, is the s.ubject of the process, e.g., a customer, an order, a patient.
- An activity label is a step in the process, e.g., receiving an order or sending a payment.
- An activity instance identifier, is the execution of a specific step for a specific case.
- A transactional life cycle stage, typical values *start* and *complete*. Other possible values are *schedule*, *suspend*, or *resume*.
- A resource identifier, can be a person or a machine associated with the activity.

An example can be seen in Figure 10, and this is an extract event log containing nine events. Each event is linked to a single timestamp. As there can be more events within a single activity, each event must have a life cycle state (in the example is the status). In addition, an activity instance identifier needs to indicate which events belong to the same activity instances.

patient	activity	timestamp	status	activity_instance	resource
John Doe	check-in	2017-05-10 08:33:26	complete	1	Samantha
John Doe	surgery	2017-05-10 08:38:21	schedule	2	Danny
John Doe	surgery	2017-05-10 08:53:16	start	2	Richard
John Doe	surgery	2017-05-10 09:25:19	complete	2	Richard
John Doe	treatment	2017-05-10 10:01:25	start	3	Danny
John Doe	treatment	2017-05-10 10:35:18	complete	3	Danny
John Doe	surgery	2017-05-10 10:41:35	start	4	William
John Doe	surgery	2017-05-10 11:05:56	complete	4	William
John Doe	check-out	2017-05-11 14:52:36	complete	5	Samantha

Figure 10: Example of an event log containing nine events.

Figure 11 shows all types of transactions, their enablement, and their effect. For example, according to the transactional life cycle model, "abort_activity" is only possible when the activity instance is running (i.e., started, suspended, or readded).

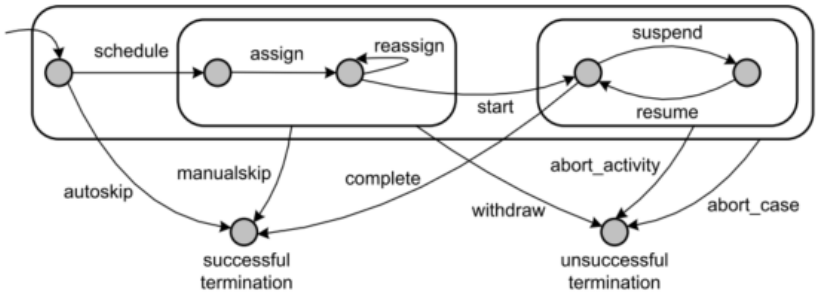


Figure 11: All transaction types.

There are some challenges in extracting event logs. Here are some of them:
- Events are grouped by cases. For this, the events need to be related to each other.
- Events need to be sorted by the case (in order of occurrence).
- Event logs can only provide a snapshot of a longer process. When the average duration of a case is short compared to the duration of the recording, it is better to remove incomplete cases.
- Company information systems may have thousands of tables with company-relevant data. Knowledge is needed to locate the required data; it depends on the available data and the questions that need to be answered.
- In many applications, the events in the event log are at a different level of granularity than the activities relevant to end users.

Finally, data quality is of utmost importance for the success of the mining process. If event data is missing or cannot be trusted, then the results of process prospecting are less valuable. In the book (Aalst W. v., 2016), it is possible to find several warnings for the most varied types of errors that can compromise the data quality.

2.3. Process Discovery

Process discovery is the leading and most studied process mining task. The biggest challenge is to find an algorithm that can discover a process model that is as representative as possible of the behavior of the event log. Petri nets are widely used to study algorithm performance because they are simple and graphical while allowing the modeling of concurrency, choices, and iteration.

For the model found to be representative of the actual behavior, there is a trade-off between the following four quality criteria:
- Fitness: The behavior seen in the event log must be the same as the reproduce model; a model with reasonable fitness can reproduce most of the traces, i.e., the paths taken by the cases in the log.
- Accuracy: The behavior discovered by the model should not be completely unrelated to what was seen in the event record and should not underperform the record; a model with poor accuracy is underfitting.
- Generalization: The discovered model should generalize the behavior seen in the event log, not overfitting the log. An overfitting model does not generalize enough.
- Simplicity: The model discovered should be as simple as possible.

For example, an oversimplified model will likely have low fitness or lack precision. Moreover, there is an apparent trade-off between underfitting and overfitting. Other properties can be applied to evaluate the model found, like soundness which assesses the existence of process bugs, but these are the main ones.

There are many algorithms, and they are numerous with the development of process mining. Only two algorithms are present, the simplest and most widely used:

The **alpha miner algorithm** was one of the first process discovery algorithms that could handle concurrency with less than three activities and could scan the event log for patterns. However, it has problems with noise, infrequent/incomplete behavior, and complex routing constructs.

The idea of the algorithm is to search for the traces to sort relationships between activities and then build a footprint matrix with these relationships. The algorithm can detect three sorting relations:
- Sequence (→): A follows B, but B does not follow A.
- Parallel (| | |): A follows B, as B follows A.
- No direct relation (#): A never follows B, and B never follows A.

Then the footprint matrix is converted into a Petri net, as an example Figure 12.
Straightforwardly, the alpha miner algorithm uses the directly following relation: start and end activities.

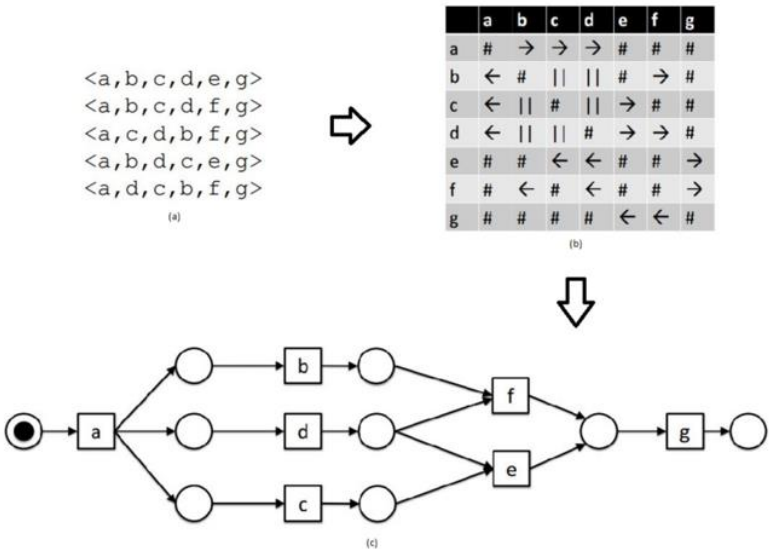


Figure 12: Process discovery using alpha algorithm, (a) Event log, (b) Footprint matrix, and (c) Petri net.

The **heuristic mining algorithm** considers the frequencies and sequences of events to build a process model. Less frequent paths are not incorporated into the model and detect short loops. It improves the alpha algorithm because it filters out noise behaviors and ignores single activities.

Figure 13, is summarized the three main steps of the algorithm. First, the frequency matrix is calculated using the number of times an activity follows another $|a >_L b|$. Then the dependency matrix, which uses the following formula, taking into account that L is the set of event logs, and $|a >_L b|$ is the number of times that activity a follows activity b in that set:

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

Where $|a \Rightarrow_L b|$ is the value of the dependency relation between 'a' and 'b'. This value is between -1 and 1. If it is close to 1, then there is a strong positive dependency between 'a' and 'b', i.e., 'a' is often the cause of 'b'. Finally, a Petri net is built using patterns and without infrequent relations between activities.

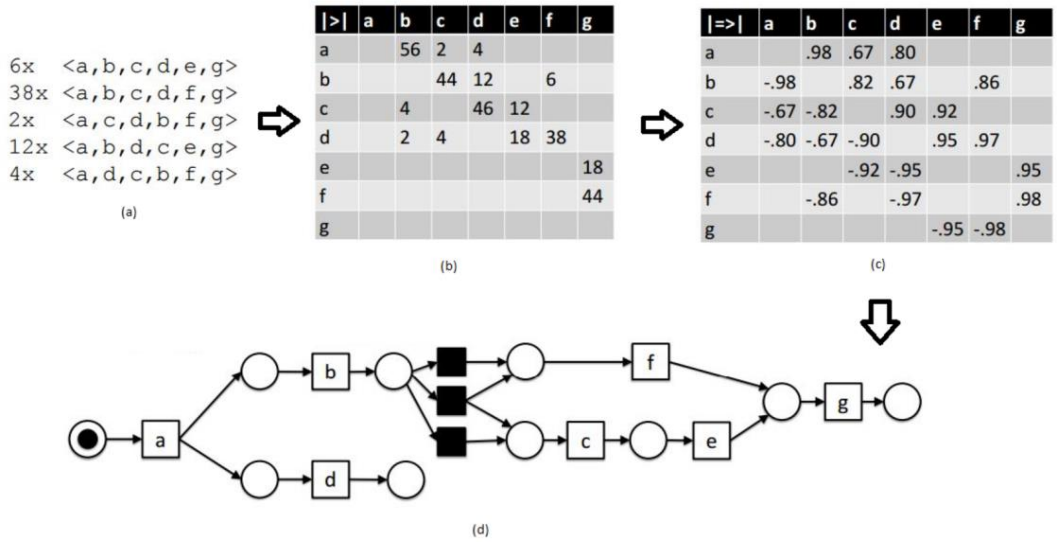


Figure 13: Process discovery using Heuristics miner (a) Event log, (b) Directly-follows matrix, (c) Dependency matrix, and (d) Petri net.

The alpha and heuristic miner algorithms provide process models in a direct and deterministic way. New approaches use iterative procedures to mimic the process of natural evolution, such as Inductive Mining ou Fuzzy Miner, which can handle vast and infrequent event data, best in flexibility, formal guarantees, and scalability.

2.4. Conformance checking and enhancement

Compliance checking can be applied when there is a process model, either discovered or built by hand, and a record of events that can be compared. The aim is to find similarities and discrepancies between the discovered and observed behavior. Compliance checking is relevant for the alignment and auditing of companies. For example, the event log can be replicated simultaneously with the process model to find undesirable deviations that suggest fraud or inefficiencies. Compliance checking techniques can also be used to measure the performance of process discovery algorithms and to repair models that are not well aligned with reality.

2.5. Software

In the last decades, a set of tools have been developed to keep up with the development of process mining techniques, both commercial and open-source tools. The big difference between commercial and open-source software is that commercial tools are more interactive, more robust in graph visualization, and have more user experience and multi-perspective nature of process analysis. In contrast, open-source tools focus on an academic study of the techniques and algorithms, which some commercial software lack. (GertJanssenswillen, 2021)

Tool	Vendor	Type	Website
Apromore	Apromore	Open source, Commercial	apromore.org
bupaR	—	Open source	bupar.net
PM4Py	—	Open source	pm4py.org
ProM	—	Open source	promtools.org
RapidProm	—	Open source	rapidprom.org
Aris	Software AG	Commercial	ariscommunity.com
Celonis	Celonis	Commercial	celonis.com
Disco	Fluxicon	Commercial	fluxicon.com/disco
EverFlow	Icaro Tech	Commercial	icarotech.com
Kofax Insight	Kofax	Commercial	kofax.com
Lana Process Mining	Lana Labs	Commercial	lana-labs.com
Minit	Minit	Commercial	minit.io
myInvenio	Cognitive Technology	Commercial	my-invenio.com
PAFnow	Process Analytics Factory	Commercial	pafnow.com
ProcessGold	ProcessGold	Commercial	processgold.com
ProDiscovery	Puzzle Data	Commercial	puzzledata.com
QPR ProcessAnalyzer	QPR Software	Commercial	qpr.com
Signavio Process Intelligence	Signavio	Commercial	signavio.com
StereoLogic Process Analytics	StereoLOGIC	Commercial	stereologic.com

Figure 14: Process mining software.

Were tested some tools like ProM, one of the most extensive and open processes mining frameworks; RapidProM, a variant of ProM that was developed on the RapidMiner framework; Celonis and Disco also tested, using a demo version with a hypothetical to order the full software with a student license. The chosen one was the latest free available - bupaR, an 'extensible set of R packages for business process analysis' developed in 2019. The bupaR packages stand for Business Process Analysis with R and provide support for different stages in process analysis, such as importing and preprocessing event data, calculating descriptive statistics, process visualization, and conformance checking (bupaR, 2019).

3. Business Understanding

Accounts payable (AP) are essential in balancing the flow of any accounting. While accounts receivable is responsible for the company's revenue, AP deals with expenses and associated purchases.

The AP department is responsible for company invoices that must be paid within a specific time limit to avoid supplier defaults. These invoices can be for products or services that have already been delivered but not immediately paid. AP can be used as short-term loans that can usually be paid within 15, 30, or 45 days after the company receives the invoice. Companies will try to extend the term to increase cash, while suppliers prefer a shorter term. Suppliers will sometimes sell at a discount to encourage quick payment if the company can pay within a shorter period. Companies that frequently exceed payment terms tend

to develop poor relationships with suppliers. Companies must balance paying their bills and maintaining good relationships with suppliers.

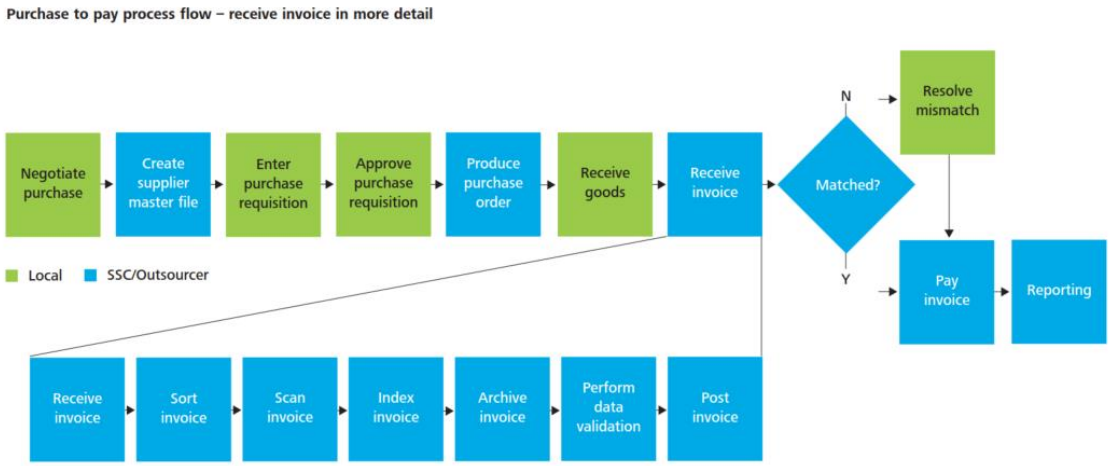


Figure 15: An example of a possible Purchase to pay process.

Although the procedure can differ from the company, here is an example of the AP process for dealing with the invoice: receipt of Purchase Order (PO) from the purchasing department, receipt of vendor invoices, matching of PO and invoice details, validation, approval, and invoice payments to the vendor. Many use a Purchase-to-Pay (P2P) procedure that covers all activities from purchase to invoice processing and supplier payments. The aim is to ensure the legitimacy and accuracy of any payment from the business. Some of the documents and definitions associated are:

- A Purchase Order (PO): is a legally binding agreement issued by the company to the supplier stating the type of good or service ordered and the agreed quantity and prices. For large companies, several entities are involved in the process, including the department requesting the goods or services, the purchasing department, the accounts payable department, the receiving department, and the supplier.
- Goods Receipt (GR): The goods receipt report confirms the purchase of the product or service ordered from the supplier. The details of the receipt report should match those of the previously issued order. This document should include the quantity of the item received, required details about the goods, and any notes of damage or issues with the shipment.
- Invoice: At the same time or upon receipt of the ordered goods, the company would receive the invoice from the supplier of the goods or services. The AP department's processing of the invoice is the sole responsibility, which checks the invoice details through correspondence between the order, the receipt report, and the supplier's invoice and schedules the payments after approval.
- Types of matching: There are different ways accounts payable departments check and match invoices. Some companies prefer two-way matching, which involves just

matching an invoice with a purchase order, but others prefer three-way matching. In three-way matching, the invoice is compared to the purchase order and the receipt report (alternatively, the goods receipt or order receipt). This way, avoid fraudulent invoices and avoid human error.

There are some common challenges along the P2P process, for instance:

- Time delays: Delays in processing an invoice and paying the supplier can snowball into late receipt of ordered items, poor credit ratings, poor vendor relations, and fees/fines.
- Matching errors: Three-way matching is necessary to ensure that all vendors are correctly paid. Discrepancies can occur when data entry errors or incorrect invoice information can take time to identify or correct. Errors lead to many problems throughout the accounts payable cycle, from friction with the vendor to delayed transactions.
- Double payment: Without proper protocols, accounting teams can accidentally double payments. Human error happens. Clear communication, software, and internal control protocols help prevent this.

Digital development has influenced processes in the business world, and the account payable process is one of them. The use of machine learning techniques and process mining has enabled the automation of invoice processes, helping employees keep a more accurate record of expenses and cash flow and can help automatically pay invoices by their due dates. Today's digital AP process includes capturing invoice data, coding invoices with the correct account and cost center, approving invoices, matching invoices with purchase orders, and posting them for payments. Therefore, the AP team has more time to devote to analytical and administrative functions that benefit the company.

Part II – The Action

4. Data Understanding

4.1. The proposed problems

The company is Klöckner Pentaplast (kp), has more than 30 locations worldwide, and is a global leader in packaging solutions under the Finance kp Shared Service Center in Porto. The shared service centers are characterized by bringing together a set of operations in a single location of a large company, thus enabling cost reduction for increased competitiveness in global markets, followed by improved quality and greater transparency. The team in the shared service center is referred to as SSC, and each company's local team will be referred to as LT.

For this report, the company provided real-life data from the software used to handle P2P processes, integrate with ERP systems, and use a Tableau Server to do OLAP analyses. It was asked to analyze the data using free choice techniques, suggest some challenges be faced and provide a wide range of insights beyond the given scope. The analysis should be driven to help quantify issues and resolve them based on the following KPIs (Key Performance Indicators) - they help the team leader monitor performance and analyze inefficiencies over time. The meaning will be explained in the following sections:

a) Receiving time,
b) Invoice senders (internal / external),
c) Reassess Line-item reading,
d) Invoices rejected,
e) Invoice lead time,
f) Suppliers by template creation,
g) Suppliers with more volume,
h) Lead time per approver,
i) Invoices and suppliers with and without PO per spend type,
j) Invoices per FTE (full-time employee),
k) Duplicate Invoices,
l) Touchless Invoices.

4.2. The data and process

This first stage of data extraction, cleaning, and transformation can take about 80% of the time. Once there is a good understanding of the concepts, and with the AP team that deals with the data daily, it is possible to choose just enough attributes to get answers from the 'pool' of available data.

Figure 16: Suppliers and companies transactions map using Tableau.

The data was collected in December and referred only to September of that year and consists of the four stages of P2P, in

Appendix I - Attributes available. This difference of months allowed the team time to deal with the invoices and to analyze only the processed invoices. As a significant data source, these are not organized and raise some doubts about their quality when confronted with other systems in the company. After the choice, it was considered valid values and had their respective meaning.

With over 500 attributes, this report mainly used the InvoiceHeader source, which contains several attributes about the invoices. The data has 27.572 invoices from 20 companies involving more than 400 million euros and 279 accounting employees, such as invoice approvals and 'goods received' (LT - local team) and manual matching (SSC – performed by the shared services center team).

General Overview

Company code	ERP	Country of organization	% of Total Number of invoices along Table (Down)	% of Total Gross Eur along Table (Down)	% of Total Distinct count of Supplier code along Table (Down)	Distinct count of Approver (FTE LT)	Distinct count of Manual matching by (FTE SSC)
9	JDE	GB	0.02%	0.51%	0.04%	1.0	1.0
12	JDE	GB	0.75%	1.91%	2.12%	5.0	3.0
21	JDE	GB	6.59%	5.64%	7.00%	4.0	3.0
61	JDE	DE	6.38%	5.93%	5.95%	1.0	4.0
62	JDE	DE	1.90%	0.99%	4.24%	1.0	3.0
68	JDE	NL	7.15%	4.09%	1.75%	1.0	4.0
71	JDE	DE	0.02%	0.01%	0.11%	1.0	1.0
72	JDE	DE	0.03%	0.00%	0.11%	1.0	1.0
78	JDE	DE	0.02%	0.07%	0.04%	1.0	1.0
101	JDE	PL	12.62%	2.99%	8.41%	2.0	3.0
1001	SAP	US	19.81%	32.98%	19.17%	49.0	5.0
1004	SAP	CA	6.46%	3.31%	5.32%	5.0	4.0
3001	SAP	DE	13.58%	21.84%	17.75%	102.0	10.0
3004	SAP	CH	2.03%	1.11%	4.35%	20.0	2.0
3005	SAP	ES	2.59%	3.07%	4.47%	14.0	3.0
3009	SAP	DE	5.71%	2.63%	7.44%	60.0	8.0
3010	SAP	PT	3.94%	5.31%	6.07%	15.0	3.0
3011	SAP	GB	2.42%	2.59%	3.42%	6.0	4.0
3103	SAP	ES	7.21%	4.93%	8.89%	21.0	2.0
7045	SAP	PT	0.74%	0.12%	1.34%	9.0	2.0
Grand Total			100.00%	100.00%	100.00%	258.0	21.0

Figure 17: Overview of the data - Invoices with completed cycles correspond to 27 572 invoices, 400 million euros, 20 subsidiaries, and 2687 suppliers.

The process starts with LT sending purchase orders (PO) to the supplier and putting them

into the system. When the supplier receives the PO, it replies with a product or service and an invoice for payment. As the company has the PO data in the system, it recognizes the data and automatically matches it when the invoice is received. If the invoice is not matched because the data is incorrect, the PO cannot be located, or for any other reason, the invoice is forwarded to the SSC team, who will match it manually. The invoice may be passed from hand to hand between the SSC team and the LT in manual matching to resolve discrepancies.

When the products are received, there may be a block payment code. If this is not filled, the payment of that invoice is made after the goods are received. However, this indicator changes over time then it was impossible to use it for this snapshot analysis.

4.3. The data quality issues

The data contains several attributes, and there are a few data quality issues that could influence the analysis result:

- Missing information: There is no information about the matching type in the available data. There is no indication if the invoice was a three-way or two-way match or if there was a before or after Goods Receipt (GR), even if there is no indication if there was a GR or whether it is a service or product purchase. Similarly, the associated attribute is not explicit in the block code, allowing knowing if there was GR. The different types of correspondence would allow different invoice flows to be recognized. Therefore, the analysis will only be on whether there is a PO. For the process mining, there is a hypothesis that there was a GR using the existence of a name in the 'Recipient' attribute, i.e., consider that someone gave the information of the reception of the product. Other activities are missing, like changing price or quantity.
- Repetitions: There are repeated names in the supplier attribute, and the same supplier can have several codes. The data with no homogenization will not give more precise analyses. Only the code will be used and not the names of suppliers, and it will be assumed that there is only one code per supplier.
- Incomplete information: The attribute 'Manual matching by' has only the last full-time employee (FTE) name. When the values of the invoice are not identical to the PO, the invoice may have to go through several FTEs. The same happens for the attribute 'Approvers'. The names were changed to a reference.
- Missing values: Some invoices are missing the Supplier and invoice number (many belong to removed invoices). Some fields are simply not filled in. For example, the approver often appears null in invoices that there was an intervention.
- Incorrect information: Unfortunately, some dates are not following what happened as checked with the team, namely Payment, Due, and Discount date. Using these data is avoided, mainly by companies using the JDE system. Another incorrect information occurs in some attributes that change over time. For example, the attrib-

ute 'Operation status' indicates why the invoice is pending, but once the situation is resolved, we are left with no information. This attribute is useful when a live analysis is done but does not allow us to know why the invoices already processed were pending.

5. Online Analytical Processing (OLAP) analysis

OLAP is the ability to manipulate and analyze a large volume of data from multiple perspectives. OLAP applications are used by managers at any level of the organization to enable them to perform comparative analyses to facilitate their daily decision-making. There are many tools on the market for OLAP analysis. The most widely used on an enterprise level is Microsoft's PowerBI. Used in this report was Tableau, as the company's P2P software system used it. In order to try to answer the company's questions, there were some adaptations to make it easier; in total, the sample was reduced to 2,420 invoices:

- For calculation, the median was used instead of the mean since the median is the value that divides the data set into two equal parts. The median is a more efficient measure of central tendency because it is not so affected by outliers.
- For the study, we use only invoices already processed. The attribute 'Invoice Status' is closed, processed, or removed, called complete cycles. The difference between the status processed and closed is that the invoices of the latter were already processed more than three months ago.
- The total gross of the invoices was in local currency. Thus, a new variable was created to homogenize this attribute in euros.
- For simplicity, only two invoice types were used, with PO and without PO, the most received by companies that use SAP as ERP, Figure 18.

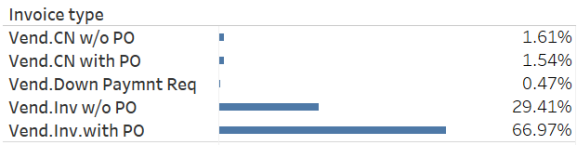


Figure 18: SAP Invoice type.

In one month, the median number of invoices daily received was 1,240, with the company 1001 being one of those that received the most, the total being 4,000 invoices. In Figure 19, we have an example of the performance of 3 companies. In the available data, no attribute is relative to points b) invoice senders, f) template, and i) spend type, and no result could be obtained.

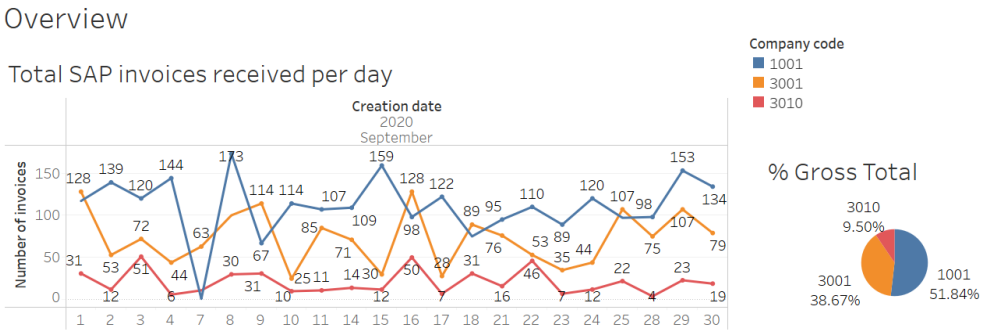


Figure 19: Overview of the tree of the companies.

a) Receiving time

Receiving time measures the difference between the time of reception of the invoice or 'Invoice creation' and the date of the invoice given by the supplier, the attribute 'Invoice date'. The objective is to detect suppliers or collaborators that delay sending the invoice for some reason, either by forgetfulness, inertia, or work method. This delay hinders the timely payment of the invoice and the application or not of the associated discount.

Figure 20 shows the median of 3 companies and the five worst suppliers without considering the gross value. For example, for company 1001, the total median is three days, with the worst supplier having a median of 431 days for the company to receive the invoice. It is only one invoice, probably a typing or system reading error. Having half of the invoices received within three days or less does not seem bad. However, for all suppliers with invoices above this amount, it may be worth the company contacting the suppliers or the associated employees to understand why there is such a long time interval to receive the invoice. Another possibility that invoices take so long is that they may have been rejected, and the suppliers sent back a new invoice with the same date as the previous one.

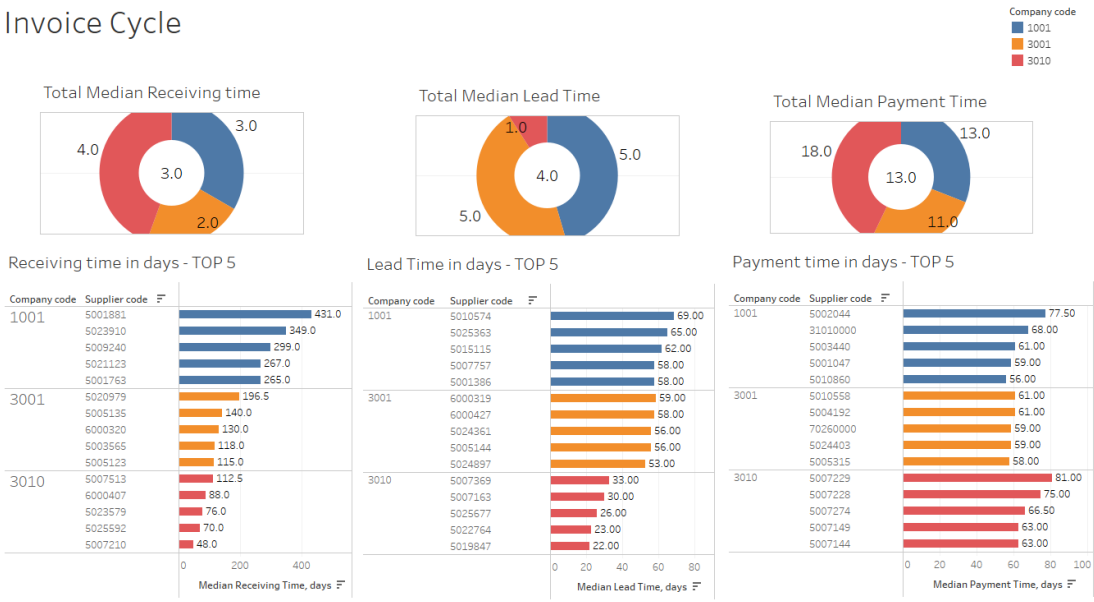


Figure 20: Invoice cycle.

e) Invoice Lead time

Again from Figure 20, it is possible to answer this metric that measures the time it takes for the invoice to be in the IT system for matching and approval before payment. It is possible to see that the median has increased relative to the arrival time of the invoice, so the matching and approver procedure should be improved by identifying the causes. Another metric is the time after the invoice leaves the system and is forwarded to an ERP software for payment. It was detected that there were problems in the system recording the payment time, but as an example, it is possible to see the suppliers with more days to perform this payment.

c) Item reading

A System automatically reads every invoice that comes into the company to the Scan provider, which has associated costs. To answer the question 'How many invoices with more than one line?' is to understand whether it is worth paying more for the system to read more lines on the received invoices. The attribute 'Nr of codingrows' provides precisely this type of reading.

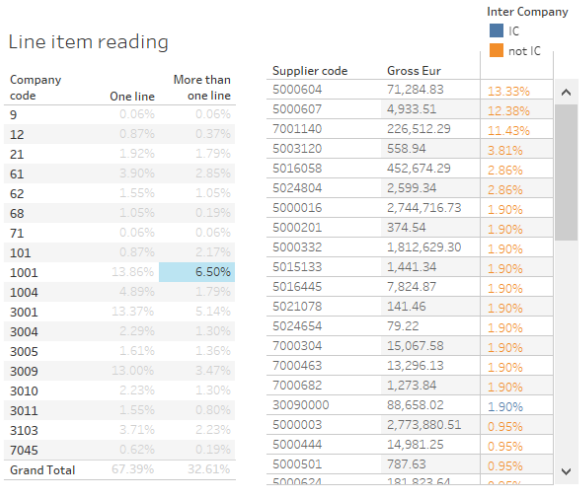


Figure 21: Number of coding rows.

More than 67% of companies receive invoices with only one line. Again, the example of company 1001 with the highest percentage of invoices and with more invoices with more than one line represents a median of 1,000 euros per invoice. In Figure 21, supplier 5000604 has the highest invoice number with more than one line. While the median invoice gross amount with only one line is 715 euros, and the median of invoices gross amount with more than one line is 4,580 euros.

d) e k) Invoices rejected and duplicated

Understanding the process is fundamental to know what attributes to look for and interpret the results. In Figure 22, all the invoices of one month were considered, and it is possible to have an idea of the whole process of the invoice combining the various attributes, the LT at the companies is responsible for 'Waiting for goods receipts' and 'In header approval'. The remaining steps are the responsibility of the SSC team.

Invoice Process

Invoice status	Invoice sub-status	Team	Reason for failed validation	Inter Company IC	Inter Company not IC	Grand Total
Received	Invalid	SSC	Invalid Value		0.01%	0.01%
	Returned	SSC	None		0.02%	0.02%
	Valid	SSC	None	0.01%	0.02%	0.02%
In matching	In manual order matching	SSC	None	0.04%	0.04%	0.08%
	Waiting for goods receipts	LT	None		0.05%	0.05%
In workflow	In header approval	LT	None	0.20%	0.47%	0.67%
In transfer	Incomplete for transfer	SSC	None		0.02%	0.02%
	Transfer failed	SSC	None	0.02%	0.02%	0.04%
	Transfer in progress	SSC	None		0.02%	0.02%
Processed	Processed	SSC	None	3.92%	36.57%	40.48%
Closed	Closed	SSC	None	0.25%	38.99%	39.25%
Removed	Removed	SSC	Duplicate	0.07%	7.17%	7.25%
			Invalid Value		0.25%	0.25%
			Mandatory Data Missing	0.01%	0.02%	0.02%
			Multiple Errors	0.02%	0.40%	0.42%
			None	6.13%	5.29%	11.41%
Grand Total				10.67%	89.33%	100.00%

Figure 22: Invoice Process where LT is the Local Team and SSC is the team from Shared Services Center.

When the data was obtained, there were still invoices to be processed three months after receipt in the System. In Figure 22, it is possible to have an idea of the several states in the process of an invoice. Many removed invoices do not indicate a 'reason for validation'. The two measures were put together, 'rejected' and 'duplicated', because there is little information on why the invoice was rejected.

Rejected Invoices

Inter Company	Processed or Closed	Removed
IC	40.09%	59.91%
not IC	85.21%	14.79%
Grand Total	80.47%	19.53%

Company code	Processed or Closed	Removed
9	100.00%	
12	73.40%	26.60%
21	89.08%	10.92%
61	77.07%	22.93%
62	87.39%	12.61%
68	22.60%	77.40%
71	100.00%	
72	50.00%	50.00%
78	100.00%	
101	91.00%	9.00%
1001	83.48%	16.52%
1004	89.36%	10.64%
3001	84.91%	15.09%
3004	84.25%	15.75%
3005	82.10%	17.90%
3009	80.95%	19.05%
3010	78.30%	21.70%
3011	79.87%	20.13%
3103	88.12%	11.88%
7045	81.72%	18.28%

Company code	Index	Supplier code	Processed or Closed	Removed
1001	1	7000259	96.12%	3.88%
	2	7000668	100.00%	
	3	5022680	64.08%	35.92%
	4	5002292	98.28%	1.72%
	5	5013234	98.00%	2.00%
3001	1	5006309	69.51%	30.49%
	2	5004906	92.65%	7.35%
	3	5004541	82.69%	17.31%
	4	5003747	100.00%	
	5	5005550	22.73%	77.27%
3010	1	6000535	26.92%	73.08%
	2	6000556	50.00%	50.00%
	3	5007228	77.78%	22.22%
	4	5007158	100.00%	
	5	5006664	100.00%	

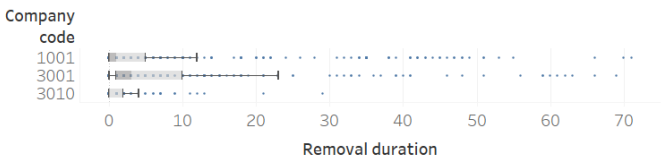


Figure 23: Rejected Invoices with Invoice remove duration.

In Figure 23, it is possible to verify that 20% of the invoices that arrive in the system are rejected. It presents the percentage for each company and the suppliers with the highest number of invoices from the three companies.

Good communication with suppliers to identify the associated problems can give advantages to companies and suppliers. Companies take less time to process invoices, can reduce staff costs, and avoid double payments while suppliers are paid on time.

FTE SSC Removal

	Match duration in days				
Manual Match User	< 5	6 to 15	16 to 25	> 25	Grand Total
Grand Total	69.51%	19.11%	5.69%	5.69%	100.00%
M-2	0.41%				0.41%
M-3	10.16%	3.25%	0.81%	1.22%	15.45%
M-4	2.85%			0.81%	3.66%
M-5	2.03%	1.22%	0.81%		4.07%
M-6	2.03%	0.81%			2.85%
M-7	0.81%	1.63%	0.81%		3.25%
M-8	1.22%			0.81%	2.03%
M-10	0.41%			0.41%	0.81%
M-12	0.81%		0.41%		1.22%
M-13	2.85%	1.22%			4.07%
M-14	0.41%	0.41%			0.81%
M-15	0.81%				0.81%
M-16	6.10%	5.28%	1.22%	1.63%	14.23%
M-17	2.03%	0.81%	0.81%	0.41%	4.07%
M-18	26.42%	1.63%	0.41%		28.46%
M-19		0.41%			0.41%
M-20	10.16%	2.44%			12.60%
M-21			0.41%	0.41%	0.81%

FTE LT Removal

	Approval duration in days				
Approver (FTE LT)	< 5	6 to 15	16 to 25	> 25	Grand Total
Grand Total	64.86%	18.92%	5.41%	10.81%	100.00%
A-135	8.11%				8.11%
A-145	8.11%				8.11%
A-245	8.11%				8.11%
A-126		5.41%			5.41%
A-2				5.41%	5.41%
A-218		2.70%	2.70%		5.41%
A-115	2.70%				2.70%
A-129	2.70%				2.70%
A-147		2.70%			2.70%
A-149		2.70%			2.70%
A-155	2.70%				2.70%
A-161			2.70%		2.70%
A-175		2.70%			2.70%
A-179	2.70%				2.70%
A-205	2.70%				2.70%
A-215	2.70%				2.70%
A-217	2.70%				2.70%
A-23	2.70%				2.70%
A-250	2.70%				2.70%

Figure 24: 'Which employees make the biggest invoice rejection?'.

Regarding FTE, from the team doing the manual matching, one stands out with many invoices being removed, M-18 doing over 26% of all rejections, and does it in less than five days. Of the approvers' team, the three that do the most rejections do it in the first five days. In both cases, the number of invoices removed in one week is less than 70%. Only the last FTE name is registered. The invoices may go from 'hand to hand' until it is decided that they should be removed.

h) e J) Employee analysis

Processing invoices as productively as possible can have an impact on working capital. Calculating how many invoices are processed per FTE can identify suitable working methods of some employees to be reproduced or organize staff for the various approaches to deal as well as possible with suppliers. Figure 25 shows two alternatives to analyze FTE's manual matching and approval. The number of invoices up to the 3rd percentile of the original data is calculated to remove outliers.

SSC Matching

Match duration in days

Manual Match User	< 5	6 to 15	16 to 25	> 25	Grand Total
M-2	1.05%	0.21%	0.12%	0.19%	1.58%
M-3	5.53%	1.75%	0.39%	0.16%	7.83%
M-4	2.51%	0.41%	0.31%	0.12%	3.35%
M-5	2.30%	1.34%	0.25%	0.12%	4.01%
M-6	16.14%	1.93%	0.08%		18.15%
M-7	4.24%	2.69%	0.53%	0.06%	7.52%
M-8	1.52%	1.13%	0.16%	0.04%	2.84%
M-9	0.16%				0.16%
M-10	0.04%	0.14%	0.06%	0.10%	0.33%
M-11	0.08%				0.08%
M-12	0.45%	0.43%	0.06%	0.27%	1.21%
M-13	5.51%	2.78%	0.47%	0.18%	8.94%
M-14	2.38%	0.90%	0.18%	0.19%	3.64%
M-15	0.37%	0.06%			0.43%
M-16	6.44%	4.11%	0.97%	0.23%	11.76%
M-17	4.91%	2.92%	0.82%	0.49%	9.13%
M-18	7.81%	0.35%	0.02%		8.18%
M-19	1.38%	0.06%			1.44%
M-20	5.02%	0.31%		0.08%	5.41%
M-21	3.87%	0.10%	0.04%	0.02%	4.03%
Grand Total	71.71%	21.61%	4.44%	2.24%	100.00%

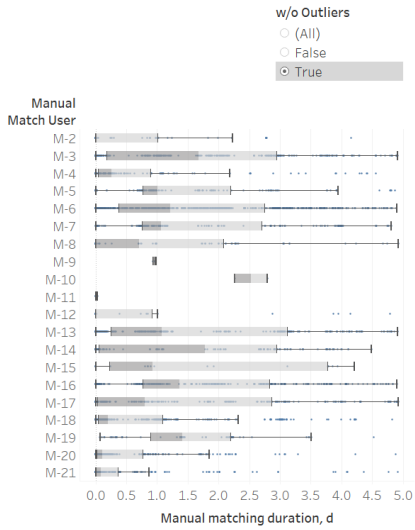


Figure 25: Percentage of invoices processed and duration by the manual matching team in one month.

Almost 30% of invoices take more than five days to match manually. M-6 stands out, solving 18% of the problems, with a median of 1,2 days. The team unable to match the invoice will likely send it to the local team that does the approval and reception of the product. The fact that the invoice takes a long time to match could mean that the data must be approved or the GR must wait for the local team, and only then is the match solved.

LT Approval

Approval duration in days

Approver (FTE LT)	< 5	6 to 15	16 to 25	> 25	Grand Total
Grand Total	65.97%	21.11%	8.59%	4.33%	100.00%
A-null	13.32%	7.95%	1.93%	2.33%	25.52%
A-57	19.82%	0.16%	0.08%	0.08%	20.14%
A-166	3.45%	4.57%	3.69%	0.40%	12.12%
A-218	1.36%	3.21%	1.69%	0.24%	6.50%
A-141	4.90%	0.40%	0.24%	0.16%	5.70%
A-30	2.81%	0.48%	0.08%	0.16%	3.53%
A-247	2.73%	0.32%			3.05%
A-55	2.89%		0.08%		2.97%
A-193	0.40%	2.25%	0.16%	0.08%	2.89%
A-175	2.49%	0.24%			2.73%
A-158	1.93%	0.16%			2.09%
A-181	1.44%				1.44%
A-99	1.20%	0.08%			1.28%
A-14	0.56%	0.48%	0.16%		1.20%
A-58	0.88%	0.16%		0.16%	1.20%
A-28	0.96%			0.08%	1.04%
A-184	0.72%	0.08%	0.08%		0.88%
A-189	0.88%				0.88%
A-120	0.48%	0.08%			0.56%
A-238	0.16%	0.16%		0.08%	0.40%
A-177	0.24%	0.08%			0.32%
A-245	0.24%		0.08%		0.32%
A-68	0.24%	0.08%			0.32%

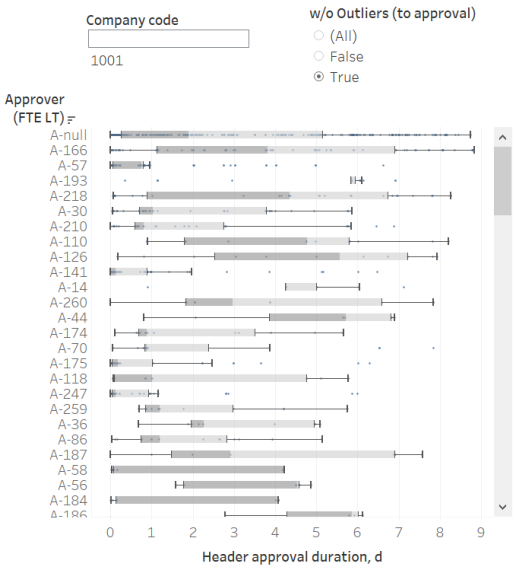


Figure 26: Percentage of invoices processed and duration by the approval team in one month.

Regarding LT, there are many invoices where the name is not recorded by the system, 25% of all invoices. However, not considering these, and giving the example of company 1001, another FTE that stands out is the A-57, which resolves almost 20% of all company invoices in less than five days, with a median of less than one day. In order to compare the teams and identify the suppliers, the following tables have the median of the manual matching and approval time for different gross amount ranges.

For invoices with PO and below 1,000 euros, the median of manual matching is five days, a high number of days in which the invoices could be processed automatically. In the other table below, we can see the list of the respective suppliers.

Gross Amount

Invoice type
- (All)
- Vend.CN w/o PO
- Vend.CN with PO
- Vend.Down Paymnt Req
- Vend.Inv w/o PO
- ☑ Vend.Inv.with PO

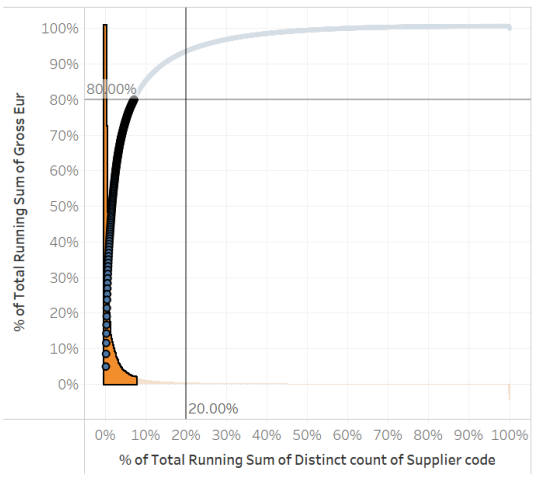
Company code	Gross Amount Range in Euros <1.000 Median Manual matching duration, d	<1.000 Median Header approval duration, d	1.000 to 10.000 Median Manual matching duration, d	1.000 to 10.000 Median Header approval duration, d	>10.000 Median Manual matching duration, d	>10.000 Median Header approval duration, d
1001	5.01	2.91	3.23	4.27	3.00	6.84
1004	1.15	5.93	1.45	4.13	6.26	7.27
3001	4.91	1.81	5.76	1.19	6.97	1.90
3004	1.85	0.43	2.17	0.77	1.10	3.71
3005	0.07	4.90	0.13	5.89	0.15	1.81
3009	1.99	4.08	2.98	1.88	1.38	3.36
3010	0.17	4.20	0.10	4.85	0.45	5.42
3011	2.86	0.92	3.27	2.99	1.12	19.07
3103	0.12	1.37	0.18	1.83	0.02	1.20
7045	0.00	0.05	0.00	0.01		
Grand Total	3.01	2.96	2.93	2.71	1.92	3.71

Supplier code	Median Gross Eur	Median Manual matching duration, d	Median Header approval duration, d
5010574	231.23		67.9
5025363	650.25	0.0	64.6
5015115	779.44	18.9	43.1
7000161	144.56	33.7	35.7
5000847	405.19	0.1	33.0
5019901	138.13	0.1	30.8
5002126	70.78	3.8	29.4
5000566	235.08	24.9	29.1
5012871	155.72	26.5	28.5
5000037	646.58	0.8	27.3
5023867	825.56	3.4	27.2
7000056	708.27	3.0	26.2
7000230	236.29	0.8	24.5
5000192	286.88	4.6	23.9
5009240	248.91	0.1	21.9
5023543	102.00	21.6	20.5
5001304	864.42	3.8	20.4
5015589	574.39	3.8	18.8
5009116	215.73	4.1	18.1
7000452	228.58	1.7	14.8
5901903	650.25	19.8	14.3
5001021	117.36		14.1
5022665	63.77	0.1	14.1
5000108	405.25	40.3	13.7

Figure 27: 'Which LT or SSC teams spend more time on an invoice?'

g) Suppliers with more volume.

Identifying suppliers with more volume allows us to give special attention to these suppliers to create an excellent relationship to facilitate the whole process. For this, a Pareto chart was made, which contains both bars and a line chart, where individual values are represented in descending order by bars, and the line represents the upward cumulative total. The Pareto principle says that about 80% of the effects come from 20% of the causes. In this case, 80% of the total gross derive from 7% of the suppliers.



Suppliers

Supplier code	Company code	Gross Eur	
5004056	3001	3,824,911.76	not IC
5000016	1001	2,744,716.73	not IC
5024390	3005	69,696.00	not IC
	3010	2,287,566.82	not IC
5006309	3001	2,054,974.24	not IC
30090000	1001	831,666.44	IC
	1004	185,226.07	IC
	3001	1,169,859.59	IC
	3005	-228,944.17	IC
	3010	-43,767.08	IC
	7045	-25,466.67	IC
5005427	3001	1,826,316.06	not IC
5000332	1001	1,815,095.34	not IC
5005412	1001	1,740,600.00	not IC
	3001	29,854.80	not IC
	3010	2,296.60	not IC
	3011	3,621.12	not IC
31140000	1001	254,258.92	IC
	3001	57,266.57	IC
	3010	1,852.24	IC
	3011	198,129.03	IC
	3103	726,576.88	IC
5008146	3001	1,214,645.30	not IC
93329	61	1,149,852.56	IC

Figure 28: Pareto Chart where 80% of the Invoices' Gross Total derive from 7% of the suppliers.

In Figure 28, the suppliers with the respective amounts and companies to which they provided services or products are identified. The 5th, 9th, and 11th largest suppliers with the highest amounts are suppliers belonging to the same corporate group.

l)Touchless Invoices

The SSC team aims to handle as many invoices as possible automatically. Software that does this is sought to reduce employee costs, shorten invoice processing time and human error, increase efficiency and professional function within the organization, and offset the cost of purchasing the software. It is also advantageous for a business to increase its touchless processing rate as it can capture discounts from suppliers by paying invoices well within early payment discount terms.

Touchless invoices are the metric that gives the percentage of invoices processed without human intervention between receipt of invoice and immediate payment; this means that for an invoice to be automatically matched, we should not have any FTE records. At the time of this report, only PO invoices can be automated, and the invoice should be handled on the same day. The system used by the company did not provide this metric. The code used to find the automatically processed invoices is as follows:

Touchless

```
If [ Manual Match User]=='M-null'
and [Approver]=='A-null'
and [Recipient]=='R-null'
and [Association type]=='Based on purchase order'
and [Creation date]==[Transfer completed time]

then 'yes'
else 'no'
END
```

Figure 29: Touchless code.

As a result, the values obtained are in Figure 30, considering all the companies. In one month, a little more than 20% of invoices were treated automatically, with company 3103 with the best performance, with 44.6%. Also, we can see the case of company 1001, with 17.8% of touchless, and the respective suppliers, notice that the third supplier with the highest number of invoices belongs to the company group, and from the 35 invoices received in one month, none was automatically processed.

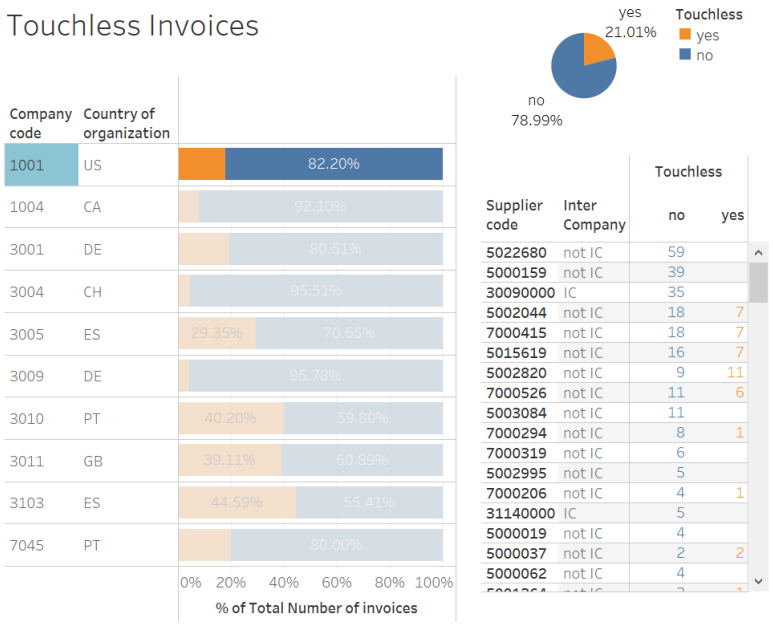


Figure 30: Number of touchless invoices by the company and their suppliers, values obtained excluding cases with human intervention, based on the purchase order and with similar entry and exit system dates.

Suppliers from the company group have poor efficiency where practices are not applied to increase invoice automation, Figure 31.

	Touchless	
Inter Company	no	yes
IC	95.45%	4.55%
not IC	88.10%	11.90%

Figure 31: Touchless invoices between suppliers in or not Inter Company.

Although the analysis seems simple, it allowed us to have an idea of the invoice process and transform the data in order to be able to apply process mining.

6. Process Mining Analysis

Process improvement is one of the components that companies have been betting on in this century. The amount of available data obtained from the various systems that support the business has allowed the analysis and improvement of all the processes involved in a field of study called process mining, which aims to discover valuable knowledge from process data. Process mining techniques allow companies to find process deviations and automate actions to correct them using event data records.

Event data analysis can be done from an organizational perspective, focusing on how people are involved. From a control flow perspective, it focuses on the flow and structuring of the process. Moreover, from a performance perspective, it focuses on time and efficiency.

This chapter presents the results and discussions on applying process mining techniques to the data analyzed. First, the data were transformed to obtain a suitable event log. Then a process discovery and process performance analysis are performed using invoices with and without PO and GR, thus obtaining different flows. In Appendix II – R programming code is possible to see the R code behind the analysis.

6.1. Event Log

The process mining starts by extracting data from one or more information systems and transforming it into event logs. Then the processing data where we can aggregate the data, removing information that is too detailed. The first two steps use the Tableau prep from the previous data analysis. Tableau Prep is an ETL tool that allows one to extract data from various sources, transform it, and output it, doing tasks such as joins, unions, and aggregations.

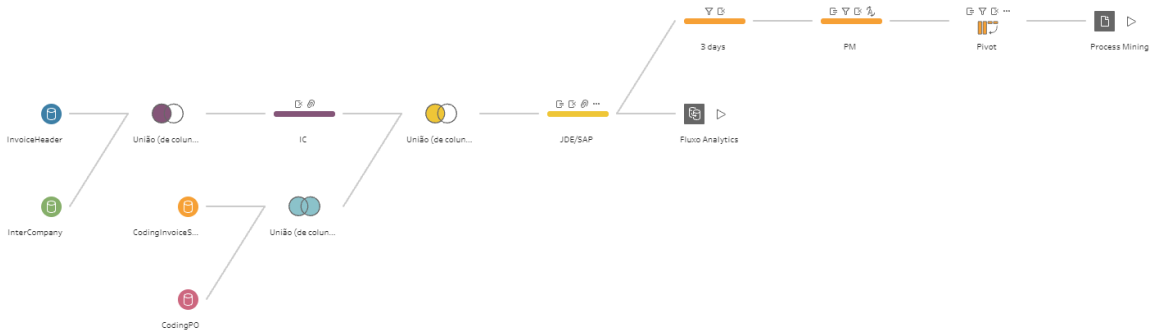


Figure 32: Flow from Tableau Prep.

The data provided had many dates, many of them prominent, others not so obvious. After some checks with the company, the most important and obvious ones were chosen to simplify the process. They are:

a) Invoice date - date of the invoice indicated by the supplier,
b) Creation date - date of reception of the invoice in the company, name simplified to just 'Creation',
c) Matched date - the final date of the invoice match with the respective PO by the SSC team. Abbreviated it to 'Match',
d) Approval time - final date of approval by the local LT team. Use only 'Approval',
e) Ready for transfer date - Date after approval and matching. Simplified to 'Ready for transfer',
f) Transfer completed time - At this date, the invoice goes to another ERP billing support system to proceed to payment. It is simplified to 'ERP Transfer'.
g) Removal date - Date of removal of the invoice. At this point, the invoice is discarded by the System for various reasons: duplication, errors, or incomplete data.

The other essential variables that were considered necessary for process mining are:

h) Manual matching count - number of users responsible for the matching of the invoice with the PO that, for some reason, may go through several users.
i) Manual matching duration, d - duration of the matching process
j) Manual matching - registers the last user to match the invoice, even if it has passed by several users.
k) Header approval count - count the approvers through which a specific invoice passed.
l) Header approval duration, d - approval duration
m) Approver - register of the last approver

During the matching and approving process, the invoices go from the SSC to the LT and vice versa to agree on the invoice data and then make the respective payment.

Invoice ID	Invoice date	Creation	Match	Approval	Ready for transfer	Removal
5ad740bbe938405097de4a7983ef7953	20/08/2020, 00:00:00	03/09/2020, 00:01:00	null	null	null	03/09/2020, 00:
4f43cea4422e4f16911d5b09e56072f1	24/08/2020, 00:00:00	03/09/2020, 00:01:00	null	null	null	03/09/2020, 00:
b20b560631ab49d9b569c42962a06405	04/09/2020, 00:00:00	08/09/2020, 00:01:00	null	09/09/2020, 00:04:00	09/09/2020, 00:05:00	null
93fdbeeb448041cf9ab36e3339d8788b	04/09/2020, 00:00:00	08/09/2020, 00:01:00	null	09/09/2020, 00:04:00	09/09/2020, 00:05:00	null
ee0f07aa409d4e6e937d619cb829e156	05/09/2020, 00:00:00	08/09/2020, 00:01:00	null	09/09/2020, 00:04:00	09/09/2020, 00:05:00	null
c55e315aabfd495c9185f2f544c8e0c7	31/08/2020, 00:00:00	01/09/2020, 00:01:00	07/09/2020, 00:03:00	08/09/2020, 00:04:00	08/09/2020, 00:05:00	null
c3da91dbcb08478a885c3731daecd817	31/07/2020, 00:00:00	01/09/2020, 00:01:00	05/09/2020, 00:03:00	null	05/09/2020, 00:05:00	null
c1720bc6928046239f91b50e19d97fc0	31/07/2020, 00:00:00	01/09/2020, 00:01:00	05/09/2020, 00:03:00	null	05/09/2020, 00:05:00	null
4e6eb6dd46914219ab1dbf3ea63ea083	31/08/2020, 00:00:00	01/09/2020, 00:01:00	05/09/2020, 00:03:00	null	05/09/2020, 00:05:00	null
24d05463327f4e9ab527f3c59f5e0e40	07/09/2020, 00:00:00	08/09/2020, 00:01:00	08/09/2020, 00:03:00	null	08/09/2020, 00:05:00	null
c196bc38335549e1bd7eca37b2d8f2a5	01/09/2020, 00:00:00	03/09/2020, 00:01:00	07/09/2020, 00:03:00	08/09/2020, 00:04:00	08/09/2020, 00:05:00	null
ba87b223584e4bdc8d717be99868a5ae	04/09/2020, 00:00:00	07/09/2020, 00:01:00	08/09/2020, 00:03:00	09/09/2020, 00:04:00	09/09/2020, 00:05:00	null
72ca260de43948cc85906830760290bc	04/09/2020, 00:00:00	07/09/2020, 00:01:00	08/09/2020, 00:03:00	10/09/2020, 00:04:00	10/09/2020, 00:05:00	null
a7890b2cceb44498818787e429ba90bd	04/09/2020, 00:00:00	22/09/2020, 00:01:00	02/10/2020, 00:03:00	null	02/10/2020, 00:05:00	null
d9ef9b16722e46798031ffcd4f2f14bd	13/08/2020, 00:00:00	02/09/2020, 00:01:00	07/09/2020, 00:03:00	null	07/09/2020, 00:05:00	null
a6863f3a8f7645bebc082d89fbdb1be5	08/09/2020, 00:00:00	22/09/2020, 00:01:00	29/09/2020, 00:03:00	null	29/09/2020, 00:05:00	null
bf11bfd929a44c97967aaa7a2ac31b97	31/08/2020, 00:00:00	01/09/2020, 00:01:00	07/09/2020, 00:03:00	null	07/09/2020, 00:05:00	null
e0dab47ab4884ba09029ce2cdd4b8c31	15/06/2020, 00:00:00	03/09/2020, 00:01:00	05/09/2020, 00:03:00	07/09/2020, 00:04:00	07/09/2020, 00:05:00	null
65d98486b3dd4426b3dd1f65395363de	04/09/2020, 00:00:00	08/09/2020, 00:01:00	21/09/2020, 00:03:00	21/09/2020, 00:04:00	21/09/2020, 00:05:00	null
fa26c61db46a4c5193705bd7c1b7abb0	11/09/2020, 00:00:00	14/09/2020, 00:01:00	null	null	15/10/2020, 00:05:00	19/10/2020, 00:

Figure 33: Data before transformation, Tableau Prep view.

As described in the last chapter, an essential question for the invoice route is whether or not to have 'goods received' (GR). Once the invoice is received, its payment may be blocked until the goods are collected. Having this indication helps to find the various flows in the process. In the provided data, there was no field with this measure, so the 'Recipient' attribute (a name) will be considered a GR indication, with a 'yes' or 'no' intervention. As each line represents an activity, an FTE was associated with the corresponding activity: for manual matching and approval.

After pivoting the data to transform each date into an activity, the number of rows went from 8 thousand to 42 thousand, increasing more than five times. The number of invoices remains the same. The point is that the more dates, the more activities, the more rows a process has, and the denser it gets.

	A	B	C	D	E	F	G	H	I	J	K
1	order	InvoiceID	date	activity	SupplierCode	GrossEur	FTE	GR	status	InterCompany	CountrySupplier
2	6880	002966a1eb6d444a8e4ce5419c342c11	22/9/20 0:00	Invoice date	7000668	18,955		yes	complete	not IC	US
3	7773	002966a1eb6d444a8e4ce5419c342c11	24/9/20 0:01	Creation	7000668	18,955		yes	complete	not IC	US
4	7948	002966a1eb6d444a8e4ce5419c342c11	24/9/20 0:04	Approval	7000668	18,955	A-57	yes	complete	not IC	US
5	8008	002966a1eb6d444a8e4ce5419c342c11	24/9/20 0:05	Ready for transfer	7000668	18,955		yes	complete	not IC	US
6	8100	002966a1eb6d444a8e4ce5419c342c11	24/9/20 0:06	Transfer completed	7000668	18,955		yes	complete	not IC	US
7	455	0041853df05240b9a888a2de1d007d90	27/8/20 0:00	Invoice date	0	1700		no	complete	not IC	
8	9760	0041853df05240b9a888a2de1d007d90	30/9/20 0:01	Creation	0	1700		no	complete	not IC	
9	10392	0041853df05240b9a888a2de1d007d90	1/10/20 0:07	Removal	0	1700		no	complete	not IC	
10	82	005cae060af145118a8673cb02c46ff0	20/4/20 0:00	Invoice date	7000069	62,5005		yes	complete	not IC	US
11	4329	005cae060af145118a8673cb02c46ff0	15/9/20 0:01	Creation	7000069	62,5005		yes	complete	not IC	US
12	5922	005cae060af145118a8673cb02c46ff0	18/9/20 0:03	Match	7000069	62,5005	M-13	yes	complete	not IC	US
13	8812	005cae060af145118a8673cb02c46ff0	28/9/20 0:04	Approval	7000069	62,5005	A-218	yes	complete	not IC	US
14	8959	005cae060af145118a8673cb02c46ff0	28/9/20 0:05	Ready for transfer	7000069	62,5005		yes	complete	not IC	US
15	9175	005cae060af145118a8673cb02c46ff0	28/9/20 0:06	Transfer completed	7000069	62,5005		yes	complete	not IC	US

Figure 34: Data after transformation, Excel view.

While in the previous chapter, each row consisted of an invoice, i.e., each row had a different ID. In process mining, each row represents an event, i.e., an invoice has as many lines as activities or events. Each row should be an event with at least the following information:

- A timestamp
- A case identifier
- An activity label
- An activity instance identifier
- A transactional life cycle stage
- A resource identifier

The type of event is called the status of the live cycle. There are many choices for the status, and by simplicity, the 'complete' was chosen for all instants. The set of all events is called the event log. Another component is resources. Resources can be the actors in the process and perform the activities. In our context, we have the employees that do the manual matching and the invoice approvals.

The event log function takes a data frame as input and returns a log object as output:

```
# To create the event log object.
invoices <- eventlog(
  Book1,
  case_id = "InvoiceID",
  activity_id = "activity",
  activity_instance_id = "order",
  lifecycle_id = "status",
  timestamp = "date",
  resource_id = "FTE"
)
```

Figure 35: Event log R function.

```
>   summary(invoices)
Number of events:  42416
Number of cases:  7697
Number of traces:  84
Number of distinct activities:  8
Average trace length:  5.510718

Start eventlog:  2013-08-31
End eventlog:  2020-11-27 00:07:00
```

Figure 36: Output Event log R function.

The number of cases represents the number of invoices, the number of activities corresponds to the number of dates chosen for the model, and the number of events is the total number of invoices in those activities. Finally, the number of traces is the specific number of paths an invoice can take. However, only extracted invoices were received in September 2020; at least one invoice has a date of 2013, Figure 36.

There are several possible combinations of paths an invoice can take. In real-life situations, many activities can occur in any order and be repeated several times. In the data, there are 84 different traces, meaning there are 84 possible paths that the invoices take.

```
> #  Each case is described by a sequence of activities, it is called a trace.
> traces(invoices)
# A tibble: 35 x 3
   trace                                                                                                  absolute_frequency relative_frequency
   <chr>                                                                                                               <int>              <dbl>
 1 Invoice date,Creation,Ready for transfer,Transfer completed                                                           571             0.236
 2 Invoice date,Creation,Match,Ready for transfer,Transfer completed                                                     399             0.165
 3 Invoice date,Creation,Match,Approval,Ready for transfer,Transfer completed                                            394             0.163
 4 Invoice date,Creation,Approval,Ready for transfer,Transfer completed                                                  384             0.159
 5 Invoice date,Creation,Removal                                                                                         344             0.142
 6 Invoice date,Creation,Match,Match,Ready for transfer,Transfer completed                                               135             0.0558
 7 Invoice date,Creation,Approval,Match,Approval,Ready for transfer,Transfer completed                                    39             0.0161
 8 Invoice date,Creation,Match,Approval,Match,Approval,Ready for transfer,Transfer completed                              31             0.0128
 9 Invoice date,Creation,Match,Match,Match,Ready for transfer,Transfer completed                                          24             0.00992
10 Invoice date,Creation,Approval,Match,Match,Approval,Ready for transfer,Transfer completed                              19             0.00785
# ... with 25 more rows
> |
```

Figure 37: Traces list.

In a more visual option, we can get the following:

Figure 38: Trace explorer, 90% of the invoices.

We can see that more than 20% of the invoices have the same trace with the five activities: 'Invoice date', 'Creation', 'Ready for transfer', 'ERP transfer', and 'Payment'.

6.2. Process discovery

Process discovery is one of the three types of process mining and aims to build a model of the process using event logs. The process model discovered it should capture the behavior of the data recorded over time. It should be precise and fit the log.

One of the functions available in bupaR is 'process_map', which generates a direct-follow graph (DFG) that shows the process activities and the flows between them, Figure 39. It is an authentic representation of the log; there is no algorithm behind it that filters or generalizes the data - an arrow between nodes represents each sequence of two tasks in the log.

The nodes' colors and the thickness of the arrows indicate the most frequent activities and process flows. Each invoice travels one and only one path, starting at the start point and ending at the endpoint.

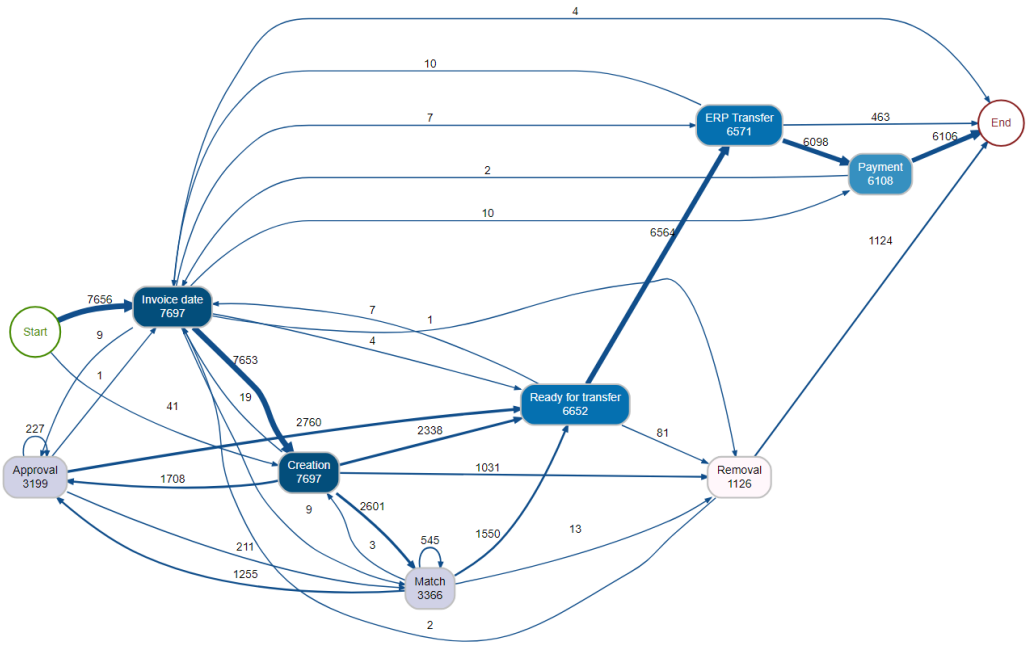


Figure 39: Direct-follow graph for all companies' invoices received in one month.

At first glance, the activities all appear as we wanted, there are loops in 'Approval' and 'Match' as expected. It seems there is some confusion with Payment, which was expected. Were applied some filters to remove the dates with errors.

The alpha miner algorithm is provided through the package pm4py; it allows an interface to a process mining library in Python. This package uses the reticulate package as a bridge between pm4py and the R package bupaR. It provides various process discovery algorithms, evaluation measures, and alignments (Interface to the PM4py, s.d.). The resulting object consists of three elements, a network, a start tag, and an end tag. The network can be visualized using petrinetR as follows. (bupaR, 2019)

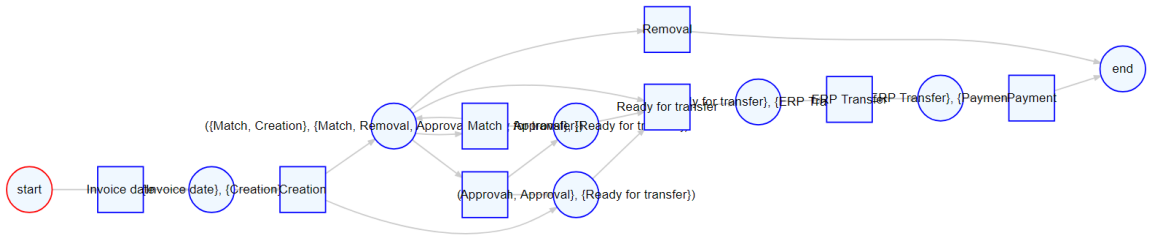


Figure 40: Petri net obtained using Alpha Miner Algorithm.

The alpha miner algorithm does not consider the frequencies of activities, so links between

activities are made two by two. The model becomes complicated and far from representing the actual behavior of the process.

The heuristicsmineR package provides the heuristics miner algorithm; this algorithm uses frequency to find the Petri net, which is an improvement of the alpha miner algorithm. The result from the precedence matrix with a dependency between activities relation of 0,8 is below:

	consequent									
antecedent	Approval	Creation	End	ERP Transfer	Invoice date	Match	Payment	Ready for transfer	Removal	Start
Approval	0	0	0	0	0	0	0	751	0	0
Creation	364	0	0	0	0	753	0	511	344	0
End	0	0	0	0	0	0	0	0	0	0
ERP Transfer	0	0	0	0	0	0	1628	0	0	0
Invoice date	0	1972	0	0	0	0	0	0	0	0
Match	387	0	0	0	0	0	0	366	0	0
Payment	0	0	1628	0	0	0	0	0	0	0
Ready for transfer	0	0	0	1628	0	0	0	0	0	0
Removal	0	0	344	0	0	0	0	0	0	0
Start	0	0	0	0	1972	0	0	0	0	0

> |

Figure 41: Efficient precedence matrix using heuristic miner.

The Petri net discovered is as follows below. Despite the filters, it is pretty extensive and presented only as an example; the application of the algorithms was not very successful.

Figure 42: Petri net obtained using heuristic miner.

The Petri nets created were not very explicit about the process, and there are not yet other process mining algorithms in bupaR available. Although, the bupaR package has another function that is interesting to see, which is an animated process map. Figure 43 is a snap-shot of the time where we can see the invoices going through the discovery process. In this case, we have all the companies' invoices, which are highlighted by color.

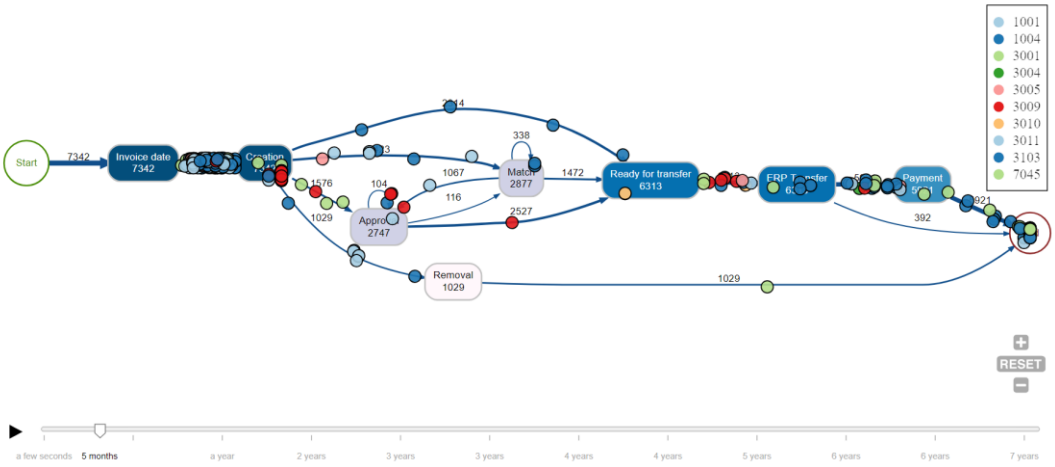


Figure 43: Animate process map, the companies in colors, and 95% of the invoices.

We get a simple drawing of the invoice process from the process map. Unfortunately, there is no indication of the GR or the type of match that would be more interesting. The only possible analysis in more detail is for invoices with or without PO and supposed 'goods received'.

Using the quality criteria listed in the Process Discovery in report Part I, it was decided to use the last criterion to choose the best process for analysis: simplicity. R calculates the other criteria, not automatically, but there is no doubt that using DFG is the best tool for this kind of process. Was created another DFG using the frequencies in percentages only for the company 1001, with the invoices of complete cycles and only with two types of invoices with PO and without PO, Figure 44.

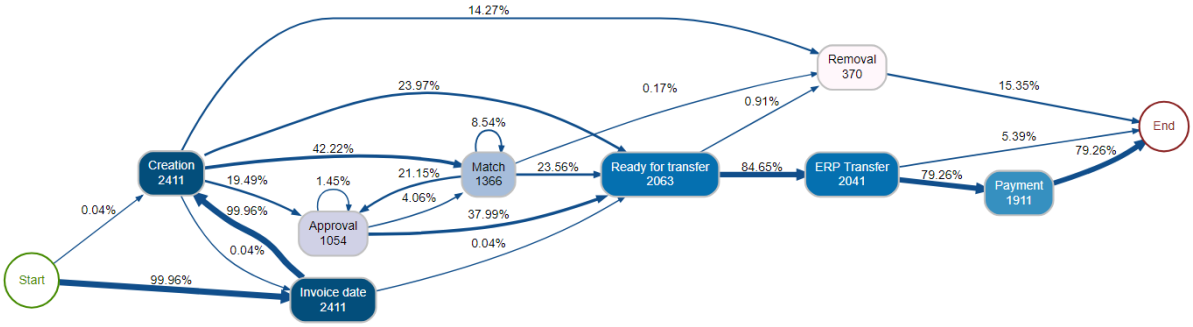


Figure 44: Direct-follow graph for all the invoices of the company 1001 received in one month. In the nodes, we can see the number of invoices and, in the arrows, their frequency.

The design of this process makes sense: most of the invoices start with their creation by the supplier ('Invoice date'), then the supplier sends it to the company's system ('Creation'), and then it has several options: it goes to be approved, it goes through manual matching, it is removed or, it goes directly to the 'Ready for transfer'. We can see that there are several loops in both the 'Approval' and 'Match' activities, and there are exchanges of invoices between them. Manual matching removed some invoices for some reason, but most go to the 'Ready for transfer'. The invoices leave the P2P system in the 'ERP transfer' activity and go to SAP or ORACLE for payment support and come back to the system with the payment date. It is possible to verify that, although the data collection has been done three months after the reception in the system, there are still 130 (5.39%) invoices that have not been paid. The process model ends as most of the invoices are paid.

Some invoices leave the 'Ready for transfer' activity for 'Removal', perhaps these invoices have had an intervention, but there is no record of this in the data.

For the 2411 invoices of company 1001, only one does not start as the others. After checking with the original data, the invoice date is after the date of receipt in the system, Figure 45. The model is working!

Invoice ID	Creation date	Invoice date
a896b7b715794d669958e35948016967	04/09/2020	05/09/2020

Figure 45: A rare case Invoice received date before the invoice date.

Out of curiosity, the respective model in BPMN is given in the figure below. It provides a graphical notation for specifying business processes, used widely by business managers and analysts.

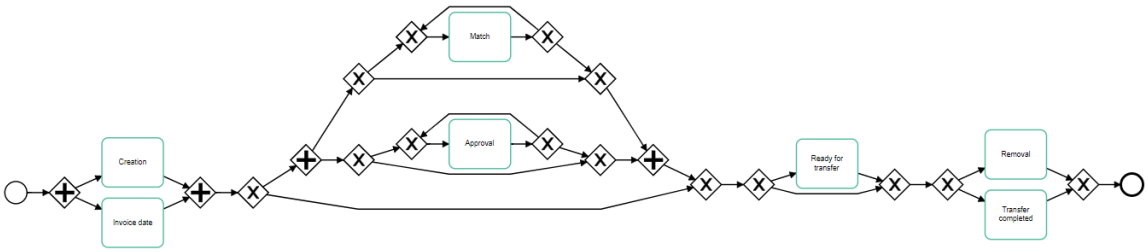


Figure 46: Designed model in BPMN.

The models were obtained by the available functions of bupaR and algorithms. Another program was used to build the BPMN model; however, there are new applications using R to generate BPMN models (process-analytics BPMN visualization in R, s.d.).

6.3. Process Performance

From a performance perspective, the performance map shows durations between activities. Allows the detection of the bottlenecks in the process. The time durations are indicated on the arcs and the number of invoices in the nodes. The performance map in Figure 47 presented durations using the median, but other configurations are possible (mean, maximum).

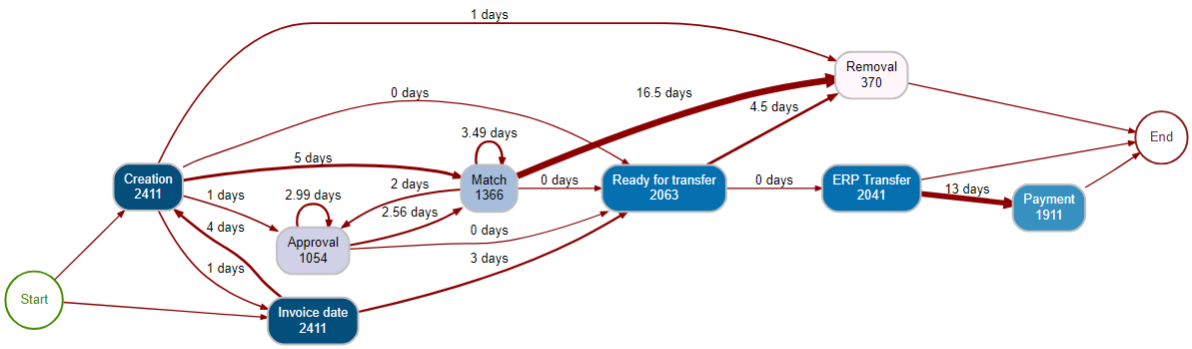


Figure 47: Performance map using median days.

Figure 48 is the process's most frequent sequence or trace, called the 'happy path'. The date of delivery of the invoice in the company system is the creation date. Therefore the activity named 'Creation', from this activity until the 'ERP Transfer', the duration is minimum, equal to zero. From the 'ERP Transfer,' the invoice leaves the system for payment using

other auxiliary business software. The company's goal is that the validation process of all invoices will have the trace and the duration of the happy path.

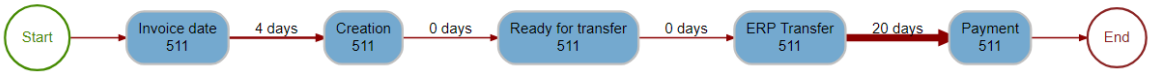


Figure 48: The 'happy path' with the median in days for performance.

Ideally, all invoices would make this path in the best possible number of days. If the invoices are automated, the process is faster and more efficient. However, it does not always work, and people are always involved in the process. In this event log, all the registration of the employee intervention is the FTE variable. As mentioned, the data obtained by the system only indicate the last user, so when an invoice has several times in the 'Match' or 'Approval' activity, we do not have several users involved; only the last one was recorded.

In order to know who executes the work, we can look at the resource labels, using the resource_labels function and frequencies_of_resources. Below, the list gives an absolute frequency per resource. The null values indicated the other activities that do not have resources associated.

```
> frequencies_of_resources
# A tibble: 52 x 3
   FTE    absolute relative
   <fct>     <int>    <dbl>
 1 NA         9332  0.793
 2 M-16        646  0.0549
 3 M-13        460  0.0391
 4 A-57        254  0.0216
 5 M-7         202  0.0172
 6 A-166       186  0.0158
 7 A-141        94  0.00799
 8 A-218        93  0.00791
 9 M-12         65  0.00553
10 A-30         48  0.00408
# ... with 42 more rows
>
```

Figure 49: Frequencies of resources.

The resource M-16 is the employee last register who realized more manual matching, and the A-57 is the approver last register who approved more invoices; this is only for the company 1001. Another aspect is the rework, where some activities are done several times for the same case, often a source of inefficiencies and waste.

From an organizational perspective, Figure 51 shows a process map of 80%. For example, M-16 has the highest number of invoices processed, a total of 644, of which 159 underwent a matching rework, 62 were passed on to approver A-166, and seven were received from approver A-218. The data available are not the best example of applying resource maps, but it still is possible to see that there is a communication network and see where an invoice passed.

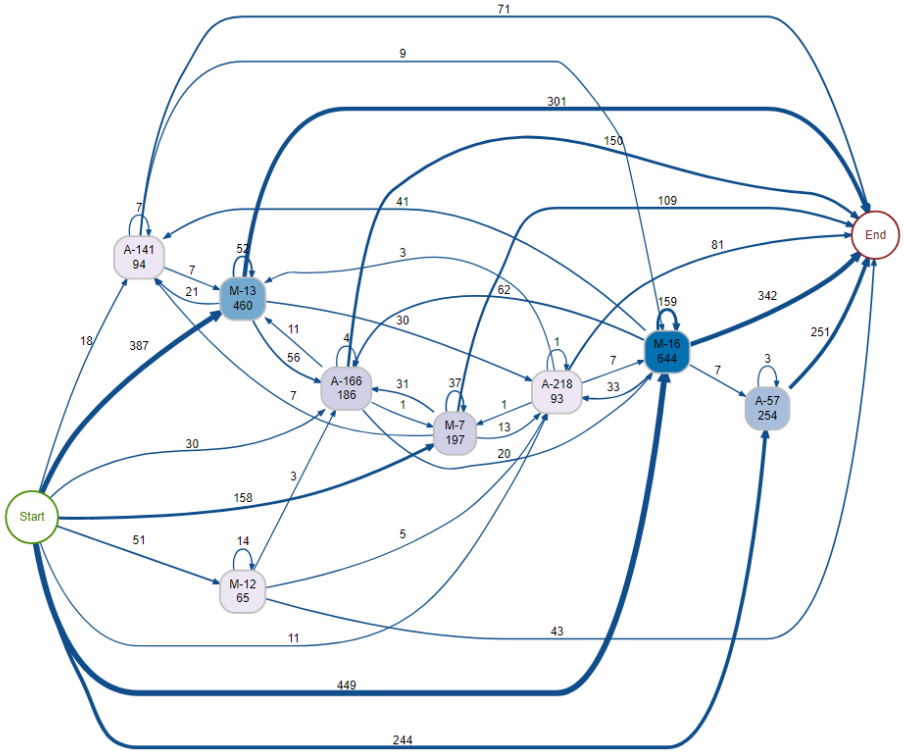


Figure 50: Resource map with 80% of the invoices

For instance, the invoice with PO with InvoiceID = 'b99cdf0bf63445f99e2b60a9602923b4' was received in one day, but between reception in the system and the manual matching took four days. In total, the invoice took 27 days or 20 working days from its creation to payment. There are not many like it either, but it is just to show the possibilities of this application.

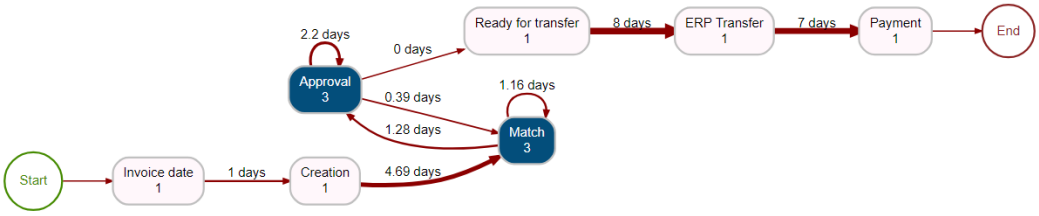


Figure 51: An example of an invoice path.

With the 'Recipient' attribute, as mentioned, It was assumed that the receipt of the goods gets recorded. In our example, there are 1839 (76%) invoices with PO and 572 (24%) invoices without PO; basically, half of the invoices have GR. Studying the invoices' path is essential since they depend on whether to wait for the GR.

	Invoice type		
GR	Vend.Inv. with PO	Vend.Inv w/o PO	Grand Total
yes	34.21%	16.59%	50.80%
no	41.83%	7.37%	49.20%
Grand Total	76.03%	23.97%	100.00%

Figure 52: Table with the percentage of Invoices with or without PO and with or without GR.

The invoices with PO can be processed automatically, while the others must go through the LT that approves the invoice data. Therefore, the path of the invoices will be different. The indication of the GR could improve the analysis to make a payment. It is necessary to wait for the product to arrive at its destination.

From Figure 53, invoices with PO take longer to process, even without GR. The median for the invoices is 30, 21, 7, and 1 day to finish the cycle, respectively. The time decreases slightly by excluding the removed ones.

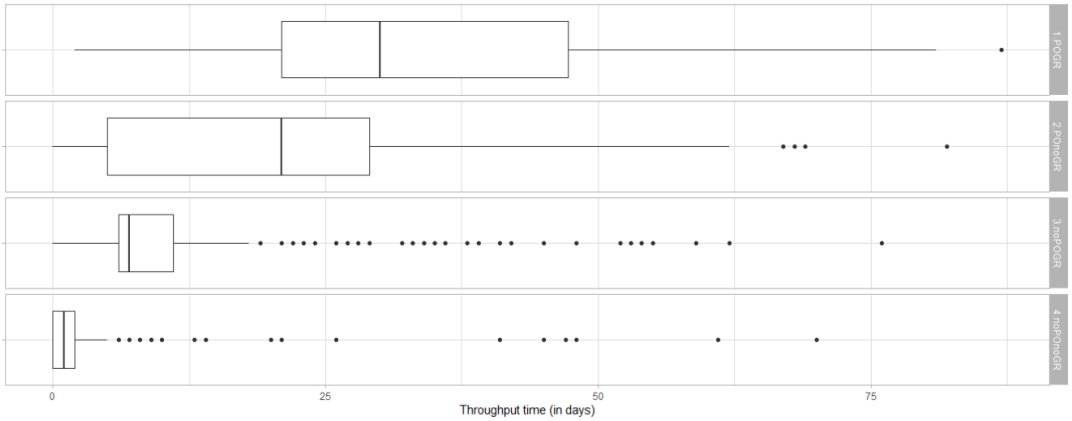


Figure 53: Boxplot comparing time performance of the invoices whit and without PO and with and without GR, from 'Creation' activity to the end of the process.

The following sections analyze these four invoice scenarios, thus obtaining different paths, only taking the example for company 1001; it can easily be extended to any other.

6.3.1. Invoices with PO and GR

The first case is the invoices with PO and GR, i.e., a PO was sent to the supplier and registered in the system. The supplier, in response, sends the GR and the invoice. Then either must wait for the goods to be received or not.

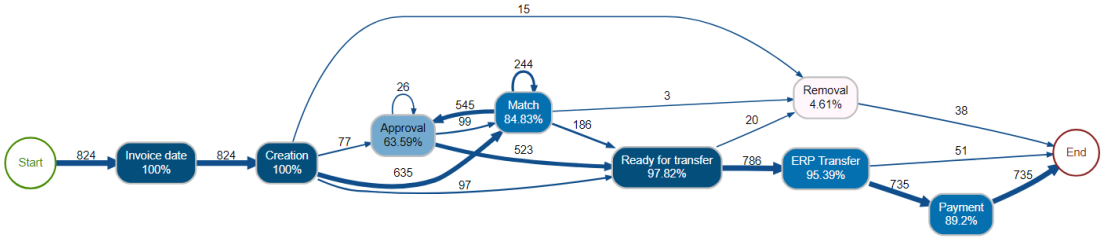


Figure 54: Process map of invoices whit PO and GR.

From Figure 54, we can see that practically all the invoices pass through the teams; 64% of the invoices passed through 'Approval' and 84.83% through manual matching. Only 97 invoices, 12%, go directly to the 'Ready for transfer' activity. However, 12 of these were removed (Figure 55), and 244 (30%) invoices suffered rework, going through several users, probably with errors that made the invoice matching with the PO difficult.

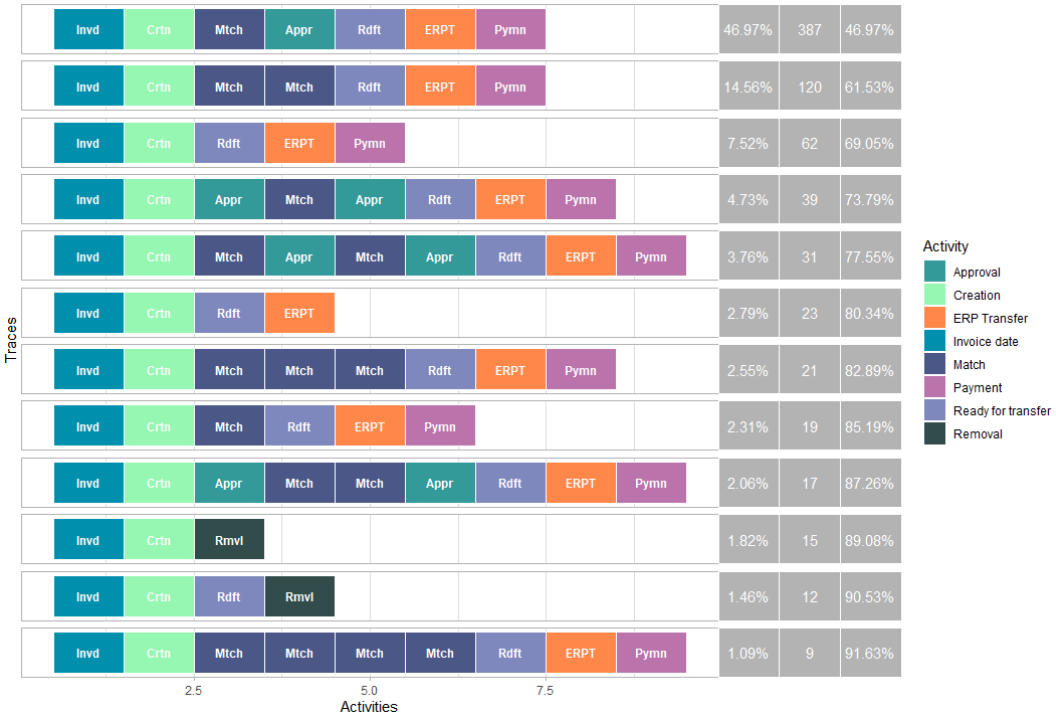


Figure 55: The different paths of 80% of invoices with PO and GR.

In Figure 55, we have 80% of the paths covered in the process, thus not considering the less frequent cases. Only 7,52% plus 2,79%, the third trace and the sixth trace, the invoices followed without any intervention. The only difference is that the invoices have not been paid yet in this last trace. There was still a tiny percentage of invoices that went directly to the approval team. These cases may have happened by mistake or the team's lack of PO insertion in the system. 4.61% of invoices were removed, the majority after manual matching and approval.

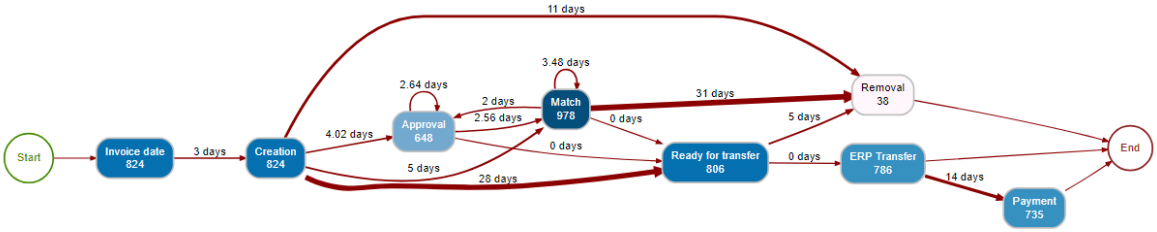


Figure 56: Invoice performance with PO and GR, median day duration.

Figure 56 shows the median of the invoice's removal duration. Even after the 'Creation' activity, the median was 11 days and 31 days for invoices removed that went from hand to hand in manual matching and approval.

6.3.2. Invoices with PO and without GR

Invoices with PO and without GR have great potential to be automatic because they do not have to wait for the GR. Indeed, more than 45% entered and left the system without intervention; they made the path 'Creation', 'Ready for transfer', 'ERP transfer', Figure 57. The performance map showed that these invoices took a median of zero days to remain in the system, which is probably an indicator that they were automated.

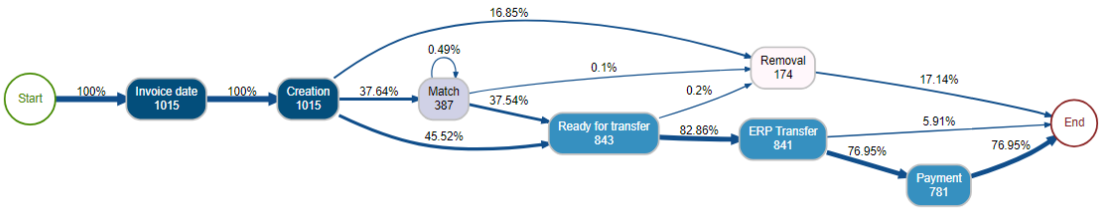


Figure 57: Process map of invoices whit PO and no GR.

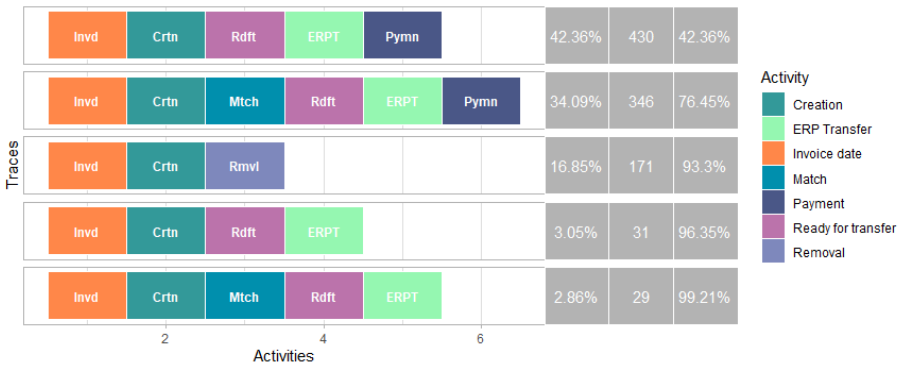


Figure 58: The different paths of 99% of invoices with PO and no GR.

The rework time for the manual matching activity is a median of 7 days, which needs to be improved for the invoice cycle performance.

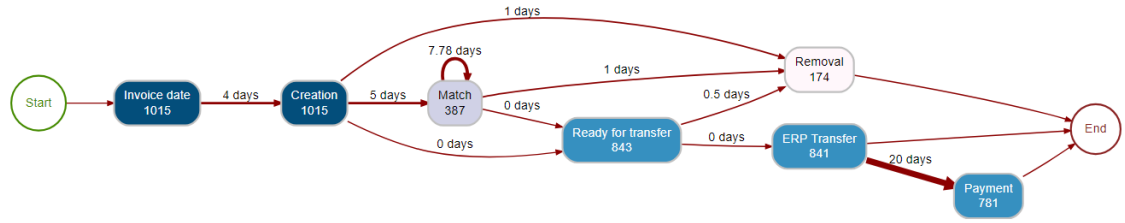


Figure 59: Invoice performance with PO and no GR, median day duration.

6.3.3. Invoices without PO and with GR

These invoices have no PO but with GR and need to be approved. Of them, 97% were analyzed by the approvers' team, except for one that may have been a data quality error and went to the manual matching team. However, four did not go through approval; these may be cases without recorder approvers.

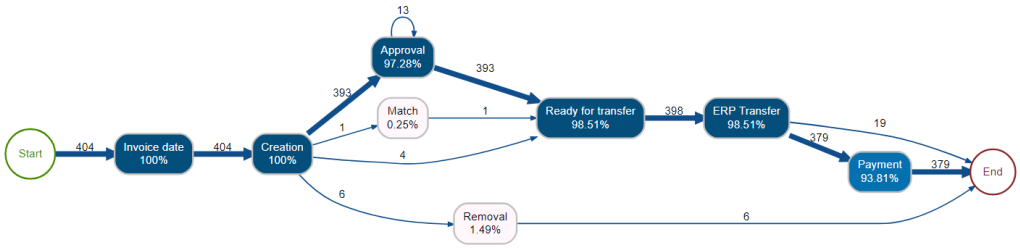


Figure 60: Process map of invoices without PO and with GR.

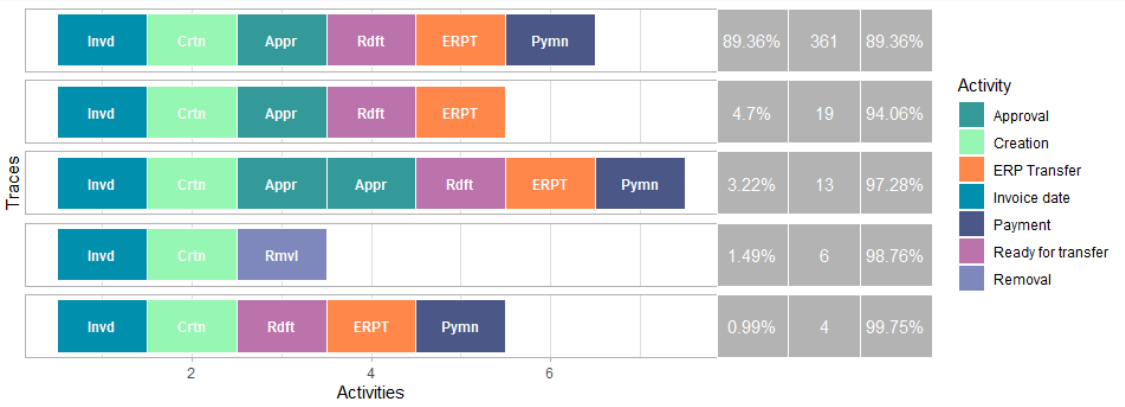


Figure 61: The different paths for 99% of invoices without PO and GR.

From the performance map in Figure 62, it is possible to see that invoices from 'Creation' to 'Ready for transfer' take a median of 33 days, perhaps waiting for the goods to be received. The duration to approve the invoices has a median of 9 days.

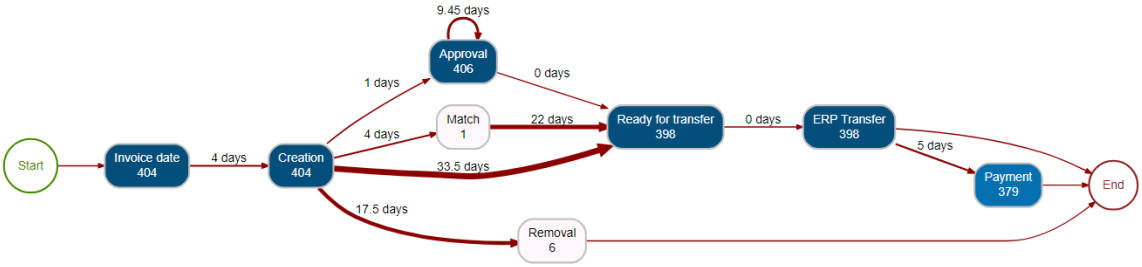


Figure 62: Invoice performance without PO and with GR, median day duration.

6.3.4. Invoices without PO and GR

Invoices without PO and GR may be services, but they must go through approval for validation. From Figure 63, 90% of the invoices were removed, but 16 were paid without intervention. These invoices seemed to have no intervention and were probably automatic. Even the removed ones were made with a median of one day.

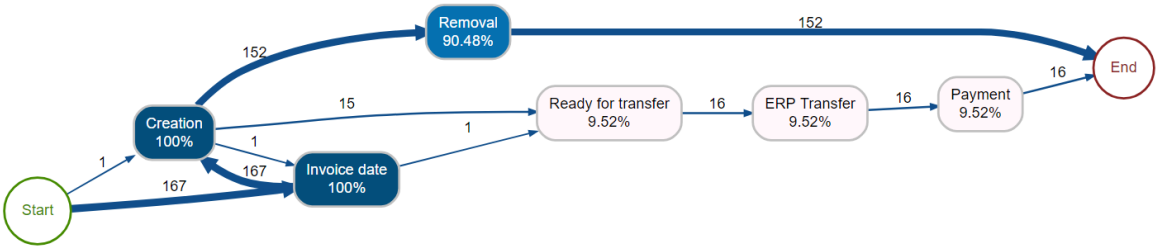


Figure 63: Process map of invoices without PO and GR.

In this situation, only three possible traces were obtained.

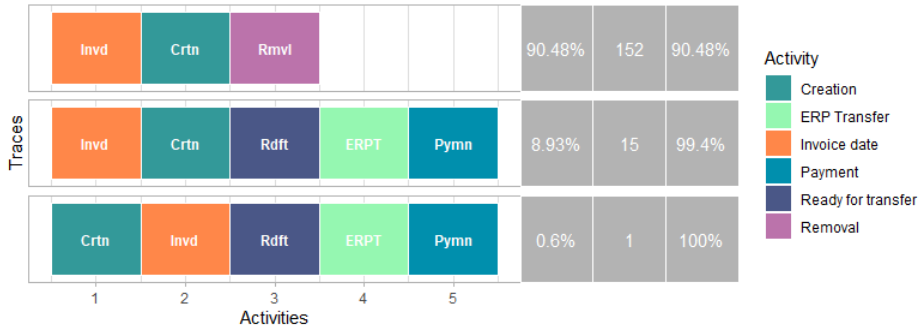


Figure 64: The different paths of all the invoices without PO and GR.

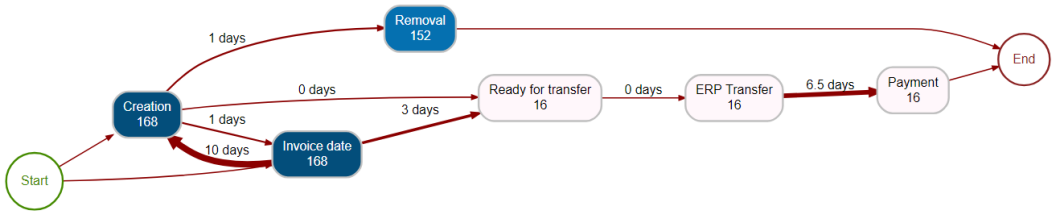


Figure 65: Invoice performance without PO and GR, median day duration.

6.4. Conformance Checking

Conformance checking aims to analyze the relationship between observed and desirable process behavior. It can repair models, evaluate process discovery algorithms, or use rules as connections between activities. Rules can be derived from process models as constraints given by the model's control flow. (Kiarash Diba, 2019) The conformance can be measured by counting the violation and satisfaction of the order of the intended activities.

In Figure 66, it is possible to check the relative-precedent type of precedence matrix, which can also be calculated with absolute, relative, or relative-consequent frequency values. The relative-precedent presents relative frequencies within each antecedent, i.e., showing the relative proportion of consequents within each antecedent, i.e., antecedent 'Approval' is followed 3.7% of the time by consequent 'Approval', 9.39% by manual 'Match' and 86.91% by 'Ready for transfer'. There is some rework: 18.23% of manual 'Match' has the same activity as the consequent.

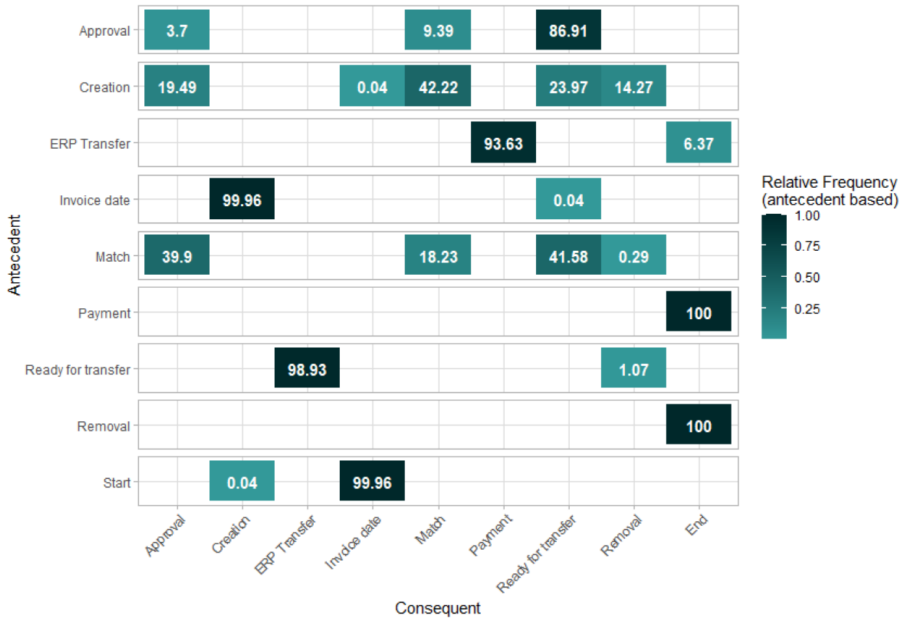


Figure 66: The precedence matrix shows the flows from one activity to another in relative frequency.

The desired path would be 'Invoice date', 'Creation', 'Ready for transfer', 'ERP Transfer', and 'Payment'. After the 'Creation' activity, 19,5% of the invoices goest to 'Approval', 42% to manual 'Match', 14% to 'Removal', and only 24% to 'Ready for transfer'. Therefore, 76 % of the invoices failed to pursue the 'happy path'.

Rules are defined for a pair of activities, and because this example has many reworks, it seems like using the trace as a study is more appropriate than only activities. The idea is to apply conformance checking to analyze *Worst* and *Best in Class*, understand the traces of the rejected invoices, and find the trace of the automatically processed ones.

6.4.1. Duplicated

Duplicate invoices prevent streamlined processing; their impact is much more significant than just within the AP department. Many of these departments, especially in large companies, have people working especially on recovering amounts resulting from duplicate payments and may use some computer software to help avoid this complication. The software uses machine learning to detect duplicate invoices and find group patterns such as similar values, suppliers, and references. For example, the same invoice may appear twice, one with a standard US date and another with an EU format, if no one notices it is paid twice.

Figure 67: Duplicated invoices.

From Figure 67, the median value between the invoice date and its reception is 95 days. This high value gives the idea that the suppliers probably forwarded the invoice because they did not get an answer to the first sending. Maybe promote invoices update or give invoice receipt and inform them how long expect it to take payment.

In the data, there was a 'Duplicated' indicator where the system marked an invoice to reject it. However, this only happened in the 'Creation' activity when the invoice was received in the system. So, there may be duplicate invoices that were removed by employees later in the process. It seems better only to consider the rejected ones for study, whether they were duplicate invoices or any other error.

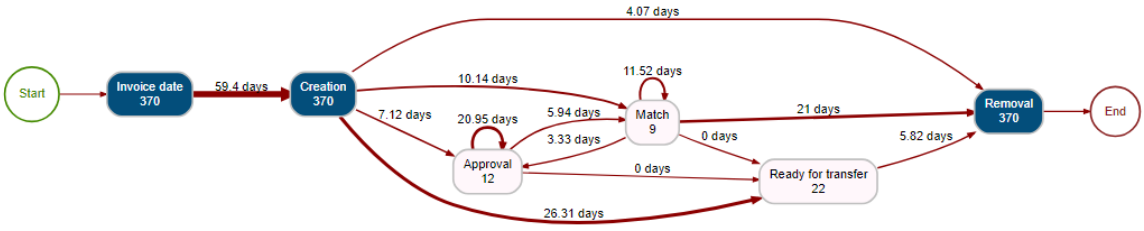


Figure 68: Performance maps of rejected invoices.

For this company, 15% (370/2411) of the invoices received were rejected, representing more than 15 million euros, and only for invoices with and without PO. Of the rejected invoices, nine went through the manual matching and twelve through the approval activity.

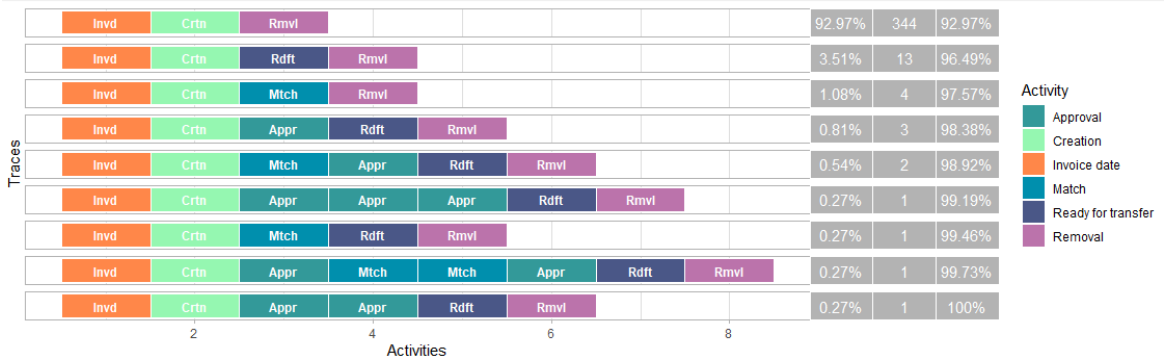


Figure 69: Traces of rejected invoices.

The first trace has a median of 1 day to remove the invoice, but the second already has 35 days. However, the majority were detected early on; of the 344 invoices, 128 were marked as duplicates, almost 40%.

Rejection Rate

Supplier code	Number of Rejections	Number of invoices	Rejection Rate	Max. Removal duration, days	Inter Company	Gross Eur	
30090000	52	113	46.02%	44	IC	2,194,246.36	DE
10040000	40	40	100.00%	14	IC	55,276.26	CA
5022680	37	103	35.92%	41	not IC	38,215.52	US
5003363	16	17	94.12%	17	not IC	36,608.95	US
5000201	14	22	63.64%	0	not IC	13,366.45	US
5003397	10	11	90.91%	1	not IC	8,677.01	US
7000259	9	229	3.93%	22	not IC	55,383.31	US
5018714	7	7	100.00%	5	not IC	Null	US
7000230	7	12	58.33%	6	not IC	4,458.02	US
5025069	5	5	100.00%	0	not IC	1,429.06	US
5025294	5	8	62.50%	1	not IC	36,378.56	US
5011232	4	7	57.14%	8	not IC	4,696.00	US
7000361	4	6	66.67%	1	not IC	13,494.60	US
5000036	3	37	8.11%	1	not IC	74,755.65	US
5000498	3	3	100.00%	20	not IC	1,976,403.19	US
5002995	3	11	27.27%	43	not IC	7,627.13	US
5007757	3	8	37.50%	2	not IC	360,424.28	KR
5011013	3	5	60.00%	42	not IC	6,227.91	US
5012216	3	5	60.00%	1	not IC	12,047.77	US
5022657	3	4	75.00%	3	not IC	130,475.00	US

Figure 70: Table with rejection rate suppliers list.

In Figure 70, it is possible to check the suppliers with the highest number of rejected invoices, the rejection rate, and the amounts involved. The first two suppliers with the highest number of rejected invoices are inter-company. Another example is supplier 5022680; in one month, sent 103 invoices, and 36% were rejected, which involves a total of 38 thousand euros.

6.4.2. Touchless

Organizations expect that purchases will be delivered and paid for on time. Making the process automatic is a goal, and finding good examples is one way to measure company

performance. There was no information in the system on which invoices had been processed automatically. The time of permanence in the system must be less than one day from 'Creation' to 'ERP transfer' activity.

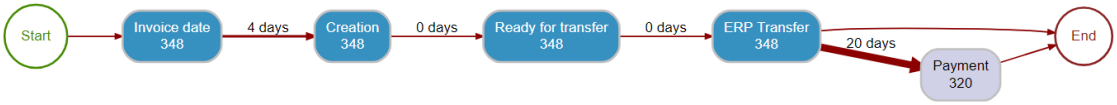


Figure 71: Touchless performance map.

As a result, 348 touchless invoices were obtained, corresponding to 14.4%, a little less than the 17.8% given by the OLAP analysis.

Touchless Rate

Supplier code	Number of Touchless	Number of invoices	Touchless Rate	Inter Company	Gross Eur	
7000259	68	220	30.91%	not IC	52,904.89	US
5013234	35	49	71.43%	not IC	129,772.56	US
5000036	19	33	57.58%	not IC	72,242.01	US
7000079	17	27	62.96%	not IC	4,931.51	US
5000604	12	25	48.00%	not IC	87,448.79	US
5002292	12	57	21.05%	not IC	138,322.57	US
5002820	12	22	54.55%	not IC	142,416.78	US
5002044	7	25	28.00%	not IC	46,945.65	US
5002099	7	18	38.89%	not IC	78,183.04	US
5015619	7	24	29.17%	not IC	52,527.50	US
7000415	7	25	28.00%	not IC	755,314.39	US
7000526	6	19	31.58%	not IC	6,003.13	US
5000013	5	26	19.23%	not IC	231,396.68	US
5000037	4	7	57.14%	not IC	14,336.09	US
7000325	4	8	50.00%	not IC	23,058.97	US
5001117	3	14	21.43%	not IC	311,380.24	US
5020926	3	9	33.33%	not IC	4,463.41	US
5000101	2	2	100.00%	not IC	1,086.29	US
5000262	2	4	50.00%	not IC	39,708.15	US
5000397	2	3	66.67%	not IC	43,228.65	US
5001134	2	2	100.00%	not IC	9,092.38	US
5001191	2	2	100.00%	not IC	1,662.56	CH
5001798	2	2	100.00%	not IC	5,847.66	US

Figure 72: Table with touchless rate suppliers list.

A list of the suppliers with the highest number of touchless invoices is calculated, with the respective touchless rate and the associated amounts (none is inter-company).

Finally, to find the invoices with the potential to be touchless by finding the number of invoices with PO that could be processed automatically but went through the manual matching without approval, and removed the rejected ones, gave 32% (786) of the invoices with an associated value of more than 30 million euros.

7. Results

The report's purpose was to apply the mining process to find problems associated with the AP department's P2P process. The process was unknown, which can change from company to company, but the preliminary OLAP analysis helped to understand. Several filters were applied to the data, and the results only concern one month, one company, and two types of invoices. The summary of all the information obtained from the data provided is:

- In one month, 12,615 invoices were received, with a median of 1,240 per day. Company 1001 had the highest number, with 2,500 invoices.
- For the 1001 company, the median time for receipt of the invoice was three days, but there are many suppliers with significant differences in dates. These differences could be due to delays in the process, and the supplier sends back the same invoice several times to get an answer from the company. Solving the problem prevents employees from unnecessary work to deal with duplicate invoices and the risk of paying twice. Maybe it is essential to improve the relationship with the supplier: make it clear how to make the process more efficient, a clear and straightforward user's guide, adapted to each country and type of supplier, or send update payment messages.
- Inter-company suppliers are not a good example of performance compared to other suppliers.
- An organization perspective analysis is done, but it is not explored because there are some doubts since data only indicate the last user.
- From the Pareto chart, we saw that 80% of the total gross amount came from 7% of the suppliers, and was identified the supplies with the highest amounts, the highest rejection rates, and the highest number of touchless.
- The processing time for invoices with PO associated has a higher median number of days than those without PO. It is the most frequent type of invoice and with a chance of being faster processed and perhaps even automatically.
- For the same company, in one month, 15% of invoices were rejected, 14% were automatically processed, and 32% had the potential to be handled automatically.

8. Conclusions and Final Thoughts

The idea of applying process mining is to obtain a model process from observed and recorded data. Just by having a recorded date, an activity, and a unique identifier for each case, it is possible to draw the trajectory of the dataset, whether it is a process in the industry, transport, accounting, tax, audit, hospital, aviation, or customer service. The data in the report is about a P2P process used by an AP department to follow the invoice process.

Process discovery algorithms deal with noise by removing infrequent behavior or ignoring incomplete cases. The alpha miner is the oldest, simplest, and most studied, and the heuristic miner is an improvement since it considers the frequency of activities and can handle loops. Both were applied. Some criteria to evaluate algorithm application are fitness, precision, generalization, and simplicity. The best choice would be a balance between the criteria and the actual behavior of the data. However, in this case, the model chosen was a direct-follow graph (DFG) application. This elementary function links one activity to another and sorts in chronological order, which was more successful than the algorithms.

Conformance checking compares the model's behavior with the observer's behavior to find deviations. The report analyzes the best possible behavior as non-contact invoices (touchless) and the worst possible behavior, which is rejected/duplicated invoices.

The tool used for process mining, the bupaR package, does not have many algorithms available to apply process discovery, and the results of the existing algorithms were not satisfactory. The academic software alternative would be to use ProM. This software would have more algorithms, including the fuzzy algorithm, one of the most efficiently used. It is not presented comparisons with other software. Still, there are clear signs that it is necessary to develop more than what is available in R. However, it is a good alternative for companies working with bupaR than some commercial software (a surprising cost). It is an excellent alternative to ProM, which according to a colleague's thesis (Silva L. F., 2014), was not worth the effort of learning the tool for such meager results. R is free and open source, allowing integration with other tools, and anyone can develop and adapt to each situation.

The bupaR package also handles XES files, which event logs are often stored, but it is not a necessary condition. The XES stands for Extensible Event Stream and is a standard adopted by the IEEE Task Force on Process Mining, conversion from other formats (CSV) is easy if the correct data are available.

When moving from data analysis to process mining analysis, it was necessary to transform the data and get more than five times more rows of data which can be pretty challenging for an extensive database. If there were more information about the process, such as the GR date, change price, disputes with suppliers, or the names of all employees, this model would be much more complicated to analyze but more attractive.

It was intended to integrate the results into Tableau. It is possible to do data mining using R in Tableau. However, it was not possible to apply for process mining, maybe due to the limitation of the student license. There are tutorials on applying the process mining functions of R in PowerBI. The significant advantage of this integration is the ease of using filters, visualization, and exploring without any code (Business Process Analysis in PowerBI using R visuals, s.d.).

This report was inspired by the BPI challenge 2019. Since 2011 the IEEE (Institute of Electrical and Electronics Engineers) Task Force on Process Mining has organized the Business Process Intelligence Challenge (BPI challenge), similar to machine learning contests in Kaggle. Participants receive a record of real-life events from a company, the process owner suggests a specific scope for analysis, and the goal is to gain insights from the process. Data analytics, data mining, process mining, and predictive analytics techniques can be used. The BPI Challenge aimed to allow researchers and practitioners to showcase their work (Conference, 2019).

I believe process mining can be a great resource and a complement for data analysis, adding the knowledge of machine learning. Proof of that is the amount of commercial software that has emerged for this theme. Process mining gives 'a helicopter view' to the process. Van der Aalst (Aalst W. v., Process Mining, Data Science in Action, 2016), suggests using process mining to improve predictive analysis. The next step of this report would be to apply predictive analytics to the rejected invoices using process mining knowledge and compare the performance with the application of data mining techniques.

Appendix I - Attributes available

The data was obtained from Basware, the kp Oporto software, to handle purchase orders and invoice processing. Four linked sources were obtained and identified the key attributes, even if some were not used in the report's analyses.

Data Source	Module / Functional Area	Description
InvoiceHeader	AP Financial, AP KPI, AP Process, Spend	Contains header level invoicing data. The user rights are module-specific. The data source can be linked with *InvoiceCodingRowMain* and *InvoiceCodingRowCustom* using the *Invoice ID* field.
CodingRowPurchaseOrderLine	Spend	Contains all purchase orders and respective lines brought in P2P for matching. Can be joined to coding row based data sources using the *ORDER_ROW_ID* field.
InvoiceCodingRowMain	Spend	Contains standard coding row level invoicing data. The data source can be linked with *InvoiceHeader* using the *Invoice ID* field and with *InvoiceCodingRowCustom* using the *CODING_ROW_ID* field.
InvoiceCodingRowCustom	Spend	Contains customizable coding row level invoicing data. The data source can be linked with *InvoiceHeader* using the *Invoice ID* field and with *InvoiceCodingRowMain* using the *CODING_ROW_ID* field.

Figure 73: Available database, source Basware help center.

The process starts with LT sending purchase orders (PO) to the supplier, putting them into the system, and recording the arrival of the products line by line - recorded in the CodingRowPurchaseOrderLine source. The invoice sent by the supplier is registered in InvoiceCodingRowMain (SAP ERP) or InvoiceCodingRowCustom (JDE ERP). As soon as the invoice is matched with the PO, it goes to the source InvoiceHeather by the system.

Invoice Header	Variable	Type	Description
Organization	Organization	Categorical	Organization unit name
	Company code	Numeric	Organization identification code
	Country of organization	Categorical	Organization unit´s country
	Organization level 2	Categorical	ERP used by the company, SAP or JDE
Supplier	Supplier	Categorical	Supplier´s name
	Supplier code	Numeric	Supplier´s identification code
	Supplier Country	Categorical	Supplier's country
Invoice	Invoice ID	Numeric	Unique identifier of invoice
	Invoice number	Numeric	Invoice reference number
	Invoice Type	Categorical	Credit memo, invoice w/ purchase order, invoice w/o PO
	Invoice Status	Categorical	Invoice status in Basware P2P
	Invoice Sub-Status	Categorical	Second level status of an invoice
	Operation Status	Categorical	Operation Status, change over time
	Disputed	Categorical	Shows if the invoice has been disputed at least once
	Association Type	Categorical	Describes if the invoice is based on a purchase order or not
	N. of Coding rows	Numeric	Number of coding rows
	Approver	Categorical	Invoice Header Approver (name)
	Recipient	Categorical	Invoice in approval (name)
	Header approval count	Numeric	Header approval count
	Header approval duration, d	Numeric	The total duration of header approval
	Reference Person	Categorical	Invoice reference person

	Supplier VAT Reg	Categorical	Supplier VAT Reg
	Purchase order number	Numeric	Purchase order number linked to the invoice header
	PO Originator	Categorical	Purchase Order originator (name)
	Payment block code	Numeric	Code, change over time
	Reason for failed validation	Categorical	Invoice validation in the reception
	Gross Total	Numeric	Gross total in invoice currency
	Currency Code	Numeric	Currency code of invoice total in invoice currency
	Gross Euro	Numeric	Gross total in euros, attribute created
	Net total	Numeric	Gross total without taxes
	Tax total	Numeric	Taxes
For Matching	Matching Status	Categorical	The outcome of matching activity on the invoice
	Matching Sub-Status	Categorical	Matching sub-status of an invoice
	Matching Type	Categorical	How invoices have been matched
	Manual Matching by	Categorical	Name of the person who has done the manual matching
	Manual matching count	Numeric	Number of manual matching tasks
	Manual matching duration, d	Numeric	The total duration of manual matching
For Discount	Supplier payment term code	Numeric	Supplier´s payment term code
	Payment term code	Categorical	Payment term in code format
	Default payment term	Categorical	Invoice payment term in the text
Dates	Invoice date	Date	Invoicing date
	Invoice cash date	Date	Date when invoice´s cash discount expires
	Due date	Date	Invoice due date
	Creation date	Date	Invoice creation time in Basware
	Date send to process	Date	Invoices send to process date
	Date matched	Date	Date matched
	Approval time	Date	Timestamp when the last invoice header approval task has been completed
	Ready for Transfer date	Date	The time when the invoice is ready for transfer to an ERP system
	Transfer completed time	Date	Date when the invoice transfer has been completed
	Payment date	Date	The time when the invoice was paid, returned to Basware P2P from the banking system, found some errors.
	G/L date	Date	Bookkeeping date
	Removal date	Date	Invoice removal date
	Voucher date	Date	Not working

Coding Row Purchase Order Line	**Variable**	**Type**	**Description**
	Order ID	Numeric	Unique identifier of the purchase order
	Order_row_id	Numeric	Unique identifier of the purchase order row
	Creation Time (order)	Date	Date of the purchase creation in the Basware system
	Goods receipt usage	Categorical	If there is GR or not, it does not work.
	Order Quantity	Numeric	Quantity order by the company
	Received quantity	Numeric	Quantity received
	First delivery date	Date	Date of the delivery of the first item in the invoice
	Last delivery date	Date	Date of the delivery of the last item in the invoice

Invoice Coding Row	Variable	Type	Description
	Account name	Categorical	Account name
	Account code	Numeric	Account code
	Cost center name	Categorical	Cost center
	Cost center code	Numeric	Cost center

Appendix II – R programming code

```
##################################################
################## Process Mining ################
#################### bupaR.net ###################
##################################################
setwd("~/Data mining Master/Tese/R Results")
#getwd()
#install.packages(c('bupaR', 'eventdataR', 'xesreadR', 'edeaR', 'processmapR', 'processmonitR', 'pm4py'))
library(readxl); library(bupaR); library(processmapR); library(processmonitR); library(dplyr); library(processanimateR); library(eventdataR); library(heuristicsmineR); library(pm4py); library(petrinetR); library(daqapo)
# pipe operator %>%
Book1 <- read_excel("pm.xlsx")
# Inspect the structure of the data
# str(Book1) or
summary(Book1)
 # To create the event log object
invoices <- eventlog(
 Book1,
 case_id = "InvoiceID",
 activity_id = "activity",
 activity_instance_id = "order",
 lifecycle_id = "status",
 timestamp = "date",
 resource_id = "FTE"
)
mapping(invoices)
# Total number of invoices
n_cases(invoices)
# Summary of the data
summary(invoices)
# Show the journey of the first invoice
slice(invoices, 1)
# Number of distinct activities
n_activities(invoices)
# The names of the activities
activity_labels(invoices)
# List of activities
# or activity_dashboard(invoices)
activities(invoices)
# A sequence of activities describes each case. It is called a trace.
traces(invoices)
# Different number of traces
n_traces(invoices)
#The percentage coverage of the trace to explore. Default is 20% most (in)frequent
trace_explorer(invoices, coverage = 0.80)
 #Which activities are always done for an invoice and which are rare?
#Which activity type occurred in the least number of cases?
invoices %>%  activity_presence() %>%  plot()
 ##########################################################
################ Process Discovery ######################
##########################################################
################### Heuristics Miner #####################
##########################################################
# Dependency graph/matrix
freqinvoices <- invoices %>%
 filter(`Company code`==1001) %>%
 filter_trace_frequency(percentage = 0.8)
dependency_matrix(freqinvoices) %>% render_dependency_matrix()
# Causal graph / Heuristics net
```

```
causal_net(freqinvoices) %>% render_causal_net()
# Efficient precedence matrix
m <- precedence_matrix_absolute(freqinvoices)
as.matrix(m)
dependency_matrix(freqinvoices, threshold = 0.8) %>% render_dependency_matrix()
causal_net(freqinvoices,threshold = 0.8) %>% render_causal_net()
# Convert to Petri net
cn <- causal_net(freqinvoices, threshold = 0.8)
pn <- as.petrinet(cn)
render_PN(pn)
############################################################
###################### Alpha Miner ########################
############################################################
discovery_alpha(freqinvoices) -> PN
PN %>% str
PN$petrinet %>% render_PN()
discovery_alpha(freqinvoices, variant = variant_alpha_plus()) -> PN
PN$petrinet %>% render_PN()
############################################################
#################### Inductive Miner ######################
############################################################
discovery_inductive(freqinvoices, variant = variant_inductive_imdfb()) -> PN
PN %>% str
PN$petrinet %>% render_PN()
############################################################
################### Direct Follow Graph ###################
############################################################
#To visualize processes using a process map.
invoices %>%
 filter(`Company code`==1001) %>%
  filter_trace_frequency(percentage = 0.95) %>%
  process_map(
    # type_edges = frequency("relative-case"),
    #type_edges = performance(FUN = median, units = "days")
  )
  # process map of median performance in hours
invoices %>%
  filter(`Company code`==1001) %>%
  process_map(type = performance(FUN = median, units = 'hours'))
  # Select top 20% of cases according to trace frequency
happy_path <- invoices %>%
  filter(`Company code`==1001) %>%
  filter_trace_frequency(percentage = 0.2)
  # Process map of absolute case frequency
happy_path %>%
  process_map(
    #type_nodes = frequency("relative-case"),
    type_edges = performance(FUN = median, units = "days")
  )

  happy_path %>%
  filter(`Company code`==1001) %>%
  throughput_time(units = "days")

invoices %>%
  filter(`Company code`==1001) %>%
  filter_activity("Invoice date",T)%>%
  group_by(InvoiceType)%>%
  throughput_time(units = "days")%>%
plot()

# Animate process map
invoices %>%
  filter(`Company code`==1001) %>%
  filter_trace_frequency(percentage = 0.95 ) %>%
  animate_process(
    #mapping = token_aes(color = token_scale("red"))
  )

invoices %>%
filter_trace_frequency(percentage = 0.95 ) %>%
animate_process( mode = "relative", jitter = 10, legend = "color",
                 mapping = token_aes(color = token_scale("Company code",
                                                         scale = "ordinal",
```

```
                                        range = RColorBrewer::brewer.pal(7, "Paired"))))

# Filter for the 80% of most common activities
invoices %>%
  filter_activity_frequency(percentage = 0.8) %>%
  process_map(
    type_nodes = frequency(),
    type_edges = performance(FUN = median, units = "days")
  )

#Filtering by activities of the process
invoices %>%
  filter_trace_frequency(percentage = 0.6) %>%
  process_map(
    type_nodes = frequency(),
    type_edges = performance(median, "days")
  )

#One anomaly - Invoice date after creation date
anomalie <- invoices %>%
  filter_precedence(
    #    antecedents = "activity that came before",
    antecedents = "Creation",
    #    consequents = "activity that came after",
    consequents = "Invoice date",
    #    precedence_type = c("directly_follows", "eventually_follows"),
    #    filter_method = c("all", "one_of", "none"),
    filter_method = "all",
  )
anomalie$InvoiceID

invoices %>%
  filter(`Company code`== 1001) %>%
  filter( is.na (CountrySupplier)== F) %>%
  group_by(CountrySupplier)%>%
  throughput_time(units = "days") %>%
  plot()
############################################################
#############  Organizational analysis #################
############################################################
#who executes the work
resource_labels(invoices)
#resources frequencies
resource_frequency(invoices)
#removing blank FTE
invoicesFTE<-filter_resource(invoices,NA,reverse = TRUE)
resource_labels(invoicesFTE)

invoicesFTE %>%
  filter(`Company code`==1001) %>%
  filter_resource_frequency(percentage = 0.80) %>%
  resource_map()


#The precedence matrix shows the flows from one activity to another
#in a rectangular format.
invoices %>%
  filter(`Company code`==1001) %>%
  precedence_matrix(type = "relative-antecedent") %>% plot()

invoicesFTE %>%
  filter(InvoiceID =="b99cdf0bf63445f99e2b60a9602923b4" ) %>%
  resource_map(
    type_edges = performance(FUN = median, units = "days")
  )

invoices %>%
  filter(InvoiceID =="b99cdf0bf63445f99e2b60a9602923b4" ) %>%
  process_map(
    type_edges = performance(FUN = median, units = "days")
  )
############################################################
############### PO and GR analysis #####################
############################################################
```

```
invoices %>% group_by(InvoiceType) %>%
  n_cases()
# 1. PO and GR
POGR<- invoices %>%
  filter(`Company code`==1001) %>%
 # filter_activity("Invoice date",T)%>%
  filter(GR=="yes" & InvoiceType=="Vend.Inv.with PO")

POGR %>%
  process_map(
    type_nodes = frequency("relative-case"),
  )

trace_explorer(POGR, coverage = 0.91)

POGR %>%
  process_map(
    type_edges = performance(FUN = median, units = "days")
  )

resource_labels(POGR)
#Remove nulls
POGRFTE<-filter_resource(POGR,NA,reverse = TRUE)
resources(POGRFTE)
POGRFTE %>%
  filter_resource_frequency(percentage = 0.80) %>%
  resource_map()

# 2. PO and no GR
POnoGR<- invoices %>%
  filter(`Company code`==1001) %>%
  filter(GR =="no" & InvoiceType=="Vend.Inv.with PO")

POnoGR %>%
  process_map(type_edges  = frequency("relative-case"))

trace_explorer(POnoGR, coverage = 0.99)

POnoGR %>%
  process_map(
    type_nodes = performance(FUN = median, units = "days")
  )
resource_labels(POnoGR)
#Remove nulls
POnoGRFTE<-filter_resource(POnoGR,NA,reverse = TRUE)
resources(POnoGRFTE)
POnoGRFTE %>%
  #filter_resource_frequency(percentage = 0.90) %>%
  resource_map()

# 3. No PO and GR
noPOGR<- invoices %>%
  filter(`Company code`==1001) %>%
  filter(GR=="yes" & InvoiceType=="Vend.Inv w/o PO")

noPOGR %>%
  process_map(
    type_nodes = frequency("relative-case")
  )

trace_explorer(noPOGR, coverage = 0.99)
noPOGR %>%
  process_map(
    type_edges = performance(FUN = median, units = "days")
  )
resource_labels(noPOGR)

#Remove nulls
noPOGRFTE<-filter_resource(noPOGR,NA,reverse = TRUE)
resources(noPOGRFTE)
noPOGRFTE %>%
  filter_resource_frequency(percentage = 0.90) %>%
  resource_map()
```

```
# 4. No PO and No GR
noPOnoGR <-invoices %>%
  filter(`Company code`==1001) %>%
  filter(GR =="no" & InvoiceType=="Vend.Inv w/o PO")

noPOnoGR %>%
  process_map( type_nodes = frequency("relative-case"))

trace_explorer(noPOnoGR, coverage = 1)
noPOnoGR %>%
  process_map(
    type_edges = performance(FUN = median, units = "days")
  )

#5. PO
PO <- invoices %>%
  filter(`Company code`==1001) %>%
  filter(InvoiceType=="Vend.Inv.with PO")

PO %>%
  process_map()

trace_explorer(PO, coverage = 0.99)
PO %>%
  process_map(
    type_edges = performance(FUN = median, units = "days")
  )
resource_labels(PO)
#Remove nulls
POFTE<-filter_resource(PO,NA,reverse = TRUE)
resources(POFTE)
POFTE %>%
  filter_resource_frequency(percentage = 0.90) %>%
  resource_map()

#6. No PO
noPO <- invoices %>%
  filter(`Company code`==1001) %>%
  filter(InvoiceType=="Vend.Inv w/o PO")

noPO %>%
  process_map()

trace_explorer(noPO, coverage = 0.99)
noPO %>%
  process_map(
    type_edges = performance(FUN = median, units = "days")
  )
resource_labels(noPO)
#Remove nulls
noPOFTE<-filter_resource(noPO,NA,reverse = TRUE)
resources(noPOFTE)
noPOFTE %>%
  filter_resource_frequency(percentage = 0.90) %>%
  resource_map(
    type_edges = performance(FUN = median, units = "days")
  )

  POGR %>%
filter_activity("Invoice date",T)%>%
throughput_time(units = "days")
POnoGR %>%
  filter_activity("Invoice date",T)%>%
  throughput_time(units = "days")
noPOGR %>%
  filter_activity("Invoice date",T)%>%
  throughput_time(units = "days")
noPOnoGR %>%
  filter_activity("Invoice date",T)%>%
  throughput_time(units = "days")


hip <-invoices %>%
  mutate(
```

```
   hypothesis = case_when(
     GR=="yes" & InvoiceType=="Vend.Inv.with PO" ~ "1.POGR",
     GR=="no" & InvoiceType=="Vend.Inv.with PO" ~"2.POnoGR",
     GR=="yes" & InvoiceType=="Vend.Inv w/o PO" ~ "3.noPOGR",
     GR=="no" & InvoiceType=="Vend.Inv w/o PO" ~ "4.noPOnoGR",
     TRUE ~ "Appropriate"
   )
 )
hip %>%
filter(`Company code`==1001) %>%
 filter_activity("Invoice date",T)%>%
  filter_activity("Removal",T)%>%
  group_by(hypothesis)%>%
  throughput_time(units = "days") %>%
  plot()

############################################################
############ Rejected / Duplicated ########################
############################################################
activities(invoices)
# invoices removed
Removed <- invoices %>%
  filter(`Company code`==1001) %>%
  filter_activity_presence(activities = 'Removal',
                           method = "all")

#filter_activity_presence(
#  eventlog,
#  activities = NULL,
#  method = c("all", "one_of", "none", "exact", "only"),
#  reverse = FALSE
#)

# Create a performance map
Removed%>%
  process_map(
    type_nodes = frequency(),
    type =performance(mean, "days"),
    type_edges = performance(mean, "days"))

Duplicated<-invoices%>%
  filter(ReasonForFailedValidation =="Duplicate")

Duplicated %>%
  process_map(
    type_nodes = frequency(),
    type =performance(mean, "days"),
    type_edges = performance(mean, "days"))

trace_explorer(Removed, coverage = 1)

# first trace
Removed %>%
  filter_trace(9) %>%
  filter_activity("Invoice date",reverse=T)%>%
  throughput_time(units = "days")

# Process map of the invoice eighth trace
invoices %>%
  filter_activity_presence(activities = 'Removal', method = "all") %>%
  filter(SupplierCode == '30090000') %>%
  process_map(
    type_nodes = frequency(),
    type_edges = performance(FUN = median, "days"))

# All invoices not removed
not_removed <- invoices %>%
  filter_activity_presence(activities = 'Removal',
                           method = "none")

#wo_removed <- not_removed %>%
#  filter_precedence(
#    antecedents ="Creation",
#    consequents = "Transfer completed",
```

```
#    precedence_type ="directly_follows",
#    filter_method ="none"
# )

#process_map(wo_removed)
# Ready to transfer and Removal ones
#ReadyToRemoval<-invoices %>%
#  filter_precedence(
#    antecedents ="Ready for transfer",
#    consequents = "Removal",
#    precedence_type ="directly_follows",
#    filter_method ="one_of"
# )
#ReadyToRemoval%>% process_map()
## table(ReadyToRemoval$InvoiceID)
# Cration to Ready to transfer and Removal
#CreationToReadyToRemoval<-ReadyToRemoval %>%
#  filter_precedence(
#    antecedents ="Creation",
#    consequents = "Ready for transfer",
#    precedence_type ="directly_follows",
#    filter_method ="one_of"
# )

##########################################################
##################### touchless ##########################
##########################################################
invoices %>%
  filter(`Company code`== 1001) %>%
  filter_activity_presence(activities = 'Removal',method = "none") %>%
  filter_resource(NA) %>%
  filter( InvoiceType=="Vend.Inv.with PO" & CycleTime < 1)%>%
  process_map(
    type_nodes = frequency(),
    type_edges = performance(FUN = median, "days")
  )

prep<-  invoices %>%
   filter(`Company code`== 1001) %>%
   filter_activity_presence(activities = 'Removal',method = "none") %>%
   filter_activity_presence(activities = 'Approval',method = "none") %>%
   filter( InvoiceType=="Vend.Inv.with PO" & CycleTime >= 1)
 # %>%
    process_map(
    type_nodes = frequency(),
    type_edges = performance(FUN = median, "days")
  )
##################################################################
##################################################################
```

References

(IEEE), I. o. (n.d.). *Task Force of Process Mining*. Retrieved from https://www.tf-pm.org

Aalst, W. v. (2016). *Process Mining, Data Science in Action.* Berlin: Springer.

Aalst, W. v. (2019). *Process Mining: Data science in action*. (Coursera) Retrieved from Coursera: https://www.coursera.org/learn/process-mining

Academic Ambassador Program. (2022). (Celonis) Retrieved from Celonis: https://academy.celonis.com/

Accounts payable. (n.d.). (Wikipedia) Retrieved from https://en.wikipedia.org/wiki/Accounts_payable

Automate Data Capture. (n.d.). (Nano Net Technologies Inc.) Retrieved from https://nanonets.com/blog/accounts-payable-process/

bupaR. (2019). *Business Process Analytics in R*. Retrieved from https://bupar.net/

bupaR Business Process Analysis with R. (n.d.). Retrieved from https://bupar.net/materials/20170904%20poster%20bupaR.pdf

Burattin, A. (2015). *Process Mining Techniques in Business Environments*. Springer.

Business Process Analysis in PowerBI using R visuals. (n.d.). Retrieved from https://community.powerbi.com/t5/Community-Blog/Business-Process-Analysis-in-PowerBI-using-R-visuals/ba-p/659401

Conference, P. M. (2019). *Business Process Intelligence Challenge*. Retrieved from BPI Challenge: https://icpmconference.org/2019/icpm-2019/contests-challenges/bpi-challenge-2019/

Deloitte Touche Tohmatsu Limited. (2011). *Shared Sevices Handbook, Hit the road.* United Kingdom: Deloitte MCS Limited.

Doxey, C. H. (2021). *The new accounts payable toolkit.* New Jersey: Wiley.

GertJanssenswillen. (2021). *Unearthing the Real Process Behind the Event Data, The case for increased process realism.* Switzerland: Springer.

Grossman, W., & Rinderle-Ma, S. (2015). *Fundamentals of Business Intelligence; Data-Centric Systems and Applications.* Springer.

Ingo Kitzmann, C. K. (n.d.). *A Simple Algorithm for Automatic Layout of BPMN Processes*. (2009 IEEE Conference on Commerce and Enterprise Computing, CEC 2009, Vienna, Austria, July 20-23, 2009) Retrieved from https://www.researchgate.net/publication/221542866_A_Simple_Algorithm_for_Automatic_Layout_of_BPMN_Processes

Interface to the PM4py. (n.d.). Retrieved from Process Mining Library: https://cran.microsoft.com/snapshot/2020-09-02/web/packages/pm4py/index.html

Janssenswillen, G. D. (2019). bupaR: Enabling reproducible business process analysis. Knowledge-Based Systems, 163, 927-930.

Kiarash Diba, S. R. (2019). *Compliance and Performance Analysis of Procurement Processes Using Process Mining.* Retrieved from https://icpmconference.org/2019/wp-content/uploads/sites/6/2019/07/BPI-Challenge-Submission-6.pdf

McCann, J. (n.d.). *Accounts payable process: Explaining the full cycle*. (Routable) Retrieved from https://blog.routable.com/accounts-payable-process/

Mohamed, A. A. (2016). *Process Mining application considering the organizational perspective using social network analysis (dissertation).* Porto: Faculdade de Economia da Universidade do Porto.

Process-analytics BPMN visualization in R. (n.d.). Retrieved from https://github.com/process-analytics/bpmn-visualization-R/

Procure-to-pay. (n.d.). (Wikipedia) Retrieved from https://en.wikipedia.org/wiki/Procure-to-pay

Silva, L. F. (2014). *Process Mining: Application to a case study (Master thesis)*. Faculdade de Economia da Universidade do Porto.

Suska, M., & Weuster, A. (2016). *Shared Services: Multiplyng Success*. Germany: PricewaterhouseCooper Aktiengesellschaft.

Tableau. (n.d.). Retrieved from https://help.tableau.com/current/pro/desktop/en-us/pareto.htm

Vercellis, C. (2009). *Business Intelligence, Data Mining, and Optimization for Decision Making*. West Sussex: John Wiley & Sons Ltd.

Vincent F. A. Meyer zu Wickern, M. J.-T.-H.-H. (2019). *Analysis and prediction of purchasing compliance using process mining*. Retrieved from https://icpmconference.org/2019/wp-content/uploads/sites/6/2019/07/BPI-Challenge-Student-Submission-4.pdf

Westland, J. C. (2020). *Audit Analytics Data Science for the Accounting Profession*. Switzerland: Springer.

FACULDADE DE **ECONOMIA**