BRNO UNIVERSITY OF TECHNOLOGY
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FACULTY OF INFORMATION TECHNOLOGY
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

DEPARTMENT OF INFORMATION SYSTEMS
ÚSTAV INFORMAČNÍCH SYSTÉMŮ

AUTOREFERAT DISERTAČNÍ PRÁCE NA TEMA

MODELLING AND ANALYSIS OF LOGISTICS PROCESSES BY APPLYING PROCESS AND DATA MINING TECHNIQUES

MODELOVÁNÍ A ANALÝZA LOGISTICKÝCH PROCESŮ POMOCÍ PROCESNÍCH A DATOVÝCH ANALYTICKÝCH METOD

AUTHOR **Mgr. JULIA RUDNITCKAIA**
AUTOR PRÁCE

SUPERVISOR **prof. Ing. TOMÁŠ HRUŠKA, CSc.**
VEDOUCÍ PRÁCE

BRNO 2022

Abstract

In this thesis, we propose an approach for modelling hidden and unknown processes and subprocesses in the example of a seaport logistics area. Having the underlying process model makes it possible to exploit more advanced algorithms since deviations and main paths are becoming visible and better controlled. The obtained model is the foundation for the core research of this work and will be enriched with key performing indicators and their forecast by applying advanced process mining, statistics, and machine learning techniques. The main difference of the approach is that we take as a target variable not any specific value, but the object – a process variant or a process type with a set of parameters. Bottleneck analysis, from one side, and predictive analysis, on the other hand, are enforced with context-aware information, especially with these additional objective process attributes. Furthermore, the support of the descriptive (*As is*) current process model with certain notation and the integration with relevant bottleneck and predictive methods compromise the advantages of the approach. The work primarily focuses on the design of algorithms and methods for supporting logistics data analysis. However, it can be adjusted and applied to other areas accordingly, which makes the approach flexible and versatile. The result of the work is the framework for unstructured process modelling and the key process parameters predictive method. This analysis of processes with their attributes might be used for decision-making systems and process maps in future.

Keywords

process mining, statistics, predictive models, data mining, nautical process, logistics process, process modelling, bottleneck analysis, process maps, context awareness

Abstrakt

V této práci navrhuji přístup k modelování skrytých a neznámých procesů a podprocesů na příkladu logistiky námořního přístavu. Základní procesní model umožňuje využívat pokročilejší algoritmy, protože odchylky a hlavní cesty jsou viditelnější a lépe kontrolovatelné. Získaný model je základem pro stěžejní výzkum této práce a je obohacen o klíčové ukazatele výkonnosti a jejich predikci použitím pokročilých technik procesního dolování, statistiky a strojového učení. Hlavní rozdíl v přístupu je v tom, že jako cílovou proměnnou neberu žádnou konkrétní hodnotu, ale objekt – variantu procesu nebo typ procesu se sadou parametrů. Analýza úzkých míst na jedné straně a predikční analýza na druhé straně jsou založeny na informacích, které jsou zlepšené pomoci *context-aware* informace, zejména těmito dalšími objektivními atributy procesu. Kromě toho podpora deskriptivního (*Jak je*) aktuálního procesního modelu s určitou notací a integrací s relevantními metodami detekce úzkých míst a prediktivními metodami doplňuje výhody tohoto přístupu. Práce se primárně zaměřuje na návrh algoritmů a metod pro podporu analýzy logistických dat. Lze je však odpovídajícím způsobem upravit a aplikovat na jiné oblasti, díky čemuž je přístup flexibilní a universální. Výsledkem práce je postup pro modelování nestrukturovaných procesů a metoda predikce klíčových parametrů procesů. Tato analýza procesů s jejich atributy může být v budoucnu využita pro systémy rozhodování a procesní mapy.

Klíčová slova

procesní dolování, statistika, predikční modely, dolování z dat, námořní proces, logistický proces, procesní modelování, analýza úzkých míst, procesní mapy, povědomí o kontextu

Contents

Chapter 1

Introduction

With the spectacular growth of data information systems play a main role in today's business processes as the digital universe and the physical universe are becoming more and more aligned. The expansion of the digital universe that is well-aligned with processes in organizations makes it possible to record and analyze events. Nevertheless, organizations have problems extracting significant value from these data, because most of the data stored in the digital universe are unstructured. In the constant competition, entrepreneurs are forced to act quickly to keep afloat. Smart and robust field planning and development are essential to ensure the profitability and success of future fields under low-price scenarios and challenging environments. Using modern mathematics methods and algorithms helps to quickly answer questions that can lead to increased efficiency, resource utilization, improving productivity and quality of provided services, minimizing environmental impact and costs. People and digital technologies collaborate, using, for instance, advisory systems (recommendation systems, decision-making systems), where decisions are made by humans based on recommendations from a decision-support system. However, there are quite new and powerful methods that are keeping data scientists' attention recently. It not only allows organizations to fully benefit from the information stored in their systems, but it can also be used to check the conformance of processes, detect bottlenecks, and predict execution problems.

An ongoing trend is in the future balancing between domain experts and data scientists. Insights and lessons learned from other industries can and should be exploited and new digital and automation technologies tailored for the logistics industry should be developed. Last decades, the question of creating hybrid professions has been raised more and more often due to the lack of specialists in related industries (the main topic of the conference HMS 2018 which was held in Budapest). Data scientists have to know not only modern analytical methods and information technologies but also understand the nature of data and processes they are going to analyze to provide correct and valuable results. Such specializations as analytics, business owners, data scientists and process miners have not separated anymore. It creates more requirements for analysts as well as flexibility and power to apply recent algorithms in a creative better way. Having deep knowledge of mining algorithms and any applying areas- logistics, industry, education, economics, etc. – brings new aspects to the process analysis. This work focuses on the application domain - logistics, especially on sea-port logistics, from the one side and process mining methods – from another side. The industry is struggling to find out how to efficiently utilize data. To elaborate, there is considerable potential in the use of process mining, statistics and machine learning techniques to transform historical data and real-time production data into actionable advice in daily operations, also denoted as short-term production optimization.

The reason to take port logistics processes is prevailing recently preconditions (for more details see section 2.2.5 Prerequisites for applying advanced mining methods), which make it possible to apply and examine prominent algorithms of process mining in a new field. Combining knowledge of logistics processes and modern methods of analytics - process mining, data mining, machine learning and statistics – can undoubtedly help improve both fields and adopt algorithms to a wider range of process types. The results of its use directly depend on the settings and understanding of the analyzed process. Therefore, we describe problematics, motivation and solutions for two sides of the research.

1.1 Goal of the Thesis. Research subgoals

The main aim of the research described in theses thesis is to extract knowledge and models from real-world raw event logs from interconnected processes and to identify problems to further improve the complex non-structured processes. All these data will be collected, transformed and analyzed to build the groundwork for predictive and bottleneck analysis of port processes. By closely working with domain experts, predictive models based on machine learning and statistics will be developed for the prediction of resource needs, activities, operations, and potential delays, and deviations. In turn, a new-found bottleneck detection method helps to highlight the risk/weak places of the model, which impacts the prediction of key parameters. The difference from other research is in applying so-called context aware information (process variants) as an input value for the proposed algorithms.

A prototype decision support system can then be developed for different participants involved in the interconnected logistics processes. Thus, we will try to combine and empower process mining results with advanced methods of data mining, machine learning and statistics fields.

The additional subgoal was to find a solution to clean the data with abnormal distribution in the correct way and to apply appropriate metrics based on the nature of the data distribution. The possible ways to solve this and other data quality issues will be described.

In terms of logistics area, the next subgoals were followed and achieved:

- To improve the understanding of the properties of the chain of nautical handling processes directed at maritime ships around seaport, from arrival to departure.
- To develop valid and practicable methods for modelling logistics processes, from a complex system perspective.
- To design strategies for further use of the obtained process model.
- Besides these goals, the thesis summarizes recent research, technologies, methods, and challenges of the current state of process mining, forecast models and bottlenecks detection methods.

1.2 Relevance. Motivation. Methodology

Relevance. According to the last articles, conferences, and webinars on the process mining topic[1]– the more we apply its algorithms to new process domains, the newer interesting challenges, which deepen our knowledge and understanding of process nature, occur. For instance, dealing with the unstructured process model and empowering the model with additional parameters still remain to be one of the important process mining challenges. The relevance of the research for process mining will be described in detail in section 2.1.7. Existing issues and challenges.

Furthermore, current European logistics projects such as SmartPort[2], SwarmPort[3], Indeep[4], IoT for Agri are looking for new methods to structure and control nautical processes to choose the most appropriate one. They describe the relevance of the research with next points:

- So far, no one has made the link between the new port performance analysis models from economics, which provide theoretical rigor, and collaborative port system designs, which describe stakeholder behavior in daily business terms.
- Until now, there has been no attempt to represent interconnected port processes in a behaviorally validated descriptive setting (e.g. single terminals or modes of transport).

[1] www.fluxicon.com/camp/2022

[2] www.smartport.nl/en

[3] www.tudelft.nl/en/2017/infrastructures/swarmport/

[4] www.indeep-project.org

- There does not exist yet a single complete tool that would allow us to study such complex systems and their dynamics. Maritime port systems represent a major challenge both from a theoretical and practical point of view and require the development of innovative tools for the analysis and synthesis of such systems.

Motivation. The motivation was to combine knowledge from interconnected areas, and studies of the latest scientific sources and to confront them with the available tools, methods and technologies. The term *Process mining* was first coined in a research proposal written by the Dutch computer scientist Wil van der Aalst ("Godfather of Process mining")[5]. Thus began a new field of research that emerged under the umbrella of techniques related to data science and process science at the Eindhoven University in 1999. Now, process mining is a promising approach having already many communities, practitioners and research groups around the world. It can fill the gap between traditional model-based process analysis (e.g., simulation and other business process management techniques) and data-centric analysis techniques such as machine learning and data mining. In their turn, researchers try to adopt the method for as many as possible processes and improve the implementation of algorithms and readability of the process model.

Methodology. We can identify several promising approaches to better analyze and synthesize complex processes with their key aspects and attributes. These approaches have not been applied rigorously to logistics processes so far. The methodology consists of five distinctive steps: logistics event log extraction, data preprocessing, process modelling for a complex non-structured process, bottleneck detection analysis and a two-step predictive model, including results of all previous steps. The scheme for the research is presented below. All steps will be discussed in the second chapter for methodology.

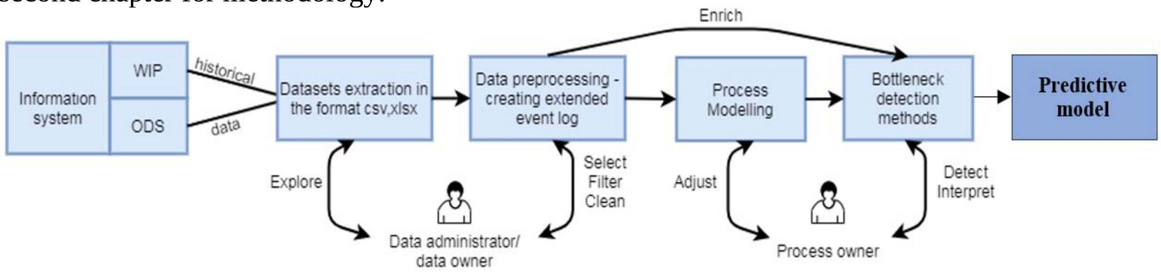

Enrich
Information system
WIP
ODS
historical
data
Datasets extraction in the format csv,xlsx
Data preprocessing - creating extended event log
Process Modelling
Bottleneck detection methods
Predictive model
Explore
Select
Filter
Clean
Data administrator/ data owner
Adjust
Detect
Interpret
Process owner

Figure 1.1: The roadmap of the current research

The results of the work are presented as a consistent set of methods and techniques, which are strongly interconnected and depend on each other results:

- a framework to deal with unstructured and unknown process modelling using additional parameters of the processes. This framework enables terminals to rapidly construct a robust, reliable and transparent model to which others, in turn, can connect their own services. In terms of process mining, the framework shows the way of revealing hidden (also unused and unknown) subprocesses and handling *spaghetti* (non-structured, non-readable, poorly regulated and entangled processes) process models as well as presenting new perspectives/types of the process. Once it is developed, the process model can be used as a foundation and be enriched with additional parameters for recommendation or forecasting systems.
- Proposal of the combined bottleneck detection method for the defined process model using context-aware information to detect weak points.
- Develop predictive models based on machine learning, statistics and process mining technologies using process types and results of bottleneck algorithm as predictor variables. The terminal operator can get the opportunity not just to better plan schedules, but also know risks, times, costs, next possible activity in the process.

[5] https://en.wikipedia.org/wiki/Process_mining

1.3 Structure of the Thesis

The thesis deals with process modelling and enhancement of this model with additional key performance parameters, which can be used in future as a significant part of a process map. To pursue this goal, the research structure is proposed as follows and concluded in five chapters:

1. In the Chapter 1, the reader is informally and briefly introduced to the goals, tasks and methodology of the current work. Main directions and questions are assigned here and will be considered in detail in the following chapters.
2. Chapter 2 presents the theoretical part used in the thesis. This part of the research will be dedicated to reviewing the state of art and understanding process mining terminology and insights. The most important information about process mining, its methods and techniques, and current research in the area are described here. Also, we will discuss widely used and applied methods of bottleneck detection analysis as well as predictive models. Since our research has a hybrid type, one section of the chapter will be dedicated to logistics processes in the seaport domain. We will form preconditions for applying the proposed methods and discuss the current situation. The main focus of the chapter is to define the benefits of existing approaches and then, to see their disadvantages, problematics and current challenges. Issues and challenges are summarized at the end of each section.
3. After the review of the state of art, the next stage is to modify and adjust existing methods to apply them to our logistics application domain. The invented framework for process modelling describes each step thoroughly. The proposed methods and algorithms for cleaning, modelling, and analyzing data will be described in the Chapter 3. Also, we take a look at their difference from existing methods and improvements made. This chapter is the core of the research and proposes the solution to issues, mentioned in the previous chapter.
4. The next step will apply the new defined methods and metrics to the logistics process. We will provide a case study and evaluate the applicability of the proposed approaches. A set of tools will be also presented in this part. This part is shown in the Chapter 4.
5. Last stage of the research will remind the initial motivation for the work – a creation of a process map. We will summarize all accomplished work and contributions, the main ideas and results. Also, possible directions for future research will be assigned in the final Chapter 5.

Chapter 2

State of the Art. Data science and process modelling

If you cannot describe what you are doing as a process,
you do not know what you are doing
W. Edwards Deming

People are increasingly surrounded by internet-connected devices that generate and exchange data in high volumes and at an unprecedented pace. Big data is an evolving term that describes any voluminous amount of structured, semi-structured and unstructured data that has the potential to be mined for information [1]. Although big data does not refer to any specific quantity, the term is often used when speaking about petabytes and exabytes of data.

Because big data takes too much time and costs too much money to load into a traditional relational database for analysis, new approaches to storing and analyzing data have emerged that rely less on data schema and data quality. Instead, raw data with extended metadata is aggregated in a data lake and machine learning and artificial intelligence (AI) programs use complex algorithms to look for repeatable patterns.

Data science is the profession of the future because organizations that are unable to use (big) data in a smart way will not survive. Clive Humby even offered to use the metaphor “Data is new oil” to emphasize what important role data plays nowadays [2]. It is not sufficient to focus on data storage and data analysis. The data scientist also needs to relate data to process analysis to understand the nature and structure of the process. One of the effective approaches to cover these needs is process mining. Generally speaking, process mining bridges the gap between traditional model-based process analysis (e.g., simulation and other business process management techniques) and data-centric analysis techniques such as machine learning and data mining (see Figure 2.1).

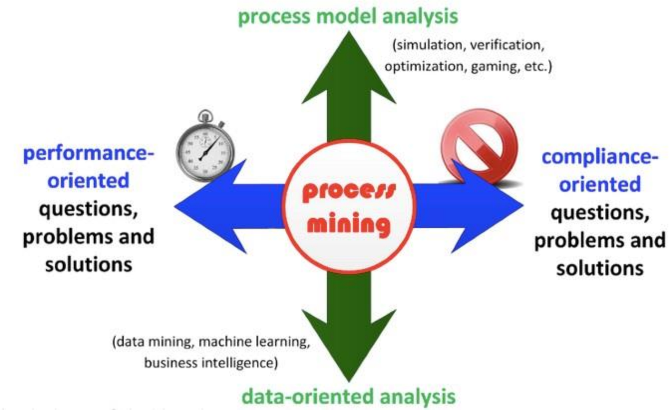

process model analysis
(simulation, verification, optimization, gaming, etc.)
performance-oriented questions, problems and solutions
process mining
compliance-oriented questions, problems and solutions
(data mining, machine learning, business intelligence)
data-oriented analysis

Figure 2.1: Positioning of process mining in the analytics [3]

More and more information about business processes is recorded by information systems and can be presented or transformed into the form of so-called *event logs*, which is a start point for process mining. Although event data are omnipresent, organizations lack a good understanding of their actual processes. Management decisions tend to be based on PowerPoint[6] diagrams, local

[6] www.microsoft.com/en-us/microsoft-365/powerpoint

politics, or management dashboards rather than careful analysis of event data. The knowledge hidden in event logs should be turned into actionable information. Advances in data mining made it possible to find valuable patterns in large datasets and to support complex decisions based on such data. However, classical data mining problems such as classification, clustering, regression, association rule learning, and sequence/episode mining are not process-centric. Talking of differences between data mining and process mining we can make a comparison like between relational databases and object-oriented – in the first case, we are working with data, in the second, objects (end-to end processes) can be handled. Approaches are different and cannot replace each other, but what is more effective, merging them can enrich the results of analysis and bring it to the next level.

The main challenge today is not to generate more data but to turn data into real value. Therefore, data science is a new engineering discipline and the main driver for innovation in the years to come, because it can help with getting knowledge from big data. Data science aims to answer questions in the following four categories:

- Reporting: What happened?
- Diagnosis: Why did it happen?
- Prediction: What will happen?
- Recommendation: What is the best that can happen?

Figure 2.2 shows what areas data science covers and one can see that it is a very broad and powerful domain [3].

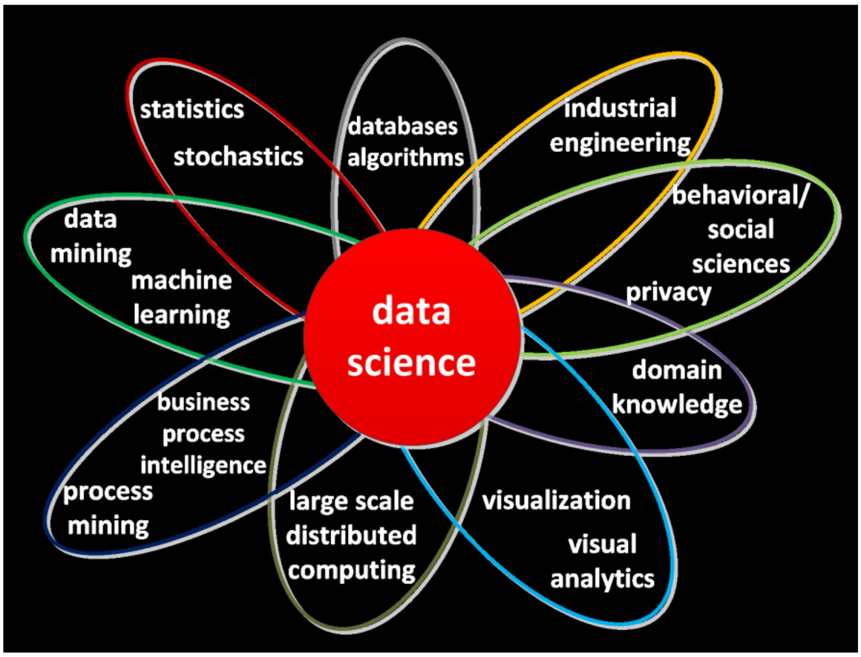


Figure 2.2: Possible analytics areas of data science

For answering above-mentioned questions there are different disciplines. Some of them can help to analyze event data, another – to handle large volumes of events. It can be said that combining advanced methods of data mining, machine learning and statistics fields will empower process mining results.

Figures 2.2 and 2.3 show few widely used areas that help to simplify work with data and their interconnection. We can see positioning of techniques, which will be used and described in following sections– data and process mining, statistics, machine learning and domain knowledge (logistics processes in our case).

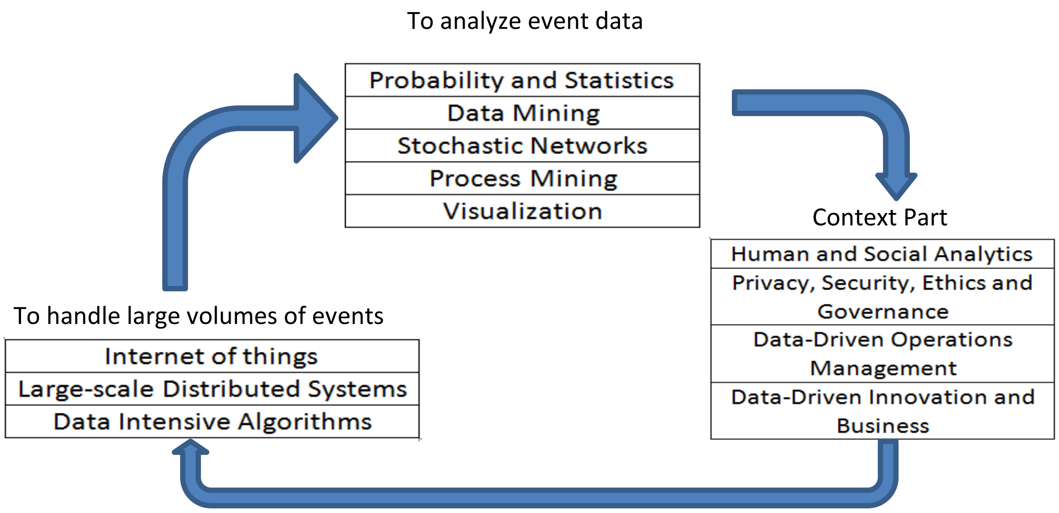

To analyze event data
Probability and Statistics
Data Mining
Stochastic Networks
Process Mining
Visualization
Context Part
Human and Social Analytics
Privacy, Security, Ethics and Governance
Data-Driven Operations Management
Data-Driven Innovation and Business
To handle large volumes of events
Internet of things
Large-scale Distributed Systems
Data Intensive Algorithms

Figure 2.3: Placement Process Mining in the common scheme of Data Science

According to the source [4], the Data Science pattern includes four main parts needed for a proper data analysis:

- Modeling. It is a graphical representation of business processes or workflows, where each individual step is presented and connected to others. Shows end-to-end overview in the context of the business field. This part, in turn, includes next methods: linear regression, k-means clustering, non-linear regression, logistic regression, hierarchical clustering, Naïve Bayes classifier, SVM (Support Vector Machine), random forest, decision tree, neural network methods, petri nets, etc.
- Optimization – is an adjustment of a process so as to optimize some specified set of parameters to get better organizational or performance results. Methods: regularization, bagging (Bootstrap aggregating), boosting.
- Feature preparation. This part is the most crucial. All algorithms and their results are useless once the appropriate well-prepared data set is unobtainable. Methods to preprocess data set and process features: discretization, feature scaling, feature imputation (dealing with missing values), feature selection, and feature extraction.
- Validation. After results are obtained, they should be evaluated by mathematician metrics or techniques like ROC curve (Receiver Operating Characteristics), confusion matrix, cross-validation, holdout method, RMSE (Root-mean-square deviation), MAE (Mean Absolute Error), MSE (Mean squared error), etc.

Most of these methods, techniques and tools are created with one purpose – minimalize big data by getting only necessary *noiseless* useful data. In the next chapters, we will be dealing with these data science parts and describe their problematics. Moreover, we look at how the results of one technique can be empowered by another.

2.1 Process modelling and context-aware information

There is a wide range of ways to define the concept of a business process. A business process is the combination of a set of activities within an enterprise with a structure describing their logical order and dependence whose objective is to produce the desired result [5]. And when this definition is considered, one realizes that business processes are everywhere in our daily lives. Processes are crucial parts of our industries and organizations [6].

Process modelling is the graphical representation of business processes or workflows. Like a flow chart, individual steps of the process are drawn out so there is an end-to-end overview of the tasks in the process within the context of the business environment [7]. Business process modelling enables a common understanding and analysis of a business process. A process model can provide a comprehensive understanding of a process. In this work, we divide process models into normative (*To be*, *de jure models*) and descriptive (*As is*, *de facto models*) types. The normative (assumed) process models are created usually by domain experts and business owners. They describe the ideal process in the company with roles and can be used for training new staff and understanding organization workflows and their cooperation. The descriptive type of model shows a current process and is made by algorithms based on raw data. The obtained model is used for optimization, improvement, and enhancement of the real process and can be the basis for recommendation or decision-making systems as well as for process maps. The focus of this work is the descriptive model and in order to obtain it, process mining techniques are applied.

2.1.1 Process Mining. Basic concepts

Process mining (*PM*) – a set of techniques, tools, and methods to discover, monitor, and improve real processes (i.e., not assumed processes) by extracting knowledge from event logs commonly available in today's (information) system [3].

PM seeks the confrontation between event data (i.e., observed behaviour) and process models (hand-made or discovered automatically). This technology has become available only recently, but it can be applied to any type of operational process (organizations and systems). The range of hodiernal applications is wide. Here are some of them listed:

- analyzing treatment processes in hospitals [8],
- improving customer service processes in a multinational environment [9],
- understanding the browsing behaviour of customers using a booking site [10],
- analyzing failures of a baggage handling system,
- improving the user interface of an X-ray machine [11].

All of these applications have in common that a dynamic behaviour needs to be related to process models.

There are two main drivers for the growing interest in process mining:

1. More and more events are being recorded, thus, providing detailed information about the history of processes.
2. There is a need to improve and support business processes in competitive and rapidly changing environments.

PM provides an important bridge between BI and BPM, data mining and workflow. It includes (automated) process discovery (i.e., extracting process models from an event log), conformance checking (i.e., monitoring deviations by comparing model and log), social network/organizational mining, automated construction of simulation models, model extension, model repair, case prediction, and history-based recommendations.

In the table below, there are pointed various directions of analysis and significant questions that PM can answer to.

Table 2.1: Process mining use cases and their focus groups

№	Use Case	Questions	Group of questions
1	Detection of real-world business processes	What is the process that actually (and not in words and not in theory) describes current activities?	coherence
2	Search bottlenecks in business processes	Where are places in the process, limiting the overall speed of its implementation? What causes these places?	productivity
3	Detection of deviations in business processes	Where the actual process deviates from the expected (ideal) process? Why are there such deviations?	coherence
4	Search fast / short cuts execution of business processes	How to perform the process of the fastest? How to perform a process for the least number of steps?	productivity
5	Prediction of problems in business processes	Is it possible to predict the occurrence of delays or deviations or risks in the performance of the process?	productivity/ coherence

So, in common, PM use-cases are focused for the next main questions:

- What is the process that people really follow?
- Where are the bottlenecks in the process?
- Where do people (or machines) deviate from the expected or idealized process?
- How to redesign a process to improve its performance?
- What is likely to happen in the future?
- What are the *highways* in my process?
- Can we predict problems (delay, deviation, risk, etc.) for running cases?
- Can we recommend countermeasures?

PM has been deployed successfully in many studies and real applications in various domains such as healthcare [11], software development [12], and education [13]. However, there are few works related to the logistics domain [14, 15] and most of the time algorithms are applied to artificially generated event logs.

Event Log

The first step to start with PM techniques is to create an event log. Event Log (*EL*) is a collection of cases, where each element refers to a case, an activity and a point in time (timestamps).

To define *event log*, we can use the next formula:

$$EL =< C, A, T_s, T_e, R, \dots > \qquad (2.1)$$

where

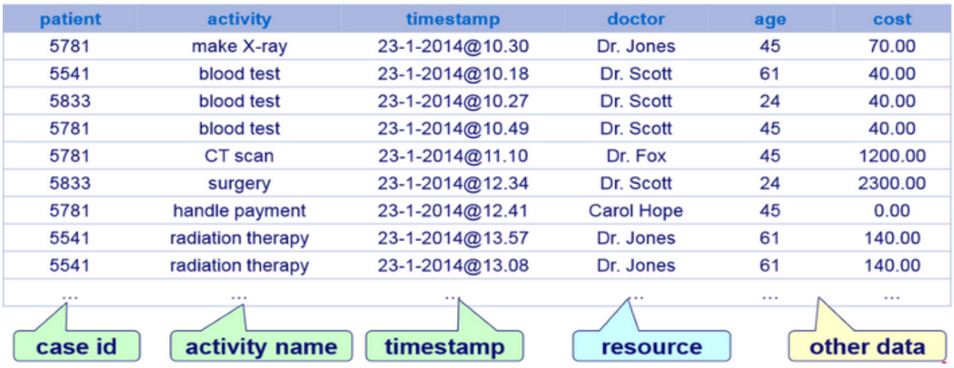
- *C* (case ID) – instances (objects), which is arranged sequence of activities. The sample of cases can be a user id on the website, a patient case number in the hospital, a document id, or even a resource name. C defines the scope of the process.
- *A* (activity name) – actions performed within the EL, or operations in the process. Each step which was recorded plays the role of activity. Activity determines the level of detail

for the process steps.

- *Ts*,*Te* (timestamps) – date and time of recording log events, the start and end of each activity accordingly. Timestamps determine the order of the activities in the process.
- *R* (resource) - holds the key actors in the log of events (those who perform actions in the EL).

Each EL should combine these main parameters to make it possible to apply process mining algorithms. Additional process attributes can help with more advanced analysis for prediction and bottleneck analysis and will be considered in the section Context awareness.

Figure 2.4 shows the sample of basic attributes of the events in the logs. The case is patient id and the process is his/her treatment in a hospital.



patient	activity	timestamp	doctor	age	cost
5781	make X-ray	23-1-2014@10.30	Dr. Jones	45	70.00
5541	blood test	23-1-2014@10.18	Dr. Scott	61	40.00
5833	blood test	23-1-2014@10.27	Dr. Scott	24	40.00
5781	blood test	23-1-2014@10.49	Dr. Scott	45	40.00
5781	CT scan	23-1-2014@11.10	Dr. Fox	45	1200.00
5833	surgery	23-1-2014@12.34	Dr. Scott	24	2300.00
5781	handle payment	23-1-2014@12.41	Carol Hope	45	0.00
5541	radiation therapy	23-1-2014@13.57	Dr. Jones	61	140.00
5541	radiation therapy	23-1-2014@13.08	Dr. Jones	61	140.00
...	...	...	...	...	...

Figure 2.4: Example of event log (the selection of the attributes depends on the purpose of analysis)[7]

Using Figure 2.4, some assumptions about event logs can be made:

- A process always consists of cases.
- A case consists of events such that each event relates to precisely one case. This assumption can be reconsidered according to last challenges in process mining which will be described later.
- Events within a case are ordered.
- Events can have attributes as well as processes themselves.
- Examples of typical attribute names are activity, time, costs, and resources.

Sources of event data can be obtained in different systems, logs, emails and so on. For instance, to create an EL with the needed attributes next sources can be exploited:

- database system,
- transaction log (e.g. a trading system),
- business suite/ ERP system/CRM (SAP, Oracle, Agile, Siebel, etc.),
- message log (e.g. from IBM middleware),
- open API providing data from websites or social media,
- CSV (comma-separated values) or spreadsheet etc.

When extracting an event log from different sources, there are certain challenges that can occur and should be processed before applying modelling algorithms:

[7] www.processmining.org/event-data.html

1. *Correlation*. Events in an event log are grouped per case. Often, this requirement is challenging as it requires event correlation, i.e. events need to be related to each other. The case should be defined in advance and be the core for further analysis.
2. *Timestamps*. Events need to be ordered per case. There are common issues with timestamps: only date information – no detailed data is available and recorded, different clocks, delayed logging and scanning.
3. *Snapshots*. Cases may have a lifetime extending beyond the recorded period, e.g. a case was started before the beginning of the event log. One should define the start and end point of the processes for the case.
4. *Scoping*. There is an issue to decide which tables to incorporate and how much data to analyze. It strongly depends on the specific cases and can be solved with domain experts.
5. *Granularity*. The events in the event log are at a different level of granularity than the activities relevant to end users.

Additionally, event logs without preprocessing have so-called *noise* and *incompleteness issues*. The first one means the event log contains rare and infrequent behaviour not representative for the typical behaviour of the process. In turn, the incompleteness is when the event log contains too few events to be able to discover some of the underlying control-flow structures. There are many methods to *clean* data and use only useful data as filtering and data mining techniques. Part 3 of this thesis will demonstrate possible solutions to deal with these issues.

Finally, it is worth mentioning some extensions of event logs:

- Transactional information on activity instance: an event can represent a start, complete, suspend, resume, and abort. This additional information is valuable to create transitional systems and predictive analysis.
- Case versus event attributes: case attributes don't change, e.g. the birth date or gender, whereas event attributes are related to a particular step in the process and change during the time.

All process mining techniques assume that it is possible to sequentially record events such that each event refers to an activity (i.e., a well-defined step in some process) and is related to a particular case (i.e., a process instance). Event logs may store additional information about events. In fact, whenever possible, the process mining techniques use extra information such as the resource (i.e., person or device) executing or initiating the activity, the timestamp of the event, or data elements recorded with the event (e.g., the size of an order). Extracting of event log for logistics will be described in the Chapter 3.

2.1.2 Types of Process Mining

To understand the capabilities of PM, it is important to understand types of this technique. There are three main types of process mining (see Figures 2.5 and 2.6) [3].

1. The first type of process mining is *discovery*. A discovery technique takes an event log and produces a process model without using any a-priori information. An example is the Alpha-algorithm that takes an event log and produces a process model (e.g. a Petri net) explaining the behaviour recorded in the log. The discovered process model is used as a basis for two other types.
2. The second type of process mining is *conformance*. Here, an existing process model is compared with an event log of the same process. The conformance checking can be used to check if reality, as recorded in the log, conforms to the model and vice versa. Process conformance focuses on the alignment between the actual process and its model. Several studies reveal that models often deviate from reality. For example, in order to react in urgent situations, processes require flexibility to take corrective actions [16].

Conformance checking illustrates where the actions and activities are altered. This kind of information can be used to improve business processes to be more corresponding with the real world. Meanwhile, in some cases, is it crucial to assure that the actual procedure is aligned with the model such as detecting abnormal processes or fraud. In this context, conformance checking methods could reveal alignment between actual procedures and models.

3. The third type of process mining is *enhancement*. The main idea is to extend or improve an existing process model using information about the actual process recorded in some event log. Whereas conformance checking measures the alignment between model and reality, this third type of process mining aims at changing or extending the a-priori model. An example is the extension of a process model with performance information, e.g., showing bottlenecks. The discovered models can be used in further analysis such as finding bottlenecks and optimizing resource allocation.

Furthermore, combining process mining types with other methods helps achieve the objective of the work's goal. The concept of a data mining, statistical, and machine learning can be integrated to discover specific questions for different process mining types. For instance, standard classification methods can be applied to find the decision point in a process model [17]. Moreover, integration with other methods can be used for prediction and recommendation at runtime.

PM types in terms of inputs and outputs are presented below:

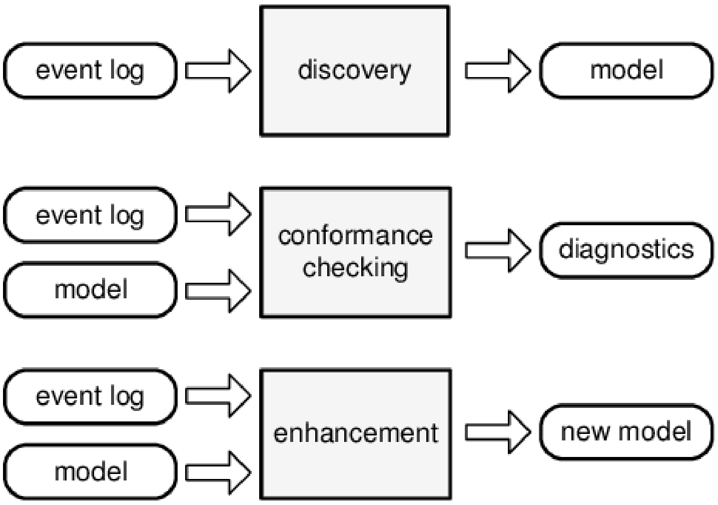


Figure 2.5: The three basic types of process mining explained in terms of input and output [26]

There are three main relations between process model and event log [3]:

1. *Play-Out* (model→behavior) – to represent the behavior of the system/process, next methods can be used: Petri nets, Workflow and Simulation engine, management games, model checking.
2. *Play-In* (logs→model) – to create the process model, different event logs are explored by α-algorithm, data mining techniques (decision tree, association rules).
3. *Replay* – to check performance and process deviations, as well as to evaluate the model, event logs can be replayed on the obtained process model.

Summarizing the above, the holistic view of process mining types and relations between process model and event log is shown in Figure 2.6.

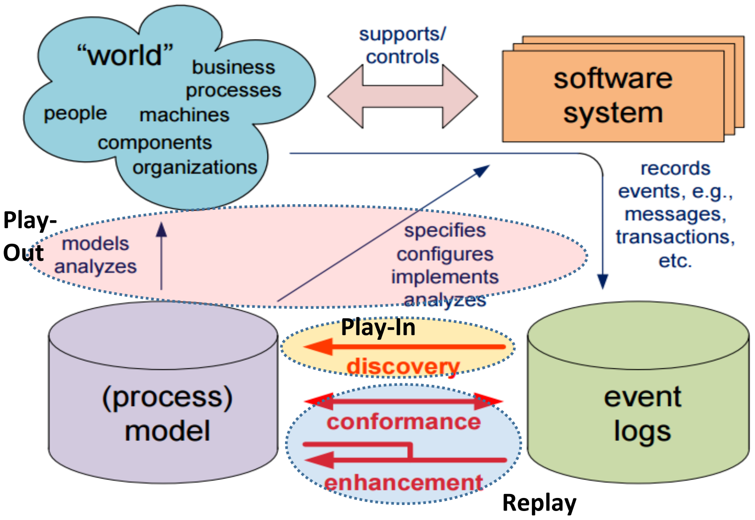


Figure 2.6: Positioning of the three main types of PM with highlighting areas of relations between a process model and event log [3]

Orthogonal to the three types of mining, different perspectives can be defined.

- *The control-flow perspective*. Focuses on the ordering of activities, the goal of mining – to find a good characterization of all possible paths (can be expressed in a Petri net or some other notation as EPCs, BPMN, UML, etc.).
- *The organizational perspective*. Focuses on the information about resources hidden in the log, goal is to either structure the organization by classifying people in terms of roles and organizational units or to show social network; to structure the organization by classifying people.
- *The case perspective*. Focuses on properties of cases.
- *The time perspective*. It's concerned with the timing and frequency of events. It makes it possible to discover bottlenecks, measure service levels, monitor the utilization of resources and predict the remaining processing time of running cases.

The common use of process mining usually ends with a control-flow perspective. However, it should be just a start point for the advanced analysis. Once modelled behaviour and real behaviour are aligned it gives the opportunity to include other perspectives of process mining, what can be beneficial in terms of online prediction, automated process improvement, resource recommendation, etc. The current work focuses mostly on control-flow, case, and time perspectives.

2.1.3 Process discovery – process modelling

Process discovery (PD) is the fundamental task for process mining. PD techniques generate process models reflecting the actual processes. These models provide specific information such as the process bottleneck, the delays, or process deviation [18]. This type of PM can help to find out a descriptive process model. Based just on an event log, a process model is constructed thus capturing the behaviour seen in the event log.

Depending on the goal of process modelling, the most common algorithms of process discovery type are [19, 20]:

- Alpha Miner;
- Alpha+, Alpha++, Alpha#;
- Fuzzy miner;
- Heuristic miner;
- Multi-phase miner;
- Genetic process mining;
- Region-based process mining (State-based regions and Language based regions);
- Classical approaches not dealing with concurrency (Inductive inference [21] and Sequence mining).

In addition, any discover technique requires representational bios. It helps limit the search space of possible candidate models. Also, it can be used to give preference to particular types of models. It is important to observe process discovery is, by definition, restricted by the expressive power of the target language. Therefore, representational bias is the selected target language for presenting and constructing process mining results. Because every notification is not universal and has its limitations (e.g. silent steps, work with duplicate activities, with concurrency and loops, etc.) and benefits, it recommends trying different variants to correct the interpretation of a process. There are commonly used notifications for representing of an obtained process model to end-user: Workflow Nets, Petri Nets, Transition Systems, YAWL, BPMN, UML, Directly-Follows Graph (DFG), Causal nets (C-nets) and Event-Driven Process Chain (EPCs).

The following are the main characteristics of PD algorithms:

1. Representational bias:

   - inability to represent concurrency;
   - inability to deal with (arbitrary) loops;
   - inability to represent silent actions;
   - inability to represent duplicate actions;
   - inability to model OR-splits/joins;
   - inability to represent non-free-choice behaviour;
   - inability to represent hierarchy.

2. Ability to deal with noise.
3. Completeness notion assumed.
4. Used approaches - direct algorithmic approaches (α-algorithm), two-phase approaches (TS, Markov model plus WF-net), computational intelligence approaches (genetic algorithm, neural networks, fuzzy sets, swarm intelligence, reinforcement learning, machine learning, rough sets), partial approaches (mining of sequent patterns, discovery of frequent episodes), etc.

Based on these characteristics, one can choose an appropriate algorithm responding to the requirements for process modelling. For the goals of the current work, the fuzzy mining algorithm was chosen as the main one to explore the process model.

Fuzzy mining

Fuzzy mining (FM) [22] is related to a heuristic mining algorithm and aims to extract less structured and complex processes. The existing algorithms of process mining like α-mining and heuristic mining encounter problems when they have to deal with processes from a restrictive environment as it is often found in reality.

In this research, the FM is deployed for our use case as it is robust to less structured and complex processes. As a result, the FM is capable of providing understandable process models from dynamic logistics processes. In contrast, α-mining and heuristic mining often encounter problems when they have to deal with processes from flexible environments. The resulted models from these algorithms when mining these types of processes are unstructured and hard to comprehend, resulting in so-called *spaghetti models*. FM provides a high-level view on the process and abstracts from undesired details.

FM involves in identifying two fundamental metrics, significance and correlation [23].

The significance metric takes the frequency of events into an account. The events which are more frequent are more important. The correlation metric is used to determine how closely related two events following one another are. The highly significant activities are contained in the simplified model. The less significant activities but highly correlated are integrated and present as a cluster in the simplified model. The less significant and weakly correlated activities are removed from the simplified model [67].

Three metrics are used for measuring significance and correlation:

- unary significance,
- binary significance,
- binary correlation.

Unary significance is composed of two primary metrics. The first one is frequency significance which measures how often a certain event class was observed in event log. Another metric is routing significance. It helps to identify important routing nodes. For instance, the higher the number of the difference between the incoming arcs and outgoing arcs, the more important the node is in terms of routing.

Binary significance relates to a precedence relation between two event classes, which can be used for selecting the edges that will be included in simplified process models. The frequency metric can be used to identify the relationship between two event classes. The more frequent two classes are observed in event logs in a sequence, the more significant their relationship is. Another metric that can be used to identify relationships is distance significance which is used for isolating behaviour of interest. Binary correlation measures the distance of events in a precedence relation, for example, how closely related two events following one another are.

Binary correlation is the key indicator used for the decision between aggregation and abstraction of less-significant behaviour.

Let N be the set of nodes in a process model, and let the matrix sig: $N \times N$, be a relation assigned to each pair of nodes $A,B \in N$. The *relative significance* of their ordering relation is as follows [67].

$$rel(A,B) = 1\ 2\ (\ sig(A.B) \sum sig(A,X)\ X \in N\ ) + 1\ 2\ (\ sig(A.B) \sum X \in N sig(X,B)\ ) \tag{2.2}$$

Relative significance can be used to determine the offset between both (A,B) and (B,A) relations' relative significances. For example, the offset gap can be expressed as follow.

$$ofs(A,B) = |\ rel(A,B) - rel\ (B,A)| \tag{2.3}$$

Relative significance, an offset between $A,B \in N$, and thresholds can be used to identify the relationship between nodes and edges of the process models. For example, if $rel(A,B)$ and $rel(B,A)$ exceed a specified preserve threshold value, *(A,B)* and *(B,A)* will be preserved for the simplified process models. Or, in the case of one of the conflicting relation's relative significance is below this threshold and $ofs(A,B)$ also exceeds the specific ratio, the relative significance of both conflicting relations differs. Thus, we can remove the less significant relation out of the simplified process model. Otherwise, i.e. if at least one of the relationships has a relative significance below the preservation threshold and their offset is less than the ratio threshold, both relations can be determined as concurrent. However, both edges are removed from the simplified process models as they do not meet the factual ordering relation. The algorithm is implemented in various software such as Disco, Prom, and Celonis and can be adjusted accordingly to the functionality. The sample of the algorithm applied in Disco can be found in Appendix A and F.

Directly-follows graph (DFG)

To construct DFG we should follow next terms [24].

An aa-event is an event that corresponds to the activity aa. A *trace* (also called *process variant*) $\sigma\sigma = \langle aa_1, aa_2, aa_3, \dots, aa_{nn} \rangle$ is a sequence of activities. $\#EL(\sigma\sigma)$ is the number of cases in the event log EL that correspond to the trace $\sigma\sigma$. Note that many cases may have the same trace. $\#\,EL\,(aa)$ is the set of aa-events in the event log EL. $\#\,EL\,(aa, bb)$ is the number of times an aa-event is directly followed by a bb-event within the same case. Without loss of generality, we assume that each case starts with a start event (denoted ▶) and end with an end event (denoted ■).

If such start and end activities do not exist, they can be added to the start and end of each case. Hence, traces (process variants) are of the form $\sigma\sigma = \{$▶$, aa2, aa3, \dots, aann{-}1,$ ■$\}$ where the start and events only appear at the start and end.

A DFG is a graph with nodes that correspond to activities and directed edges that corresponds to directly-follows relationships. There are three parameters, i.e., $\tau\tau vvvvvv$, $\tau\tau aaaaaa$, and $\tau\tau dddd$, that define thresholds for the minimal number of traces for each variant included (based on $\#\,EL\,(\sigma\sigma)$), the minimal number of events for each activity included (based on $\#\,EL\,(aa)$), and the minimal number of direct successions for each relation included (based on $\#\,EL\,(aa, bb)$).

1. Input: event log EL and parameters $\tau\tau vvvvvv$, $\tau\tau aaaaaa$, and $\tau\tau dddd$.
2. Remove all cases from EL having a trace with a frequency lower than $\tau\tau vvvvvv$, i.e., keep all cases with a trace $\sigma\sigma$ such that $\#\,EL\,(\sigma\sigma) \geq \tau\tau vvvvvv$. The new event log is EL'. Note that the number of cases may have be reduced considerably, but the retained cases remain unchanged.
3. Remove all events from EL' corresponding to activities with a frequency lower than $\tau\tau aaaaaa$, i.e., keep events for which the corresponding activity aa meets the requirement $\#EL'(aa) \geq \tau\tau aaaaaa$. The new event log is EL''. Note that the number of cases did not change, but the number of events may be much lower.
4. Add a node for each activity remaining in the filtered event log EL''.
5. Connect the nodes that meet the $\tau\tau dddd$ threshold, i.e., the activities aa and bb are connected if and only if $\#\,EL''(aa, bb) \geq \tau\tau dddd$.
6. Output the resulting graph. The nodes are decorated with the activity frequency $\#\,EL''(aa)$ and edges are decorated with the directly-follows frequency $\#\,EL''(aa, bb)$. Nodes and edges can also be decorated with timing information. Note that an edge connecting activities aa and bb corresponds to $\#\,EL''(aa, bb)$ observations of activity aa being followed by activity bb. It is easy to compute the sum, mean, median, minimum, maximum, and standard deviation over these $\#EL''(aa, bb)$ observations.

The sample of the notation can be found in Appendix A and F.

Quality criteria.

There are some issues regarding the process discovery algorithm.

The first one is noise and incomplete data as a common issue. In this context, noise data does not mean incorrect data. Noise refers to infrequent behavior or outliers. Process discovery techniques should be able to present the majority of traces and filter noise. The approaches which are robust to noise are heuristic mining, genetic mining, and fuzzy mining [3, 25]. On the contrary, an incomplete data refers to having too little data. Similarly, in the data mining and machine learning, we cannot assume to have seen all possibilities from training data, particularly when the amount of training data is limited. Process models can be evaluated by various dimensions.

Completeness and noise refer to qualities of the event log and do not say much about the quality of the discovered model. Commonly, there are four dimensions for the determination of the quality of the resulting models: fitness, precision, generalization, and simplicity [3, 25].

- *Fitness*. The discovered model should allow for the behavior seen in the event log. A model with a good fitness allows for most of the behavior seen in the event log. A model has a perfect fitness if all traces in the log can be replayed by the model from beginning to end.
- *Precision*. The discovered model should not allow for behavior completely unrelated to what was seen in the event logs (avoid underfitting). Underfitting is the problem that the model overgeneralizes the example behavior in the log (i.e., the model allows for behaviors very different from what was seen in the log).
- *Generalization*. The discovered model should generalize the example behavior seen in the event logs (avoid overfitting). Overfitting is the problem that a very specific model is generated whereas it is obvious that the log only holds example behavior (i.e., the model explains the particular sample log, but a next sample log of the same process may produce a completely different process model).
- *Simplicity*. The discovered model should be as simple as possible.

In Figure 2.7, it is presented the key criteria for evaluating the quality of process models with appropriate answers that explain the meaning of each criteria.

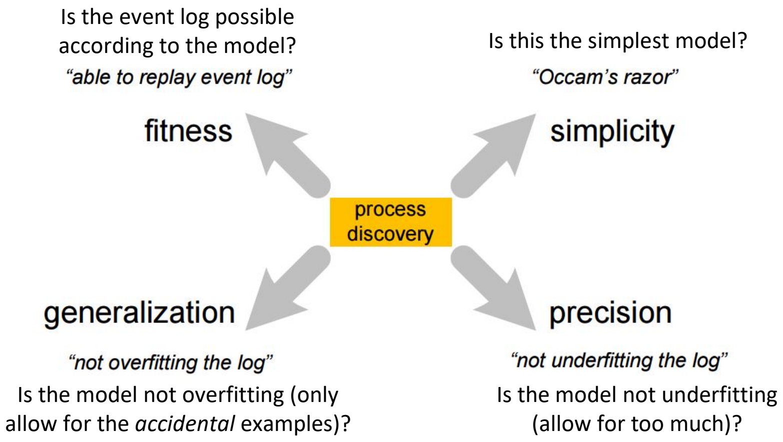

Is the event log possible
according to the model?
"able to replay event log"
fitness
Is this the simplest model?
"Occam's razor"
simplicity
process discovery
generalization
"not overfitting the log"
Is the model not overfitting (only allow for the *accidental* examples)?
precision
"not underfitting the log"
Is the model not underfitting (allow for too much)?

Figure 2.7: The main quality criteria for process discovery algorithms and the questions they raise [3]

Balancing fitness, simplicity, precision and generalization is challenging. This is the reason that most of the more powerful process discovery techniques provide various parameters. Improved algorithms need to be developed to better balance the four competing quality dimensions. Moreover, any parameters used should be understandable by end-users.

2.1.4 Refined process mining framework

A process model is extended or improved using information extracted from event logs. As seen before, event logs contain much more information that goes far beyond just control-flow, namely information about resources, time, and data attributes, etc. The organizational mining can be used to get insight into typical work patterns, organizational structures, and social networks. Timestamps and frequencies of activities can be used to identify bottlenecks and diagnose other performance-related problems. Case data can be used to better understand decision-making and analyze differences among cases.

The different perspectives might be merged into a single integrated process model for the next simulation and *what if* analysis to explore different redesigns and control strategies. In the Figure 2.8, it is presented an approach to come to a fully integrated model covering the organizational, time, and case perspectives [25].

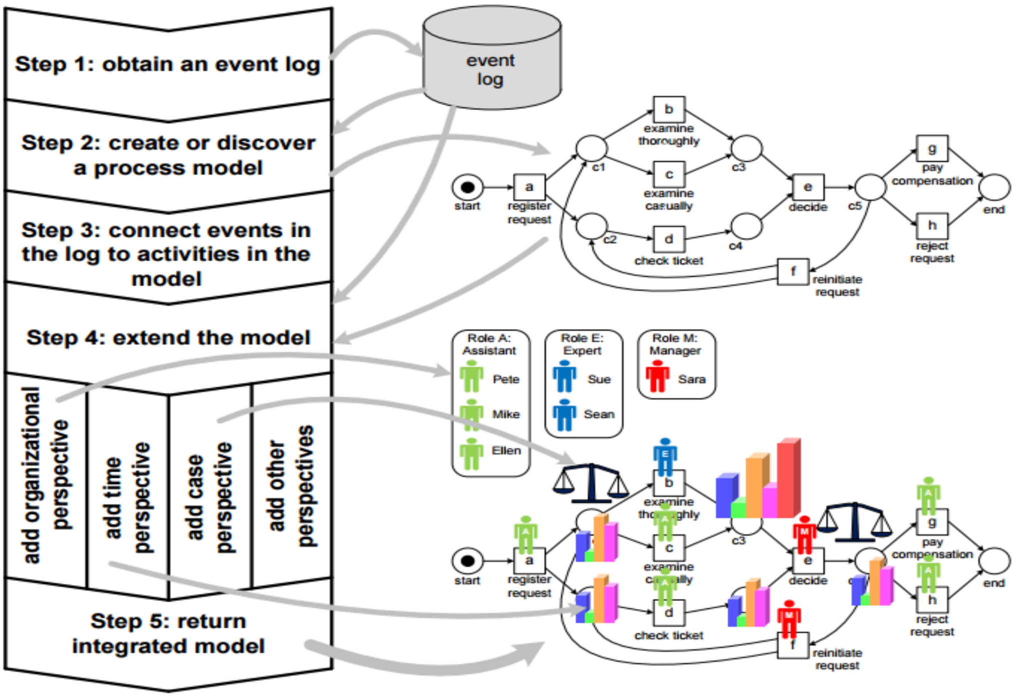


Figure 2.8: Connecting event log and model (with extension) [3]

Today, many data are updated in real-time and a sufficient computing power is available to analysis events when they occur. Therefore, PM should not be restricted to off-line analysis and can also be used for online operational and value stream support.

Figure 2.9 shows a refined PM framework, which can be extended. A provenance refers to the data that is needed to be able to reproduce an experiment. Data in event logs are portioned into *pre mortem* and *post mortem*. *Post mortem* – information about cases that have been completed and can be used for process improvement and auditing, but not for influencing the cases. *Pre mortem* – cases that have not yet been completed and can be exploited to ensure the correct or efficient handling of the cases.

PM can be seen as the *maps* describing the operational processes of organizations. Group *Cartography* includes three activities:

- *Discover*. This activity is concerned with the extraction of (process) models.
- *Enhance*. When existing process models (either discovered or hand-made) can be related to events logs, it is possible to enhance (extend and repair) these models.
- *Diagnose*. This activity does not directly use event logs and focuses on classical model-based analysis.

Group *Auditing* obtains a set of activities used to check whether the business processes are executed within certain boundaries set by managers, governments, and other stakeholders.

- *Detect*. Compares de jure models with current *pre mortem* data. The moment a predefined rule is violated, an alert is generated (online).

- *Check*. The goal of this activity is to pinpoint deviations and quantify the level of compliance (offline).
- *Compare*. De facto models can be compared with de jure models to see in what way reality deviates from what was planned or expected.
- *Promote*. Promote parts of the de facto model to a new de jure model.

And the last category is *Navigation*. It is forward-looking and helps in supporting and guiding process execution (unlike the *Cartography* and *Auditing*).

- *Explore*. The combination of event data and models can be used to explore business processes at run-time.
- *Predict*. By combining information about running cases with models, it is possible to make predictions about the future, e.g., the remaining flow time and the probability of success.
- *Recommend*. The information used for predicting the future can also be used to recommend suitable actions (e.g. to minimize costs or time).

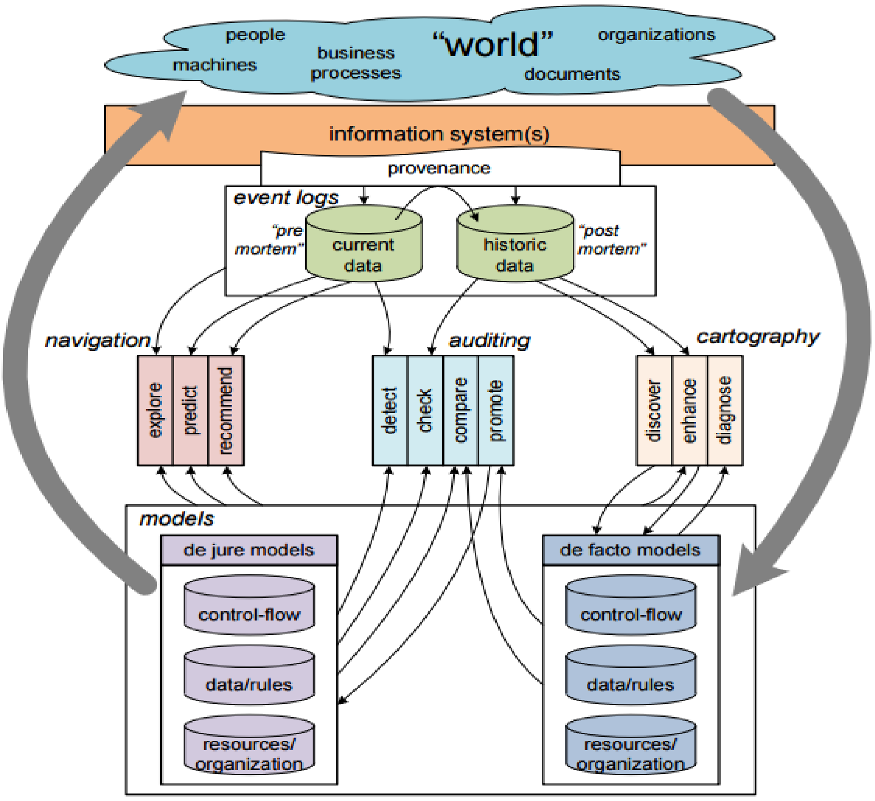


Figure 2.9: Refined PM Framework [3]

Using Figure 2.9, we are focusing on the path: historic data → discover a descriptive model→ control flow &data/rules perspective → promote → all navigation category.

Depending on the goals of our analysis, the navigation category plays the main role in the work. In terms of operational support, it brings important insights to the process analysis and enhances the process model potential. Operational support can be considered in three directions [25] and be applicable to the current *online* data [68].

Detection

Figure 2.10 illustrates a type of operational support. Users are interacting with some enterprise information system. Based on their actions, events are recorded. The partial trace of each case is continuously checked by the operational support system, which immediately generates an alert if a deviation is detected. The same alarm can be applied for bottleneck detection, which will be described in the next section.

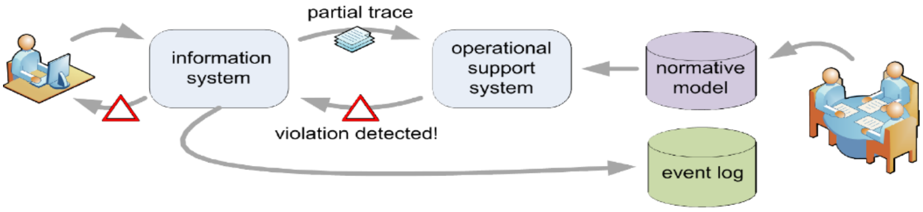
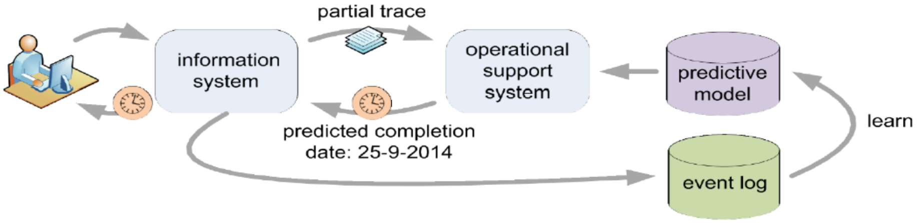


Figure 2.10: Detecting violations at run-time [3]

Prediction

Taking the same setting in which users are interacting with some enterprise information system, but from prediction part (Figure 2.11). The events recorded for cases can be sent to the operational support system in the form of partial traces. Based on such a partial trace and some predictive model, a prediction is generated. Further, we will use one additional parameter – context-aware information about process variants.



partial trace
information system
operational support system
predictive model
learn
predicted completion date: 25-9-2014
event log

Figure 2.11: Both the partial trace of a running case and some predictive models are used to provide a prediction [3]

Recommendation

The setting is similar to prediction. However, the response is not a prediction, but a recommendation about what do next (Figure 2.12). To provide such a recommendation, a model is learned from *post mortem* data. A recommendation is always given with respect to a specific goal. For example, to minimize the remaining flow time or the total cost, to maximize the fraction of cases handled within weeks, etc.

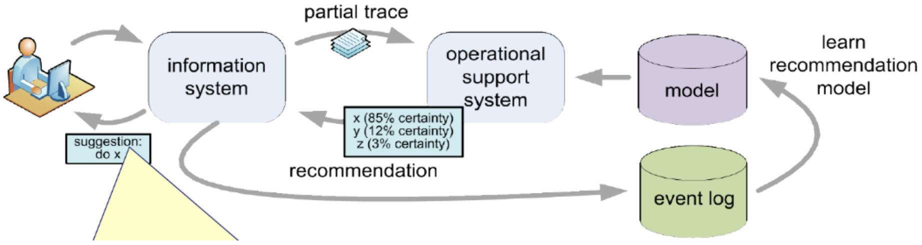


Figure 2.12: A model based on historic data is used to provide recommendations for running cases [3]

2.1.5 Process Mining tools

The main process mining tool for academics remains to be PROM. It is an extensible framework that supports a wide variety of process mining techniques in the form of plug-ins. Availability of adding own programming modules, makes this tool flexible and powerful but the same time slow and vulnerable. Nowadays already exist 600 plug-ins and this amount grows up.

The main characteristics of PROM:

- Aims to cover the whole process mining spectrum.
- Notations supported: Petri nets (many types), BPMN, C-nets, fuzzy models, transition systems, Declare, etc.
- Also supports conformance checking and operational support.
- Many plug-ins are experimental prototypes and not user friendly.

There is also commercial software Disco that has following characteristics:

- Focus on discovery and performance analysis (including animation).
- Powerful filtering capabilities for comparative process mining and ad-hoc checking of patterns.
- Uses a variant of fuzzy models, etc.
- Does not support conformance checking and operational support.
- Easy to use and excellent performance.

Both these tools were used on the phase of process modelling, especially in parts of process exploration, performance analysis and process variants detection.

New tools are appearing nowadays (Celonis, ProcessGold, myInvenio, PAFnow, Minit, QPR, Mehrwerk, Puzzledata, LanaLabs, StereoLogic, Everflow, TimelinePI, Signavio, and Logpickr, etc.) [24]. They exploit the same basic algorithms to discover and analyze processes, which were described above. Also, process mining functionality is being embedded in more business intelligence, business process management, data mining suites, etc. For instance, Qlick add-in for Power BI, process mining modules for Knime, SAP systems, 6 Sigma concept (lean management), open sources Python scripts[8] to develop own software and web applications.

Figure 2.13 shows commonly used process mining tools in terms of process and data perspectives, as well as variety of functions.

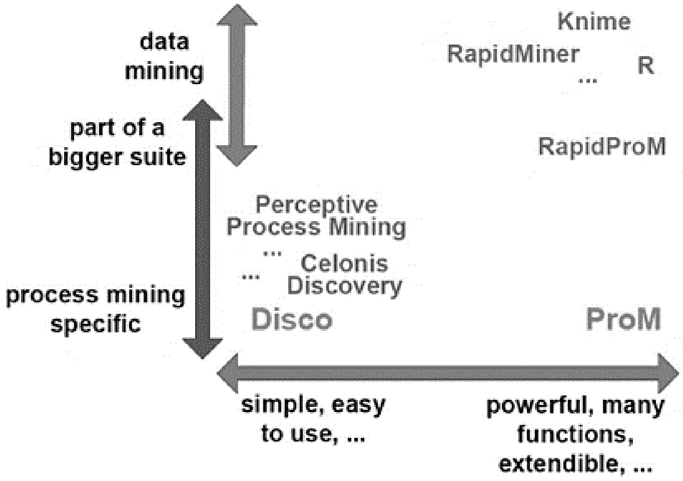


Figure 2.13: Spectrum of process mining tools in terms of usability and process orientation [3]

[8] https://pm4py.fit.fraunhofer.de/

For a quick overview of possible abilities of different tools, Table 2.2 presents advantages and disadvantages of some of them.

Table 2.2: Example of process mining tools with analysis of prons and cons

Product name	Noted advantages	Noted disadvantages
Reflect One	Supported BPM life-cycle, reflect user-friendliness, scalability, support organizational mining by social networks, Discovery algorithms: on a genetic mining, on a sequential model.	Not support conformance checking and prediction
Disco	Focus on high performance, support seamless abstraction and generalization using the cartography, deal with complex (Spaghetti-like) processes, have Nitro, that can recognize different time formats and automatically maps these onto XES or MXML notation. Algorithm based on the fuzzy mining.	
Enterprise Visualization Suite	Focus on analysis of SAP supported business processes Process discovery algorithm inspired by α-algorithm and heuristic mining	
InterStage BPME	Not need to install, focus on process discovery, able to seamlessly abstract from infrequent behavior, able to analyze performance using indicators such flow time	Unable to discover concurrency, not support prediction, recommendation and conformance checking
ARIS Process Performance Manager	Focus on performance analysis (drilling down to the instance level, benchmarking, dashboards), support organizational mining	Not support prediction, recommendation and conformance checking
Genet/Petrify Rbminer/Dbminer	Use state/based regions, support control-flow discovery, rely on Prom for conformance checking	
Service Mosaic	Analysis of service interaction logs, discovers transition systems, focus on dealing with noise and protocol refinement	Unable to discover concurrency

2.1.6 Types of processes

In terms of complexity, structure and representation, processes can be divided to two groups:

1. *Lasagna processes* are relatively structured, repetitive and regular. The cases flowing through such processes are handled in a controlled manner. Therefore, it is possible to apply all of the PM techniques presented in the preceding sections. Lasagna process can be defined as a process, for which it is possible to create with limited efforts an agreed-upon process

model that has the fitness of at least 0.8, i.e., more than 80% of the events happen as planned and stakeholders confirm the validity of the model. Figure 2.14 presents the example of lasagna process in Petri net and BPMN notations.

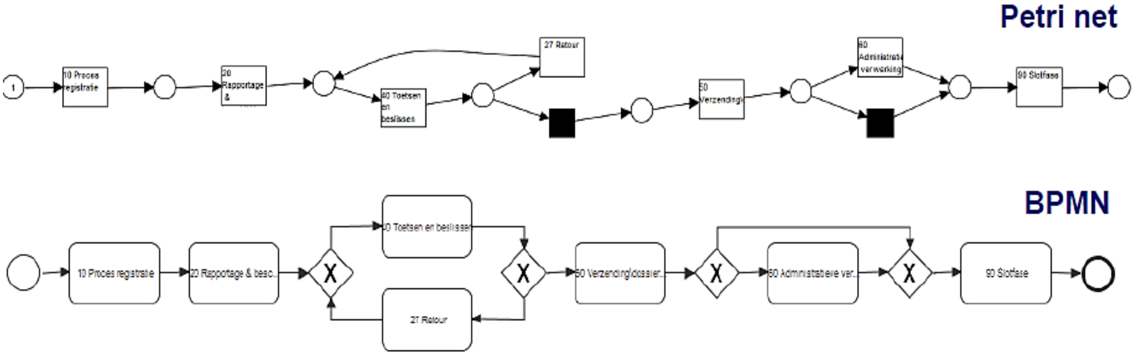


Figure 2.14: Example of usual *lasagna* process

The main characteristics of lasagna processes are:

- Easy to discover, but it is less interesting to show the *real* process. The resulting process model will be close to the expectation; however, all deviations will be visible immediately.
- Whole process mining toolbox can be applied.
- Added value is predominantly in more advanced forms of process mining based on aligning log and model.

2. *Spaghetti processes* (see Figure 2.15) are unstructured, irregular, flexible and variable processes, only some of the process mining techniques can be applied. The figure 2.15 shows why unstructured processes are called spaghetti processes. There are different approaches to getting valuable analysis from this type of process. For example, the method *divide and conquer* (by clustering of cases), which will be described in the next chapter, or showing only the most frequent paths and activities.

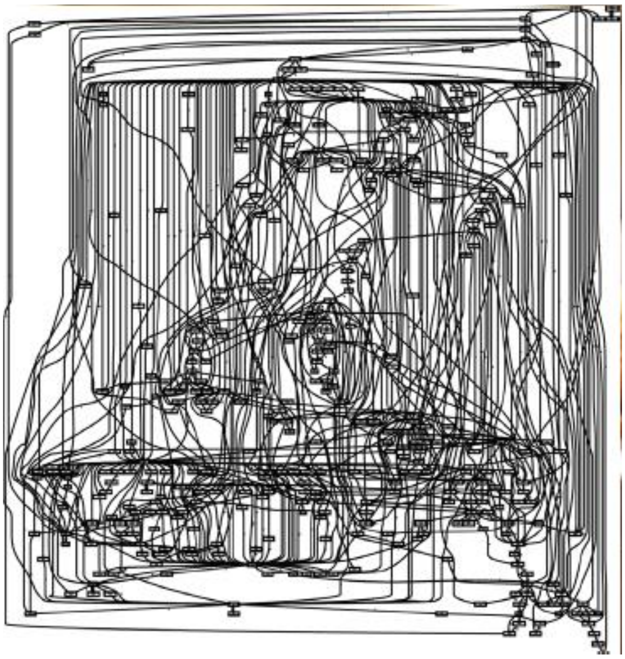


Figure 2.15: Example of *spaghetti* process [3]

The figure 2.16 depicts the overview of the different functional areas in a typical organization. Lasagna processes are mostly encountered in production, finance/accounting, procurement, logistics, resource management, and sales/CRM. Spaghetti processes are typically encountered in product development, service, resource management, and sales/CRM.

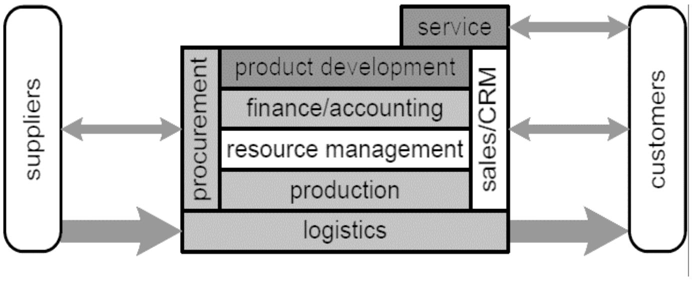


Figure 2.16: Application of spaghetti (white cells), Lasagna (grey cells) processes and both (dark-grey cells) [3]

Nevertheless, spaghetti processes are very interesting in the view of PM as they often allow for various improvements. Highly structured well-organized process analysis is less effective in this respect; it is simpler to apply PM techniques but there is also not significant improvement potential. However, the practice shows that functional area does not define the process type as much as feature selection. Also, choosing different case and process perspectives can radically change the whole picture and representation will be more confusing.

2.1.7 Existing issues and challenges

As a quite immature field, process mining currently has a set of challenges, issues and open questions. According to [26], the next challenges are still relevant and require more research:

C1: *Concept drift*. The term concept drift refers to the situation in which the process is changing while being analyzed. Algorithms need to recognize processes based on additional parameters like weather, seasons, economic crisis, epidemics, etc.

C2: *Finding, Merging, and Cleaning Event Data*. This issue has a crucial impact on all analysis and even best algorithms are not able to give appropriate results without wrong input data.

- Data may be distributed over a variety of sources.
- Event data are often *object centric* rather than *process centric*.
- Event data may be incomplete.
- An event log may contain outliers, i.e., exceptional behavior also referred to as noise
- Events occur in a particular context (weather, workload, day of the week, etc.)

C3: *Dealing with complex event logs having diverse characteristics and cases.* If the case is the item inside the package in the parcel, how can the model show only the process for the item.

C4: *Balancing between quality criteria such as fitness, simplicity, precision, and generalization.* Improved algorithms need to be developed to better balance the four competing quality dimensions (fitness, simplicity, precision and generalization). Moreover, any parameters used should be understandable by end-users.

C5: *Cross-organizational mining and providing operational support.* Process mining should not be restricted to off-line analysis and can also be used for online operational support.

C6: *Combining process mining with other types of analysis.* Both fields (operations management and data mining) provide valuable analysis techniques. The challenge is to combine the techniques in these fields with process mining. Also, it is desirable to combine process mining with visual analytics. By combining automated process mining techniques with interactive visual analytics, it is possible to extract more insights from event data.

***C7:** Improving usability and understandability non-experts*. The challenge is to hide the sophisticated process mining algorithms behind user-friendly interfaces that automatically set parameters and suggest suitable types of analysis.

Moreover, there are prominent issues with process models' behaviour:

- *There are no negative examples.* An event log shows what has happened but does not show what could not happen. For example, we only know about sick customers that complained afterwards.
- *Due to concurrency, loops, and choices the search space has a complex structure, and the log typically contains only a fraction of all possible behaviors*. In this case, we can talk about spaghetti process models, which should be treated carefully in order not to lose important deviations.
- *There is no clear relation between the size of a model and its behavior* (i.e., a smaller model may generate more or less behaviour although classical analysis and evaluation methods typically assume some monotonicity property).

Some of these challenges and issues were partly solved, while others are still in progress. As a fast-growing area, process mining getting new challenges for academics and analysts. Recently, novel process mining capabilities have been identified [27]. These provide new scientific and practical challenges (see Figure 2.17).

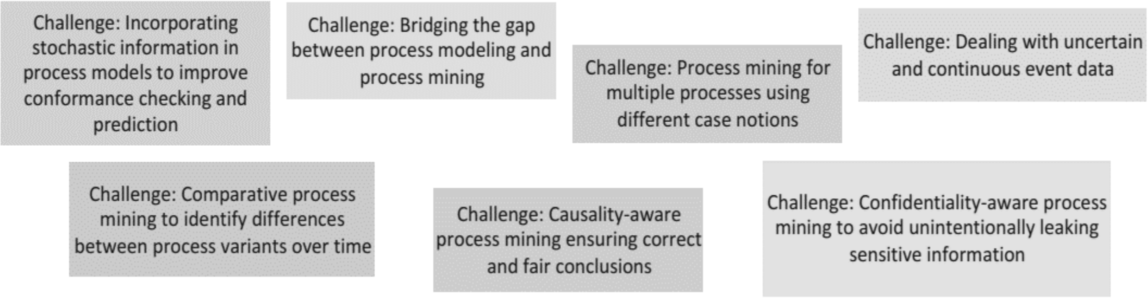

Challenge: Incorporating stochastic information in process models to improve conformance checking and prediction
Challenge: Bridging the gap between process modeling and process mining
Challenge: Process mining for multiple processes using different case notions
Challenge: Dealing with uncertain and continuous event data
Challenge: Comparative process mining to identify differences between process variants over time
Challenge: Causality-aware process mining ensuring correct and fair conclusions
Challenge: Confidentiality-aware process mining to avoid unintentionally leaking sensitive information

Figure 2.17: Some of the challenges getting more attention in research and thus showing the anticipated development of the PM discipline [27]

- Bridging the gap between process modelling and process mining. Many organizations use tools for modelling processes. With the uptake of process mining, it becomes clear that these models do not correspond to reality. Although such idealized models are valuable, the gap between discovered and hand-made models needs to be bridged. For these goals, the hybrid process models can be used as a backbone formally describing the parts of the process that are clear and stable, and less rigid data-driven annotations to show the things that are less clear [28].
- Incorporating stochastic information in process models to improve conformance checking and prediction. Frequencies of activities and process variants are essential for PM. A highly frequent process variant (i.e., many cases having the same trace) should be incorporated into the corresponding process model. This is less important for a process variant that occurs only once. One can view frequencies as estimates for probabilities, e.g., if 150 cases end with activity reject and 50 cases end with activity accept, then the data suggests that there is a 75% probability of rejection and a 25% probability of acceptance. Such information is typically not incorporated in process models.

- PM for multiple processes using different case notions. Traditional PM techniques assume that each event refers to one case and that each case refers to one process. Since the data in event logs need to be *flattened*, it's not available to reflect the more complex reality [29]. For example, the information system holds information about customer orders, products, payments, packages, and deliveries scattered over multiple tables. An object-centric process mining relaxes the traditional assumption of process mining each event relates to one case. The relation *1 to 1* is quite rare in the world. If we look at the supply chain process from relations perspective, it is as follows:

  Order → (one-to-many) → Item → (many-to-one) → Package →
  → (many-to-many) → Route

  This challenge is explored more in [30].
- Dealing with uncertain and continuous event data. The starting point for any PM effort is a collection of events. Normally, we assume that events are discrete and certain, i.e., we assume that each event reported has actually happened and that its attributes are correct. However, due to the expanding scope of PM, other types of event data are encountered. Events may be uncertain, e.g., the time may not be known exactly, the activity is not certain and there may be multiple candidate case identifiers. Future PM tools and approaches will need to be able to deal better with uncertain event data and measurements that are continuous in nature.
- Comparative PM to identify differences between process variants over time. Processes change over time and the same process may be performed at different locations. This requires techniques to support the visual comparison of process variants and machine learning techniques using process-centric features. One of the research works [31] described the way to make automatically groups/clusters of process variants.
- Causality-aware PM ensuring correct and fair conclusions. Analytics talk about *FACT* issue (fairness, accuracy, confidentiality and transparency). Data mining and machine learning techniques might be able to uncover root causes for performance and compliance problems. Such a combination of techniques yields a powerful approach to automatically diagnose process-related problems. Relations between process features can be made explicit using a mixture of domain knowledge and statistical evidence.
- Confidentiality-aware PM to avoid unintentionally leaking sensitive information. Event data are potentially very sensitive. A few timestamped events are often enough to identify a customer or employee. Current research aims at dedicated anonymization and encryption techniques. One of possible solution to encrypt data can be found in [32].

In the case of comparing these two lists of challenges, we can see that the field of PM rapidly grows and develops and requires more and more research work. Both of them focus on the importance of combining process mining with other types of analysis like statistics, data mining and machine learning to improve results, fix issues and overcome some challenges. The aim of the work is C6: Combining PM with other types of analysis, considering FACT problematic and feature selection.

2.1.8 Context-aware information and feature selection

A *context-aware system* is a system that is able to provide suitable service or information with regard to a user's requests [33, 34]. There are many authors who define the meaning of context differently. We consider the following definition: "Context is characteristics which describe the situation of an entity. An entity refers to a person, place, or object which is related to the interaction between user and application" [35]. The logistics and manufacturing domains are ideal candidates to include context awareness features because industrial and supply chain activities happen in very heterogeneous environments with a multitude of information available besides the actual observed process. Context data can be presented as time, location, and frequency of events as well as related communication, tools, devices, or operators.

Feature selection is a time-consuming step and needs to be carefully performed because all algorithms are sensitive to input data and the results of the analysis are based on which features, we are using from the beginning. Features of the process can be not only selected but created as well. We can differ three groups of features for analysis. These groups can be extended in terms of the type of the process. The examples of these types in relation to the logistics process are as follows:

- *standard features* of the logistics process: wind speed, ship size, day of arrival (which day in the week), country of origin, traffic jams, SLA (Service Level Agreement) measurements (window time for each level of service, cost trade-offs, frequencies), etc.;
- *quantitative features*: # of ships/containers/operators/resources on the line;
- *qualitative features*: productivity or performance (processed containers/hour, processed ships/hour), velocity(activities/hour).

Additional context parameters – information about the system, types of the process (process variants), set of risk/weak points, number of operations, start/end operation, source of data, rework, deviations, skipped process activities – play an important role in additional advanced analysis. Also, such psychological parameters as the operator is concerned and its way of working can be considered as features with a high impact on the whole process. In the Chapter 3, we will offer the framework for using context awareness in terms of process mining types.

Summary of the section

Process modelling is necessary in order to display the entire process from the beginning to end with various levels of detail and the ability to enrich the model with additional attributes in the future. To date, there are many different methods for advanced analysis of the process. However, the process mining approach has a lead role and differs from others by having flexible and powerful tools. It provides an opportunity to get a descriptive process model based on raw data from different sources like documentation, ERP systems, Intranet, etc. The descriptive type of model shows the current process and is made by algorithms based on raw data. The obtained model is used for optimization, improvement, and enhancement of the real process and can be the basis for recommendation or decision-making systems as well as for process maps.

In this section, we introduced process mining terminology and insights. The most important information about process mining, its types, algorithms, and techniques, which are related to the current research, are described here. Based on the type of analysis used for this work, we chose Fuzzy mining algorithm and DFG notation. Fuzzy mining is deployed for our use case as it is robust to less structured and complex processes. As a result, fuzzy mining is capable of providing understandable process models from dynamic logistics processes. Fuzzy mining gives a high-level view on the process and abstracts from undesired details. In a turn, DFG notation is also highly recommended to work with *spaghetti* models and brings transparent understanding. It's important to note, that both techniques have limitations and should be applied carefully to other types of processes.

In this thesis, we focus more on *Discover* and *Enhancement* types of process mining. Once modelled behaviour and real behaviour are aligned it gives the opportunity to include other perspectives of process mining, what can be beneficial in terms of online prediction, automated process improvement, resource recommendation, etc. The current work aims mostly to control-flow, case, and time perspectives. Talking about the refined process mining framework (see Figure 2.9) the path of the work can be defined as follows: historic data → discover a descriptive model → control flow &data/rules perspective → promote → all navigation category.

As a fast-growing area, process mining getting new challenges for academics and analysts. Solving issues brings up more perspectives and a deeper view of the process nature, which creates more challenges. One of the listed above relevant challenges focuses on the importance of combining process mining with other types of analysis like statistics, data mining and machine learning to improve results, fix issues, and create recommendation systems and process maps. The aim of the work is the challenge 6: *Combining process mining with other types of analysis, considering FACT problematic and feature selection* – context awareness. In the Chapter 3, there will be discussed how to deal with spaghetti models by context awareness in terms of different types of process mining. And, after, use the obtained process model to enhance it with the results of another analysis.

Moreover, there is a limited amount of studies regarding process mining in logistics and only a few cases in logistics applications. Therefore, this research objective is to discuss the application of process mining, its advantages, as well as the challenge in the logistics domain. The next section is dedicated to the domain knowledge – logistics processes.

2.2 Current state and development prospects of modern seaport logistics

Logistics is an industry consisting of process-oriented business that focuses on managing the flow of resources, both material and abstract resources, between the point of origin and the point of destination. The process of logistics service providers remains to be highly human-centric, flexible and complex, which leads to a set of uncertainties. Standard operating procedures in the form of normative process models are commonly developed by logisticians in order to control the workflow and keep a high level of service. Hence, research has demonstrated the existence of a discrepancy between the procedures in the designed processes and the actual processes [36]. This discrepancy might result in serious operational risks. Therefore, logistic companies could benefit from acquiring a full insight into the actual process executions in order to enable logisticians to improve and revise the designed processes.

The optimization of the port operations can be archived in two ways:

1. reorganizing currently existing resources of infrastructure or the construction;
2. developing the new resources/infrastructure.

In the first case, optimality is gained through a set of organizational and technical procedures and small investments, which allow speeding up in the shortest possible time, in the second - due to large investments, long-term constructions and commissioning of new facilities. For row reasons (including economic ones) the first way of optimization is the most acceptable. In addition, the share of the transport component in the final product cost reaches 30% or more, as a result of which the state is interested in developing ports and accelerating the processing of the fleet in order to increase competitiveness in the world market. Reducing the layover time of vessels under processing, even with additional costs, can lead to significant fleet savings.

Contemporary logistics information systems record detailed information about the events happening in the environment. These events are occurrences of importance in the context of logistics management, e.g. the arrival of a certain ship. Consequently, the collection of events, i.e. event log, contains an untapped reservoir of knowledge about the logistics processes [14]. In this part of the work, we will present the nature of logistics processes, their members and characteristics. Also, problematics and prerequisites of applying advanced process mining techniques are shown in detail.

2.2.1 Seaport logistics process description

Seaports are important interfaces in the supply chain between sea and land transportation and a component of freight distribution as the entrance of produce, merchandise, and passengers to a country. To better understand the place of a seaport in the supply chain, the cargo transportation scheme and its important participants are presented in Figure 2.18. Based on Figure 2.18, three main roles can be recognized:

- *Agent* is a person/company, who is held responsible for handling shipments and cargo.
- *Freight forwarder* – person/company that organizes shipments for individuals or corporations to get goods from producer to a market.
- *Customs broking* – a profession that involves the *cleaning* of goods through customs barriers.

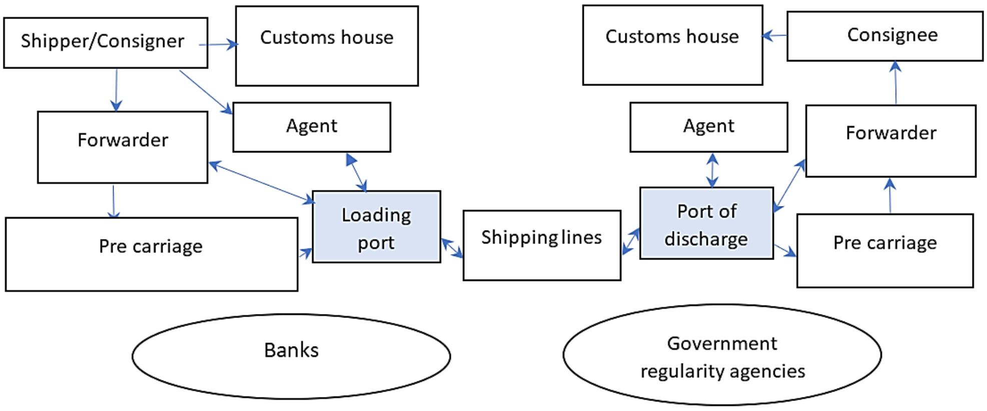


Figure 2.18: General Cargo transportation scheme

Every seaport has its own type of operations, but, in common, there are main operations of each seaport:

- Processing all kinds of general loading cargo.
- Reloading processes.
- Warehouse's operations.
- Carriage of passengers.
- Berthing.

The main task of the harbor is *loading*, *unloading* and *transshipment of cargo* to and from the vessels. Reloading processes are the main port activity, which includes all loading and unloading operations for different kinds of transport, such as ships, vans, wagons and cars (see Figure 2.19). In order to implement freight handling with significant freight traffic, the port is equipped with cargo gear that has different load capacity, dock system, warehouses, transport routes net, and so on.

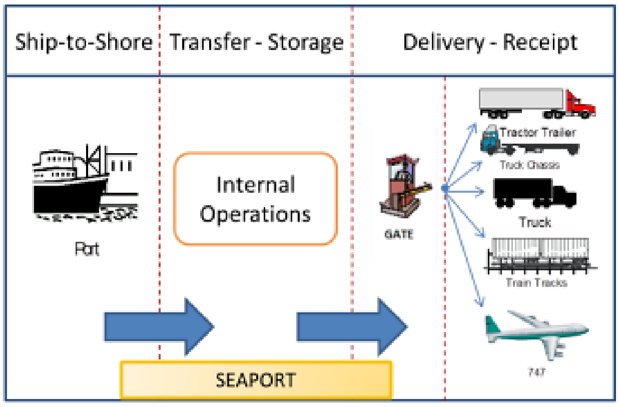


Figure 2.19: The operations for container seaport divided for processes parts[9]

[9] www.logistics.gatech.pa

Generally, in seaports, it is required considerable acceleration of reloading processes, reduction of the transport laytime and decrease of complex costs for ships and port. At the same time, the complexity and the particularity of loading and unloading operations require certain improvements in reloading process management using information technology, intelligent algorithms and methods. When loading and unloading processes, it is important to keep the principle of the minimum route and identical time for all transports. Figure 2.20 presents only a high-level view of the process inside the port. Most of the time, lower levels of the process with detailed information remain to be hidden and represent the so-called *black box*.

The ship, standing in the port, is needed to complete reloading processes as soon as possible and leave the berth (parking place). In this case, the ship gets a bonus, so-called dispatch money. Otherwise, it will be fined (in other words, it gets demurrage). First of all, for the unloading of freight flows it is necessary to determine the dock so that its depth would correspond to the ship's tonnage. Also, this dock needs to have a port crane, corresponding to a special kind of cargo flow. Then, one needs to determine the warehouse for cargo storage, in preference, this warehouse needs to be the closest one to a certain berth. Thus, the route will be minimal and, as a result, it will be reduced energy consumption and fuel. The most important of reloading operations is to organize transport and use every machine with a maximum load capacity.

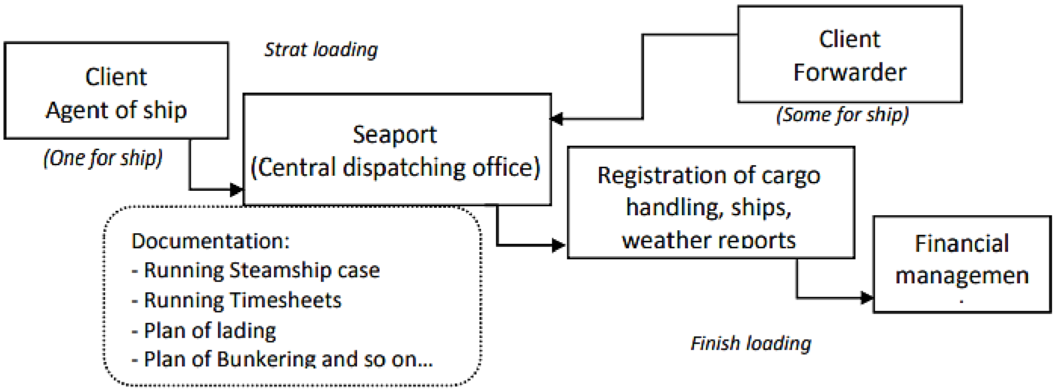


Figure 2.20: Workflow inside the port (including documentation inside the central dispatching office)

Acceleration of processing vehicles at the port and reducing the laytime promote the increase in seaport competitiveness. That is why every seaport needs to improve the technology and organization of loading and unloading processes. Throughout our study, we focus on the loading/reloading process, which can be called the ship-handling process.

2.2.2 The urgency of the problem

The growth in container transportation and the limited capacity at the major terminals in the port require new approaches for process optimization and control. Nowadays, long waiting times for vessels at the major terminals (up to 48 hours) affect the transit times of the containers. The delays made barge operators decide to raise their transport tariffs by about 10-20%. A quick solution to the problems is not expected so far [37, 38].

The lack of coordination and information sharing results in little transparency and a general lack of trust between the parties involved. For authorities, the lack or inaccuracy of information poses a security risk and an increased workload. Similarly, the inaccuracy of information and delays also affect operational aspects along the supply chain. Many of these challenges point to issues of governance and business processes for inter-firm coordination that do not properly account for the complexity in the supply chain of international trade [39].

The uncertainty in the sojourn time of a ship in the port nowadays is mainly determined by uncertain waiting and handling times at berths. For a terminal operator, it is important to utilize the terminal resources as efficiently as possible. The uncertainty in arrival times of barges implies uncertainty in the quay schedules and the risk of idle time of the quay resources. Moreover, uncertainty in the quay schedules causes uncertainty in the processes that precede the barge handling, e.g., the stacking of containers at the quays. Delays at one operation cause the appearance of a number of operations that significantly increase the total handling time and affect the operations of other berths. This is a kind of dominoes effect, which is reinforced when barges have planned terminal visits close after each other [38]. Current European logistics projects such as SmartPort, SwarmPort, Indeep, IoT for Agri keep looking for new methods to structure and control nautical processes to choose the most appropriate one.

One of possible solution can be developing the process map for nautical processes. A particular advantage of the implementation of the unified process map is the ability to get instantly access to data of any seaport substructures. That makes it possible to promptly respond to current challenges and to successfully predict the further operational activities of the port complex.

Thus, summing up the above, the motivation for logistics processes can be defined as follows:

Well-organized chain of nautical services → Involvement of port capacities (reduce downtime percentage) → Turnaround time reduction → Service price reduction → Improving the quality of service → Improving the competitiveness of the transport node.

2.2.3 Data issues and limitations

According to the interview, which was taken with the questionary (see Appendix E) from key domain experts of logistics processes, challenges in seaport logistics can be determined as follows:

- Nautical process is a conservative field, where there is strong resistance to the implementation of new methods as well as following customs and politics in most countries. The example of such situation is following. Of the 40 days containers travel from central China to central Europe, they spend around 24 days actually on the move, and 16 going nowhere," says Tim de Knegt of the Port of Rotterdam, the party that supplies information and facilitates the logistics process. "One factor is the extensive paperwork, with the consignment note as the main registration and control document. But it is primarily due to the fact that the parties involved do not share information in real-time. If a ship enters the port half an hour earlier or later, the lorry driver is unable to alter his own process."
- The need for investments to adopt new methods and train staff discourages managers who are used to working with traditional techniques.
- Reluctance to share data with other logistics process members because of trade competition. "The biggest problem with logistics is interoperability - that is, the sharing of data and real-time status between parties," confirms TNO senior advisor Wout Hofman. "A transport operation involves at least 20 to 25 parties: customs, port authorities, stevedore, freight forwarder, road carrier, shipper, consignee, bank, and so on. And these parties do business not just with Rotterdam, but with other ports as well. That is why we would like to move towards having an open environment."
- A low level of standardization in some regions.
- Data is stored in fragmented systems, in various formats and not easily accessible – companies will have to aggregate data sets standardize the format for easy analysis and make them available to the right people; data ingestion and cleaning is the key process.
- Combining data analytics knowledge with deep domain knowledge.
- Interorganizational information sharing systems are outdated and manual processes still prevail in large parts of the supply chain. As authorities often do not have all information readily available, containers stand still almost half of the time of their journey [40].

2.2.4 Research gap in port service systems. Review of relevant logistics process modelling

Research on the topic of port service systems is emerging but is developing alongside two separate streams.

- The port economics literature has recently provided a formal theoretical model showing that collaboration between stakeholders in port service chains is vital to maintaining overall port service performance levels [41]. This research on port performance analysis departs from a microeconomic perspective and provides an interesting theoretical basis to study collaboration within port service chains. As the models include stylized behaviour and business models, their applicability to practice is still limited.
- On the other hand, there is also recent port management research that focuses on design approaches for collaborative port planning and management [42, 43]. These approaches are closely linked to practice and implementation focused. Although valuable when it comes to practicable innovations, the designs are generally not evaluated using rigorous frameworks or quantitative models.
- Agent-based models (ABM's) for port systems are a recent development and still a scarce phenomenon [38]. The approaches presented in the literature have important limitations that prohibit studying ports as complex, self-organizing systems. Until now, there has been no attempt to represent interconnected port processes in a behaviorally validated descriptive model. The emphasis has been on crude, meta-level processes in normative models.
- There are few other types of research which offered to use two-stage stochastic model based on Automatic Identification System (AIS) data [44]; self-adapting multi-agent systems and swarm intelligence [45]; combining evolutionary game theory and Reinforcement Learning [46, 14].

However, further research using descriptive models is needed to extend and improve the behavioral validity of the port system analysis, making it amenable for real-life applications. My research is aimed at filling this gap. In general, there is no existing literature on the mathematical optimization of the structure of the nautical chain of a port. More specifically, the investigation of statistics, data mining and process mining techniques with context-aware information leads to another level of research. It is important to note that the process mining approach is flexible and scalable, so it can be merged with the approaches described above to enhance their results.

2.2.5 Prerequisites for applying advanced mining methods

The reason to take sea-port processes is prevailing recently preconditions, which make it possible to apply and examine prominent algorithms of process mining in a new field.

To apply the algorithms and methods of PM to the logistics domain, the next prerequisites were formed recently:

- Digitalization of many process steps and their parameters. Here we can use recent trend concepts such as Big data (data that contains greater variety, arriving in increasing volumes and with more velocity)[10], Industry 4.0 (the current trend of automation and data exchange in manufacturing technologies)[11], Internet of things (physical objects with sensors, processing ability, software, and other technologies that connect and exchange data)[12], and Smart Cities (a technologically modern urban area that uses different types

[10] www.techtarget.com/searchdatamanagement/definition/big-data

[11] www.i-scoop.eu/industry-4-0/

[12] www.en.wikipedia.org/wiki/Internet_of_things

of electronic methods and sensors to collect specific data)[13]. The data volumes produced by various systems, devices and sensors are continually growing. Databases are collecting and storing data from scanners, documents, forms, information systems and emails, etc. All these data can impact the accuracy of PM methods as well as enhance the detailing of the process map. A great example of this prerequisite is APM Terminals Maasvlakte II located in Rotterdam and opened in 2015. It still remains the world's most advanced fully automated terminal [47]. Thus, building fully automated terminals/workplaces, increasing capacities to collect data and using of last methods to work with Big Data- all of this help to apply Process Mining in a more effective way.

- Search and collecting of relevant data. To build decision-making systems, it is very important to have the maximum number of parameters that affect the process. Also, to work at the global level, these parameters must be unified and have standards around the world. Currently, data collection centers are gaining more and more popularity and combine information from urban infrastructure (traffic jams, emergencies, resource reduction) and basic process documents like bill of lading or notices. The development of such centers, or Port Community Systems (PCS), is facing some obstacles because of the issue of confidentiality and trade competitiveness. It affects the ability to share the collected data with other process participants. Also, data sets are provided by governments and nonprofit organizations via open data portals. Nowadays, local city and state governments maintain data portals containing data sets about road networks, urban infrastructure, city services, amenities, transportation services, weather conditions, etc. Examples of such companies – DNV GL[14] in Germany (standardization), DBH[15] in Germany and Smart Data Factory[16] in the Netherlands (collecting data from all possible sources).
- The ability of PM techniques to provide a clear transparent model on different levels and information for the process owner. All players involved in the port logistics process (such as terminal operators, shipowners, forwarders, port IT, rail, port authority and customs operators) are connected in a complex port communications network and exchange information with each other. Building the readable process model shows all connections and can be presented to all participants of the process. It helps to improve the process model by reducing noises and incorrect inputs and defining special types of process scenarios. Open-source software PROM[17], commercial Disco, Celonics or add-in module for Power BI – Qlick – all these products can create a basic process model from raw data, which is a start point for our research as well.
- Stevedoring companies in the port area share the same port resources. Any infrastructure improvements require significant capital investment. However, there is an option to reorganize and restructure inner processes. At this point, it is mandatory to know not just normative processes and subprocesses but also the ones, which reflect reality. It can help better plan and share port capacities/resources. An example is an idea to create a board game in TNO with seaport operations and to show different participants of the logistics process how collaboration and smart sharing of resources can impact profit for all sides [69].

[13] www.en.wikipedia.org/wiki/Smart_city
[14] www.dnv.com
[15] www.dbh-germany.com
[16] www.smart-data-factory.com/
[17] www.promtools.org

2.2.6 Sources for event logs and context-awareness information

We already discussed possible data sources to extract event logs for process mining in Section 2.1. Here, we want to point out some documents (see Table 2.21), which can be used to obtain needed data for analysis or create new features [48]:

1. *The commercial invoice (CI*) is a banking/transactional document that serves the purpose of providing an accounting record and stating the terms of a transaction between the buyer and the seller.
2. *The Bill of Lading (BoL*) confirms that the goods have been received on board the carrier's vessel. It is a receipt of the goods by the carrier obliging him to deliver the goods to the consignee. It contains generic information about the goods and identifies the vessel and the port of destination. It also serves as proof of the contract of carriage and title to the goods, meaning that the holder of the BoL is the owner of the goods (European Commission). The BoL is created with information from the Packing List, which essentially holds more detailed information about the goods being shipped. The sample is presented in Appendix C1.
3. Countries have differing trade agreements based on whom they are trading with. The *Certificate of Origin (CO*) serves the purpose of applying for tariff preferential treatment. The tariffs depend on the trade relationship of the countries involved in the transaction. In some cases, importing from specific countries requires licenses, permits, authorizations from other governmental agencies, pre-shipment inspections etc.
4. *Import and Export declarations* serve the need of gathering information so that customs authorities can perform the three above-described documents.
5. *Statement/Timesheets*. Statement of Facts (SOF) is a detailed chronological description of the activities of the vessel during the stay in a port: taking the sea pilot onboard, hailing in, locking through (if applicable), mooring, preparing the loading and unloading operations, the actual loading and unloading operations, the amount of load transfer, unmooring and departure). Timesheet is recording all operations with a ship from the moment of its entry into port to its departure. Based on this document, all financial calculations between ship and port organizations are made. The sample is presented in Appendix C2.

Table 2.3: Sources for event log and context-awareness with roles [48]

Document	Requestor	Issuer	Securer
Commercial invoice	Authorities, Importer	Exporter	Forwarder
Bill of Lading	Authorities, Forwarder	Carrier	Forwarder
Certificate of Origin	Import Authorities	Chamber of Commerce	Exporter
Export declaration	Export Authorities	Exporter	Forwarder
Import declaration	Import Authorities	Importer	Inwarder

In this research, we are extracting data from timesheets to create the process model. Chapter 3 is presenting the way of forming an event log for logistics processes and context-aware information to handle complex process models.

Summary of the section

Logistics companies used to pay attention to the optimization of the business processes within their organization – changing infrastructure, dismissal of personnel, reducing investments in innovative projects and so on. This way does not bring so appreciable impact on strategic long-term goals. Some companies already started to integrate the advanced method of process modelling, since deviations between the procedures in the normative processes model and the actual processes might result in serious operational risks. However, the alignment of operations of different companies often requires the sharing of information or giving up part of the control over the operational processes. For many companies, these are difficult issues since misuse can threaten one's competitive position in the chain [38].

Logistics intelligence covers the set of techniques that seek to improve logistical operations with their abilities to reduce the uncertainties and risks in logistics. Being aware of the actual logistic processes and deviations from the planned processes in a dynamic environment is essential for companies in order to gain additional flexibility. Current European logistics projects such as SmartPort, SwarmPort, Indeep, IoT for Agri (all references are on the page 4) keep looking for new methods to structure and control nautical processes to choose the most appropriate one.

In turn, the logistics sector is quite conservative and there is an opinion that any changes in the work process entail unjustified risks, unnecessary waste of time and investments. Nowadays, many maritime businesses operate with a set of proven mature tools that are most often highly human-centric. Thus, the decision-making process is reduced to expert opinion, intuitive assumptions, and a set of available parameters from not connected different information sources. The combination of these factors greatly affects productivity and efficiency of work. Combining knowledge of logistics processes and modern methods of analytics - process mining, data mining, machine learning and statistics – can undoubtedly help improve both fields and adopt algorithms to a wider range of process types. After all, process mining is a quite new, flexible, and progressive technology. Moreover, there are recent preconditions, which make it possible to apply and examine prominent algorithms of process mining in a new field. The results of its use directly depend on the settings and understanding of the analyzed process. A prototype decision support system can be developed for different participants involved in the interconnected logistics processes. Maritime port systems represent a major challenge both from a theoretical and practical point of view and require the development of innovative tools for the analysis and synthesis of such systems.

The main advantage of a building process map for the seaport is the possibility of deep analysis of the existing statistics within a common standard of the enterprise, as well as the possibility of efficient management of the entire system on the basis of predicting the possible problems associated with the transport of cargo through the port infrastructure.

Throughout my study, we focus on the loading/reloading process, which can be called the ship-handling process. In this research, I am extracting data from timesheets to create the process model. Other documents, which we discussed in this part can be the source for additional features for advanced analysis. Chapter 3 is presenting the way of forming an event log for logistics processes and context-aware information to handle complex process models.

Benefits of using the descriptive process model for logistics processes are follows:

- ✓ Discover the real process model of ship handling in the port.
- ✓ Setting main attributes significantly affecting this process.
- ✓ Time and delay prediction.
- ✓ Improving ship scheduling accuracy.
- ✓ Using a descriptive process model to strengthen existing MAS (multi-agent systems) and operation management.
- ✓ Optimization of the ship-handling process by delays detection and causes of them.
- ✓ Improving our understanding of the nautical process by examining of the descriptive process model.

2.3. Bottleneck analysis

Bottleneck detection is the key to improving production efficiency and stability in order to increase capacity [49]. The prerequisite for improving the bottleneck is to find the bottleneck in the first place – bottleneck detection. Hence, before searching for the bottleneck, it is important to first define clearly what a bottleneck is. Some common notions of denoting a bottleneck are as follows [50]:

- Bottlenecks are processes that influence the throughput of the entire system. The larger the influence, the more significant the bottleneck.
- The bottleneck is defined as a stage in a system that has the largest effect on slowing down or stopping the entire system.
- A bottleneck in the process is often defined as the operations/resource whose production rate in isolation is the smallest among all the operations in the system. In any case, the influence of the process on the overall system performance depends heavily on the speed of the process.

The bottleneck in the process can limit working hours, information flow, materials, and products. Thus, the biggest bottleneck in the process trace significantly impacts the capacity and throughput of the whole system. Comparing bottleneck occurrence through the process to the road traffic map, we can state that the capacities of the trace before and after the bottleneck are not important, the throughput of the process will be defined by the capacity of the bottleneck itself (see Figure 2.22).

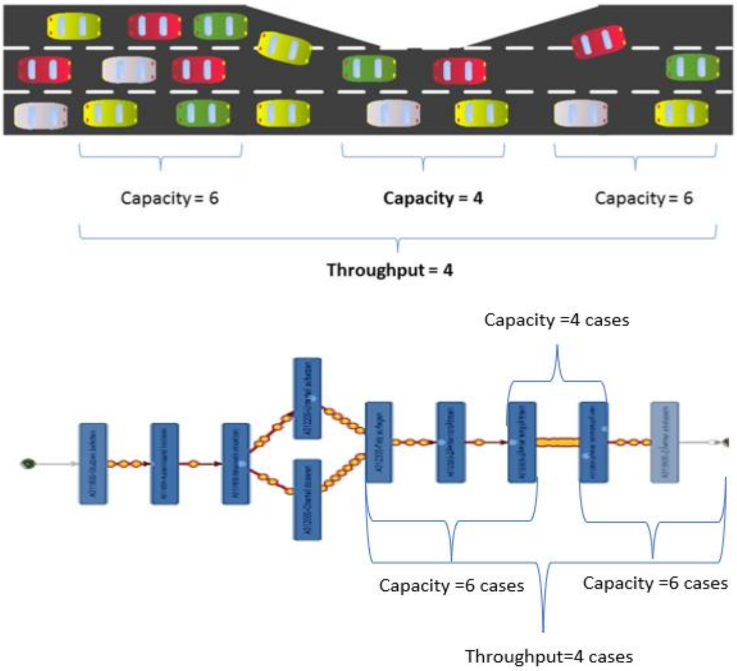


Figure 2.22: Comparing traffic bottlenecks to process bottlenecks in terms of throughput. The names of operations in the process model are not important since the goal is to show the bottleneck.

In terms of the origin, we can differ two kinds of bottlenecks in a process – *long-term* and *short-term* bottlenecks[18].

Long-term bottlenecks are those that keep on occurring in a company for a long time and have not been touched. It can include the postponements that usually occur in financial auditing,

[18] https://www.latestquality.com/bottlenecks-in-a-process/

monthly payroll handling, or inappropriate leave management of the organization's personnel. Exploring long-term bottlenecks leads to improving the processes and organizational workflows. However, processing long-term bottlenecks takes more time and effort because it requires reconsidering and restructuring the usual way of work.

On the other hand, *short-term* bottlenecks are those that take place all of a sudden owing to inadequate planning and automation deficiency. It can include an official person unexpectedly going on leave, inadequate knowledge about a venture or a missing file. This kind is more temporal and doesn't significantly impact the whole process parameters. Also, once it is detected, the issue can be fixed promptly.

Nowadays, automated lines and scanners can help to detect bottlenecks forthwith. Otherwise, in the case of not known and hidden processes, bottleneck detection is not so effortless procedure. Creating a process model for all processes in an organization and pointing out their time can be very useful in the identification of bottlenecks. There is clearly a bottleneck when the duration is time-consuming as compared to the estimated time. Timestamps of the event log can be used to track and estimate whether a bottleneck has occurred during the process. Having the process model as a basis helps to monitor, detect, and predict weak points of the process as well as to suggest possible reasons or roots. Once, a system is defined by the weakest place, bottleneck detection becomes a crucial part of the analysis.

2.3.1 Theory of constraints and Lean management

In my work, I am using two insights related closely to the process mining approach and bottleneck detection – the theory of constraints and Lean management.

The *theory of constraints* (TOC) is a management paradigm that views any manageable system as being limited in achieving more of its goals by a very small number of constraints. There is always at least one constraint, and TOC uses a focusing process to identify the constraint and restructure the rest of the organization around it. TOC adopts the common idiom *a chain is no stronger than its weakest link*. That means that organizations and processes are vulnerable because the weakest resource or part can always damage or break them, or at least adversely affect the outcome [51].

A *constraint* is anything that prevents the system from achieving its goal. There are many ways that constraints can show up, but a core principle within TOC is that there are not tens or hundreds of constraints. There is at least one, but at most only a few in any given system. Types of (considering the system as internal) constraints:

- *Equipment*. The way equipment is currently used limits the ability of the system to produce more salable goods/services.
- *People*. Lack of skilled people limits the system. Mental models held by people can cause behaviour that becomes a constraint.
- *Policy*. A written or unwritten policy prevents the system from making more.

Thus, according to the TOC the detection of the biggest and the most serious bottleneck of the process is enough to improve all other parameters. I will use this theory for one of the proposed methods for the bottleneck detection. However, it has limitations and can be applied not to complex processes.

Lean is a continuous improvement philosophy focused on eliminating waste while maximizing customer value. Lean management is the holistic approach of lean methods, the strategic implementation and the consideration and integration of the cultural level [52]. So, it can be described as a balanced use of resources, systems and tasks that brings to the whole process the best KPI (key performance indicators) like cost, time, risk.

One of the tools of lean management is *DMAIC* – a *data-driven improvement cycle* used for improving, optimizing and stabilizing business processes and designs. It is a five-phase method – *Define, Measure, Analyze, Improve and Control* – for improving existing process problems with unknown causes.

Since, the concept of PM is responding to both concepts, DMAIC is presented for process mining in the Figure 2.23.

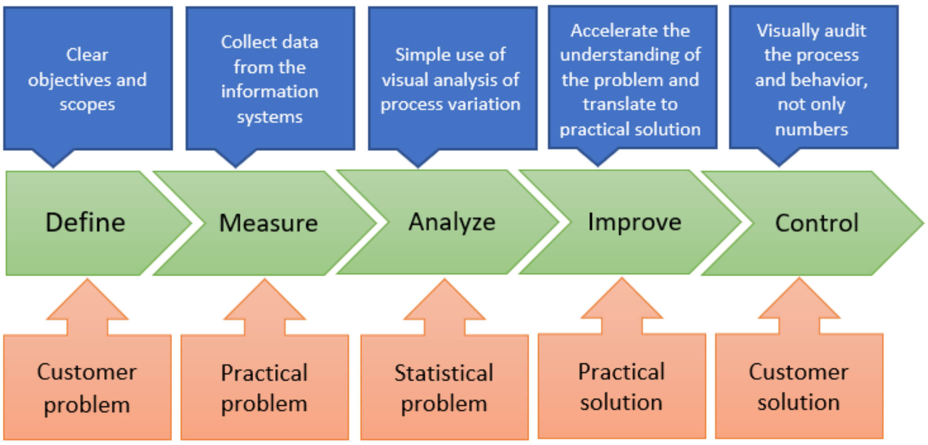

Clear objectives and scopes
Collect data from the information systems
Simple use of visual analysis of process variation
Accelerate the understanding of the problem and translate to practical solution
Visually audit the process and behavior, not only numbers
Define
Measure
Analyze
Improve
Control
Customer problem
Practical problem
Statistical problem
Practical solution
Customer solution

Figure 2.23: DMAIC in terms of process mining use [70]

Another comprehensive problem-solving instrument of Lean management is the *5 Whys technique*. After the bottleneck was detected, this technique helps to track roots and reasons, asking domain experts correct questions.

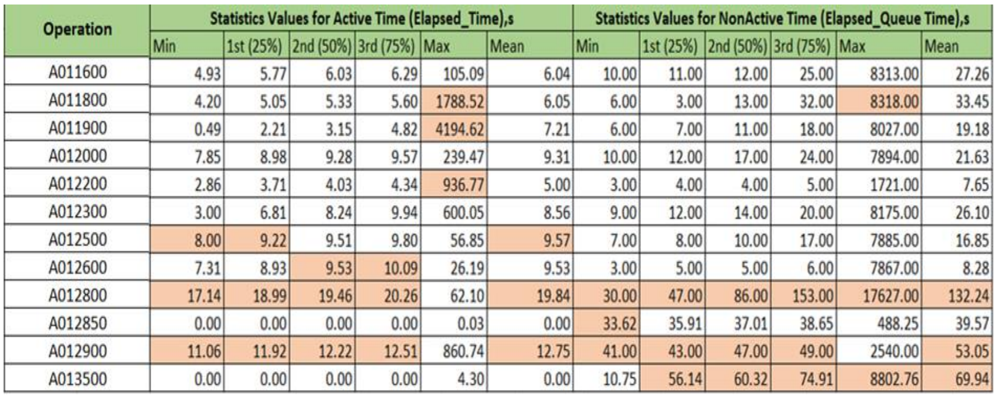
Both philosophies focus on improvement and advocate techniques to control the process flow. Both have demonstrated dramatic results of implementations—profitability skyrockets, inventories and lead times are slashed, and operations are drastically simplified. At the same time, both movements recognize that in order to achieve and sustain such improvement trends, new tools and methods should be involved. As a result, TOC and Lean movements have expanded their scope to encompass principles and practices of the entire system to enable continuous systemwide improvement [53]. Recently, PM became one of the tools for Lean management since its main goal is to reduce *waste* in the processes and to make them transparent.

2.3.2 Common methods of bottleneck analysis

In this section, I consider widely used bottleneck detection methods, which ideas became the core for proposed algorithms in the Chapter 3.

Statistical bottleneck detection

Statistical metrics are widely used and do not require sophisticated calculations. To detect the most common roots of bottlenecks, statistical characteristics like min/max values, data quartiles (three values that split sorted data into four parts, each with an equal number of observations – 25, 50, 75% of data set size), and mean (average) values can be calculated. Working with statistics, one needs to inspect the structure of data first. In case of abnormal distribution, there is a recommendation to take a median (or 2nd quartile) since this parameter is not impacted so much by outliers as a mean value [54]. The time can be considered both for the activity/operation (active, *elapsed* time) itself and for delays between operations in queues (non-active, waiting, *elapsed in queue* time). Reasons for bottlenecks can have different natures – an absence of resources, breakdown, unique cases, etc. Also, the longest time for the operation can indicate its complexity and specificity and should not be recognized as a bottleneck without confirmation from the process owner. According to the historical data, results can be represented in the next view (see Figure 2.24) and the longest active and waiting time can be highlighted in the column for each position of the process.



Operation	Statistics Values for Active Time (Elapsed_Time),s						Statistics Values for NonActive Time (Elapsed_Queue Time),s					
	Min	1st (25%)	2nd (50%)	3rd (75%)	Max	Mean	Min	1st (25%)	2nd (50%)	3rd (75%)	Max	Mean
A011600	4.93	5.77	6.03	6.29	105.09	6.04	10.00	11.00	12.00	25.00	8313.00	27.26
A011800	4.20	5.05	5.33	5.60	1788.52	6.05	6.00	3.00	13.00	32.00	8318.00	33.45
A011900	0.49	2.21	3.15	4.82	4194.62	7.21	6.00	7.00	11.00	18.00	8027.00	19.18
A012000	7.85	8.98	9.28	9.57	239.47	9.31	10.00	12.00	17.00	24.00	7894.00	21.63
A012200	2.86	3.71	4.03	4.34	936.77	5.00	3.00	4.00	4.00	5.00	1721.00	7.65
A012300	3.00	6.81	8.24	9.94	600.05	8.56	9.00	12.00	14.00	20.00	8175.00	26.10
A012500	8.00	9.22	9.51	9.80	56.85	9.57	7.00	8.00	10.00	17.00	7885.00	16.85
A012600	7.31	8.93	9.53	10.09	26.19	9.53	3.00	5.00	5.00	6.00	7867.00	8.28
A012800	17.14	18.99	19.46	20.26	62.10	19.84	30.00	47.00	86.00	153.00	17627.00	132.24
A012850	0.00	0.00	0.00	0.00	0.03	0.00	33.62	35.91	37.01	38.65	488.25	39.57
A012900	11.06	11.92	12.22	12.51	860.74	12.75	41.00	43.00	47.00	49.00	2540.00	53.05
A013500	0.00	0.00	0.00	0.00	4.30	0.00	10.75	56.14	60.32	74.91	8802.76	69.94

Figure 2.24: The sample of statistical characteristics for each operation [68]

The disadvantage of the statistical method. Unfortunately, statistics is a data-centric approach and does not consider objects (processes) and their relations. Besides the process flow perspective, it also does not include important process parameters such as concept drifts (process changes), process variants, loops, and concurrency. This method is not able to follow the flexibility of logistics supply chains. Mathematical models are made for common cases and are useless for the specific minority of processes. Also, the method strongly depends on historical information, more than on expert knowledge. In addition, outliers in data impact result to a great extent and should be processed separately. However, historical data analysis can be beneficial for process model enhancement.

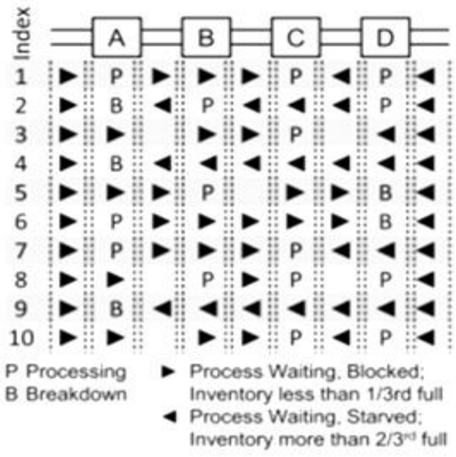
Bottleneck walk method

This approach is extensively described in [55]. The *bottleneck walk* is based on observations of different process and resource states. These data are collected during a walk along the flow line. The gathered data are evaluated in a systematic process. The result of these two steps is a ranking of bottleneck sets that limit the output of the flow line during the period observed (see Figure 2.25).

When observing a process, it cannot be determined by one observation alone if the process is the bottleneck. If the process is working, it may or may not be the bottleneck. If the process has an ongoing breakdown, it may or may not be the bottleneck. If the operator is absent, it may or may not be the bottleneck. However, it can be clearly stated when it is not the bottleneck. Whenever the process is waiting, it cannot be the bottleneck, since the process is waiting on another process. The process could work more but is slowed down by the bottleneck. Furthermore, from this observation of a waiting process, it can be determined in which direction the bottleneck needs to be searched next. There are three possible system states and the conclusion about the bottleneck:

- *May be the bottleneck working*; breakdown; setup; maintenance; scheduled break, etc.
- *Starved bottleneck is upstream* – the inventory is empty or rather empty.
- *Blocked bottleneck is downstream* – the inventory is full or rather full.

During the bottleneck walk, the observer walks along the line, writing down the inventory levels and process states in one line of the data sheet each round. For practical purposes, the process states are abbreviated with *W* for waiting, *P* for processing, *B* for breakdown, and so on. Subsequently, for every buffer or process where the direction of the bottleneck can be determined, an arrow is drawn on the datasheet in the direction of the bottleneck. The bottleneck then must be between the arrows pointing toward each other [55].



Index		A		B		C		D	
1	▶	P	▶	▶	▶	P	◀	P	◀
2	▶	B	◀	P	◀	◀	◀	P	◀
3	▶	▶		▶	▶	P		◀	◀
4	▶	B	◀	◀	◀	◀	◀	◀	◀
5	▶	▶	▶	P		▶	▶	B	◀
6	▶	P	▶	▶	▶	▶	▶	B	◀
7	▶	P	▶	▶	▶	P	◀	◀	◀
8	▶	▶		P	▶	P		P	◀
9	▶	B	◀	◀	◀	◀	◀	◀	◀
10	▶	▶		▶	▶	P	◀	P	◀

P Processing
B Breakdown
▶ Process Waiting, Blocked; Inventory less than 1/3rd full
◀ Process Waiting, Starved; Inventory more than 2/3rd full

Figure 2.25: Sample of the bottleneck detection walk method for ten observations [55]

The method is suited for practical use by supervisors and operators. The direct observation of the bottleneck also gives additional information about the underlying causes of the bottlenecks, simplifying the improvement of the system capacity.

The disadvantage of the method. The method is functional, and it is applied to a *flattened* process. Once, the process is getting complex, the bottleneck walk is confusing. Moreover, it finds queues and weak points manually by counting cases/units before and after the operation. In this case, another issue occurs, when a manual resource takes out the production line/process trace or adds new items in a certain part of the production line to rework. Therefore, items just emerge in the production log through the process, and it is hard to follow them. Also, the method is meant to be working with a direct flow process, not processing conjunctions and parallel flows, or loops. The detection can be automated using sensors, production logs and information about the buffer of items before and after the operation. However, the absence of knowledge of the descriptive model will require manual work when every little change in the process occurs and measurements should be recalculated from a scratch.

Bottleneck analysis with PM and animation

Once, the adjusted descriptive process model was created, statistical metrics can be added to it and give an integrated perspective. In order to get execution times for all steps of the production process, both start and completion timestamps in the data set should be identified.

Performance analysis gives the opportunity to see the mean, median, min/max time, and frequency of cases through the process. Filtering by process variants, start/end timestamps, case duration, etc. helps to focus on specific cases with a detailed view, and, thus, drilling down in the process. When one wants to intuitively spot and highlight bottlenecks in the process, process animation can be applied. Figure 2.26 presents two different views on the same process – animation of process flow and performance analysis – on the DFG notation. Each rectangle refers to the activity, yellow dots are cases, arches represent connections between operations and start and finish are defining the scope of the chosen process accordingly. A lot of cases, which accumulate in certain paths, reveal possible weak spots in the process. Real-time monitoring shows when many cases pile up on a certain arc and are causing congestion. Process animation[19] groups these cases into larger *bubbles* (see Figure 2.26 (left))

[19] Animation from Disco tool (www.fluxicon.com)

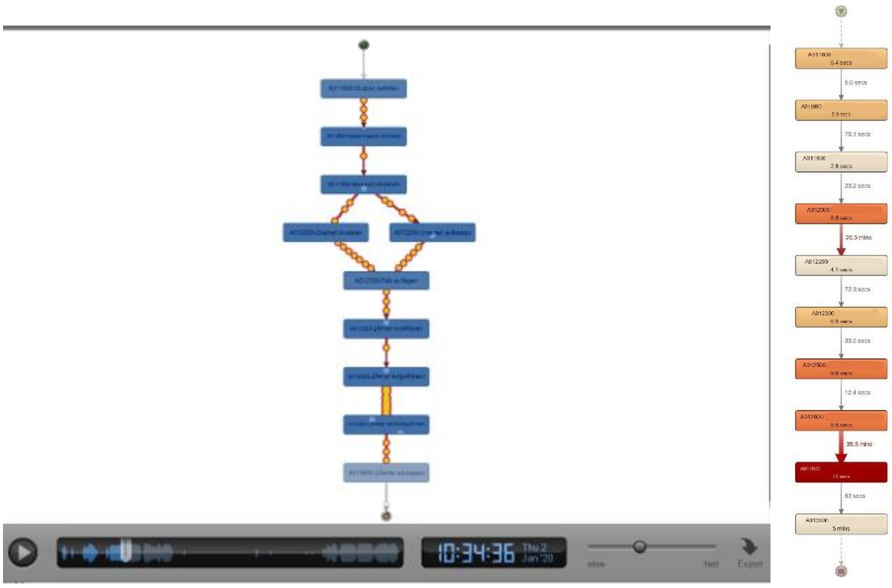


Figure 2.26: Process animation to detect bottlenecks in real-time (Disco) – on the left side, Process map with Performance perspective – on the right side. The names of operations are not important, since the figure shows the sample of different perspectives [68]

Another process map view, which is shown on the Figure 2.26 (right), presents performance characteristics for each operation and the paths between them. Red colors detect the longest times in comparison with other parts. The process model enriched with information about times makes the detection more transparent and comprehensible. Figure 2.26 presents an abstract view of the process bottlenecks and the names of operations do not play an important role here. The performance view will be discussed more specifically in the Chapter 3 for a proposed bottleneck detection method. Both schemes were obtained from real data and all operation names were hidden in compliance with FACT statement – the confidential policy since order and names of operations can be easily recognizable by competitors.

The disadvantage of the approach. Although the process model combined with statistic metrics gives a better understanding of the process and possible bottlenecks, the detection in this method is relative. All activities and paths are compared with each other based on historical data. Therefore, if one operation takes 25 min (and this time is regular for the operation), but another one takes 5s, then the first operation will be detected as a weak place in the process. This situation can be omitted by time processing and considering each operation as instant. In this case, the waiting time between two operations will combine the duration time of the operation (active time) and time in the queue (waiting time). However, in this case, it will be problematic to differ bottleneck roots – breakdown of a resource or delays on the path between activities. Also, finding the root of the bottleneck requires additional analysis and cannot identify the direction of the issue as well as guess possible reasons.

Summary of the section

The bottleneck detection is a crucial part of the continuous improvement of processes. The bottleneck in the process can limit working hours, information flow, materials, and products. Thus, the biggest bottleneck in the process trace significantly impacts the capacity and throughput of the whole system. Two concepts – theory of constraints and lean management – focus on detection and reducing possible bottlenecks/weak points by reconsidering process management. It leads not just to improving KPIs, but also to knowing risks and threat points in the system.

The common methods of bottleneck detection – Statistical, Bottleneck walk and Bottleneck analysis with PM methods – were discussed in the current section. In spite of their advantages, the flexible and appropriate method of bottleneck detection stays relevant. I described the drawbacks of each method. Most of the time present bottleneck detection methods are data-centric, not aligned with the process model, and do not consider loops and concurrencies. However, this situation can be improved by mixing the advantages of each method. Chapter 3 will present how these approaches can be improved with the help of the process model. There will be two bottleneck detection methods proposed. One is effective in terms of the theory of constraints. The second one is more detailed and helps with not just bottleneck detection but also guessing the reasons.

2.4 Predictive models

Prediction of KPI parameters plays an important role in the planning and helps to reduce *wastes* in warehousing, and processes, as well as to improve the quality of service for customers. *Predictive modelling* is a statistical approach that explores data patterns to predict future events or outcomes. These outcomes can be used then for recommendation and decision support systems or process maps. Predictive models solve various groups of tasks: clustering, classification, forecasting, threat predictions, outliers /anomalies identification, and time series models. To obtain better results some of them can be combined. Also, each group of tasks can be transformed into another one according to the initial request. It is important to note that algorithms are taught on historical data and input affects the parameters of the predictive model significantly. Thus, data cleaning and selecting features takes usually 70% of the analysis.

In this part, I describe some widely used predictive models and methods for the evaluation of the quality (validation). The problematic of every method is defined at the end of each section. All figures from this chapter are presenting an application of the real use case from Chapter 4 and are used to show the main parameters of the models. These parameters will be used later in the application part. Evaluation metrics with comparison will be described in the last chapter.

2.4.1 Time Series analysis

Most of the time, predictive models are needed to estimate times: flow time, the end time of the process, suspend time, delay time, etc. As the name of the analysis suggests, the time series model comprises the sequence of data points using time as an input. Using historical data – timeline – one can create long- and short-term predictions. Without knowing the process model and its changes, it is recommended to create short-term prediction, since for the long-term it gives worse accuracy results and is quite sensitive to any possible changes. For instance, in terms of epidemy or economic crises, the long-term model will not be useful.

So, *time series* (TS) is an ordered sequence of values of a variable at equally spaced time intervals [56]. Examples of time series can be found in a variety of fields ranging from medicine to economics. The investigation of the time series makes it possible to control the process that generates time series, clarify the mechanism, situated in the basis of the process, clear a number of outliers, and also make predictions for the future based on knowledge of the past. The main characteristics of TS, which make it different from a regular regression problem, are time-dependence and seasonality trends (i.e., variations specific to a particular time frame). The basic assumption underlying the analysis of time series is that the factors, affecting the object, in the present and in the past, will affect it in the future. The goal of TS analysis is to find dependencies or patterns and express them in the form of mathematical expressions. To achieve this, various mathematical models have been developed, designed to study the TS components. Thus, it it makes possible to forecast future values according to the historical data that can help in the development of plans and development strategy of the organization. I consider the most important components of TS (trend, seasonal variation, cyclic changes, irregular factors) [57] to build the most suitable model in terms of our time series.

The models which are described are exponential smoothing (single, double and triple), and two more advanced models – auto-regressive moving average (ARMA) and ARIMA with seasonal component (SARIMA). The choice of a suitable model depends both on the distribution of the data and on the useful information that it will bring. Experiments for every model are carried out with different values of the model parameters to find the best fitting ones [71].

Time series components

Each time-series method requires the understanding and description of the timeline. Prior to applying algorithms, it is recommended to conduct the next steps:

- *Define* the time period and provide an explorative analysis of this timeline to see the behavior of the system with different frequencies.
- *Determine* the normal distribution by applying specific statistical tests.
- *Identify* trend and stationarity.
- *Decompose* the time series to see cyclic changes and irregular factors.

Before choosing the method of prediction, it is necessary to provide an explorative analysis of the timeline. Often, this part can impact the type of time series analysis as well as coefficients. First, time series should be visualized with different frequencies (days, weeks, months) (see Figure 2.27). In this work the language Python (in particular library statsmodels) was used for time series exploration and visualizations. To get a general idea of the data, time series should be visualized (see Figure 2.27).

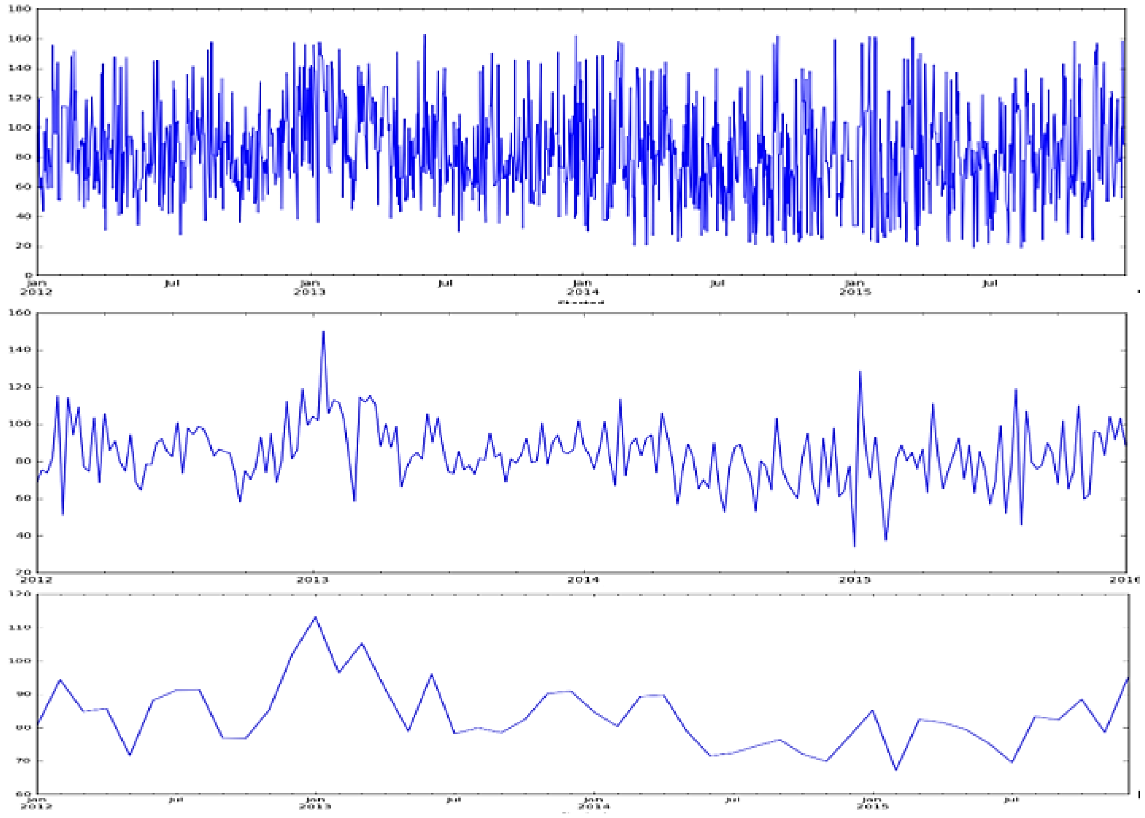
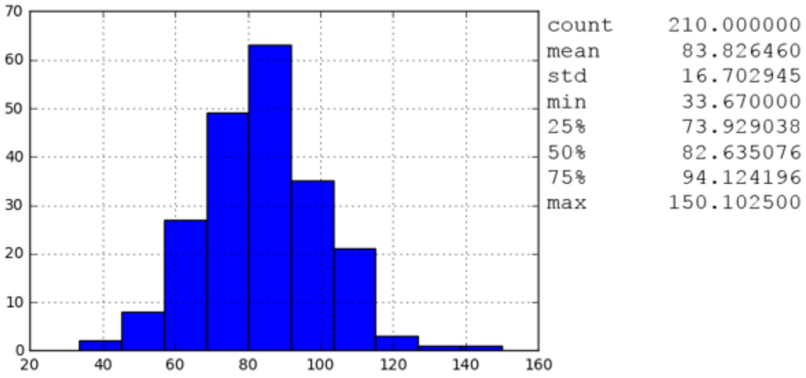


Figure 2.27: Ship handling duration time plot with different frequencies (from top to bottom – *TSd(days), TSw(weeks), TSm(months)*) [71]

Figure 2.27 shows two peaks for TSw (January 2013 and 2015) and one peak for *TSm* (January 2015), but this figure is not specific enough to discuss whether the data has a distinct trend or seasonality.

On the other hand, time series should have a normal distribution. This goal can be achieved by filtering and preprocessing inputs. Ideally, TSw indicators and histogram should look as it is depicted in Figure 2.28. So, the homogeneous TS with a relatively small dispersion is a suitable object for the models. It is also evidenced by the coefficient of variation (the ratio of standard deviation to average value) that equals 0,199. Moreover, to determine the normal distribution the Jarque-Bera test can be conducted, and if the p-value is higher than 0,05 then the null hypothesis of the normality is true (in described case p-value=0,11), what means the series has a normal distribution [58].



```
count    210.000000
mean      83.826460
std       16.702945
min       33.670000
25%       73.929038
50%       82.635076
75%       94.124196
max      150.102500
```

Figure 2.28: TSw bar chat and its main indicators [71]

After the explorative analysis and normal distribution detection, another four main components which can impact the predictive models for time series are:

- trend (long-term change in the mean level),
- seasonal variation (the periodic fluctuation in the series within each year),
- cyclic changes (variation across a fixed period due to some physical cause),
- irregular factors.

These parameters can be defined by the time series decomposition method shown in Figure 2.29.

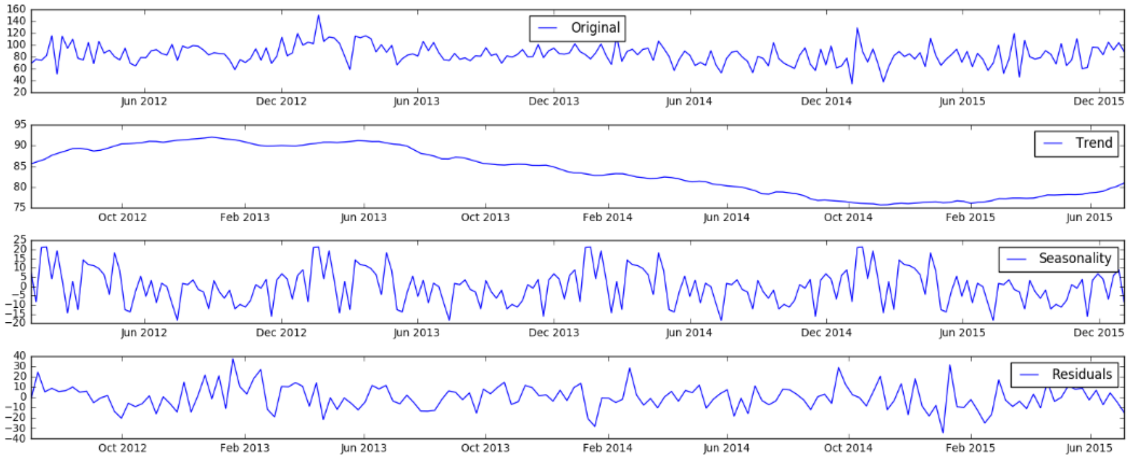
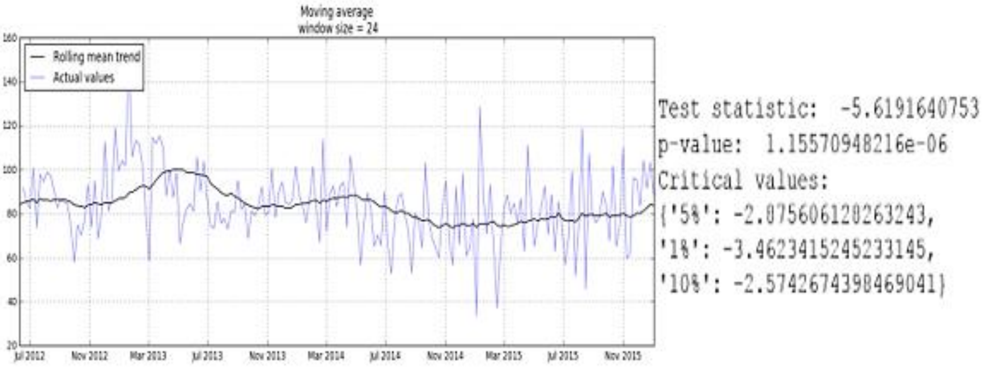


Figure 2.29: Decomposition of TSw [71]

On the Figure 2.29 there is seasonal variation, which can be caused by changes in the process. The explored time series is stochastic and discrete, the trend is not significant.

Nevertheless, in the first place, time series stationarity should be checked to find out if it is suitable for statistical techniques. Besides, the techniques related to stationary series are more mature and easier to implement as compared to non-stationary series. TS is said to be stationary if its statistical properties such as mean, variance and autocorrelation structure remain constant over time. There are two basic methods for checking the stationarity: plotting rolling statistics (moving average) and Dickey-Fuller test [59].



```
Test statistic:  -5.6191640753
p-value:  1.15570948216e-06
Critical values:
{'5%': -2.875606128263243,
'1%': -3.462341524523314 5,
'10%': -2.5742674398469041}
```

Figure 2.30: Moving average and Dickey-Fuller test [71]

According to Figure 2.30, the mean value appears to be varying with time slightly and the test statistic is much smaller than 1% critical values so we can say with 99% confidence that this is a stationary series.

In order to ensure TSw stationarity in another way, *ACF* (Autocorrelation function that presents a correlation between different lag functions) and *PACF* (Partial autocorrelation function) are shown in the figure below. For stationary time series, both functions should decrease to zero as the lag increases, as evidenced be the presented functions.

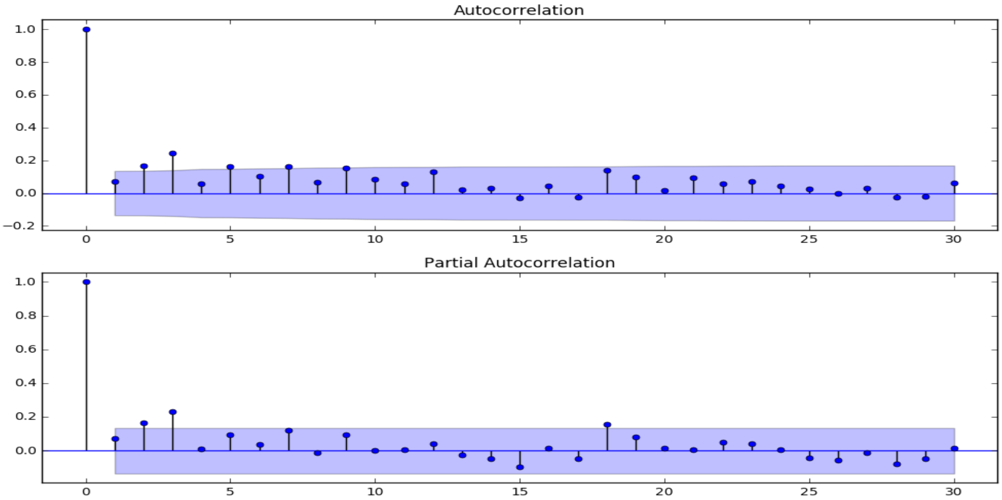


Figure 2.31: ACF and PACF of TSw [71]

Almost none of practical TS is stationary and, probably, data preprocessing makes it possible. If even after data transformation TS stays non-stationary, there are three main approaches to change it:

- detrending (remove the trend),
- differencing (model the difference of the TS),
- seasonality.

The differencing is also called the integration part in ARIMA model and seasonality is SARIMA.

Exponential smoothing

Exponential smoothing of time series data assigns exponentially decreasing weights for newest to oldest observations. In other words, the older the data, the less priority (*weight*) the data is given; newer data is seen as more relevant and is assigned more weight. *Smoothing parameters* (smoothing constants) — usually denoted by α — determine the weights for observations.

- *Simple exponential smoothing*. It can be represented as the following formula:

$$S_t = \alpha y_{t-1} + (1-\alpha)S_{t-1} \tag{2.4}$$

where

S_t – overall smoothing;

y_t – observations;

α – the smoothing constant, a value from 0 to 1 (when α is close to zero, smoothing happens more slowly);

t – time period.

This method uses a weighted moving average with exponentially decreasing weights. It is not recommended to apply it to TS with trend and seasonality [60].

- *Double exponential smoothing*. Here are the two equations, which define the method:

$$S_t = \alpha y_{t-1} + (1-\alpha)(S_{t-1} + b_{t-1} \tag{2.5}$$

$$b_t = \beta(S_t - S_{t-1}) + (1-\beta)b_{t-1} \tag{2.6}$$

where

b_t – trend smoothing;

β is a constant that is chosen with reference to α. Like α it can be chosen through the Levenberg–Marquardt algorithm.

This method is used for analyzing time series that show a trend.

- *Triple exponential smoothing*. This method can be defined as follows:

$$I_t = \gamma * \frac{y_t}{S_t} + (1-\beta)I_{t-L} \tag{2.7}$$

$$F_{t+m} = (S_t + mb_t) * I_{t-L+m} \tag{2.8}$$

where

I_t – seasonal smoothing;

F_t – the forecast at *m* periods ahead.

The resulting set of equations is called the *Holt-Winters* method after the names of the inventors. The method is applied for data that shows a trend and seasonality. All formulas are taken from the source [60].

The disadvantage of Holt-Winters methods. For this method, the main problem is to determine the best coefficients. Finding the appropriate coefficients is calculated based on the error value. It requires computer capacity is and very sensitive to process changes. Also, the resulted model should not be overfitting or generating.

ARMA model

ARMA stands for Auto-Regressive Moving Averages and provides a description of the stationary stochastic process in terms of two polynomials, one for the autoregression and the second for the moving average. This model is used more in practice than simple *AR* and *MA* models. The *AR* part involves regressing the variable on its own lagged (i.e., past) values. The *MA* part involves modelling the error term as a linear combination of error terms occurring contemporaneously and at various times in the past [56].

SARIMA

SARIMA is a seasonal autoregressive integrated moving average model. It is an extension of ARMA model with a seasonal component. The model can be defined as follows:

$$ARIMA(p,d,q)(P,D,Q)_S \tag{2.9}$$

where

p – order of the *AR* process (*AR* terms are just lags of dependent variable);

q – order of the *MA* process (*MA* terms are lagged forecast errors in prediction equation);

d – order of differencing (in our case, d=0, because TS_w is stationary)

P, D, Q – order of seasonal *AR*, differencing, and *MA* respectively.

Often time series possess a seasonal component that repeats every s observations. For monthly observations s = 12 (12 in 1 year), for quarterly observations s = 4 (4 in 1 year).

Disadvantages of time series models. Time series analysis is not multicriteria and could not be used for global planning as well as for long-term prediction, since every single process case is considered a black box with just one parameter – the time duration. In this type of prediction, many significant features of the process like case attributes (size, type, cargo, etc.) are staying without needed attention. Context awareness also cannot be applied and enrich the results. The forecast can be called short-term because it is based on historical data and any changes in the organization of processes or in the infrastructure will lead to a new analysis of the time series. The process of the prediction is not transparent and the reasons for predicted values should be always discussed with the process owner.

2.4.2 Regression models

Having big data as a source opens new opportunities for predictive models and requires them to be multivariate. Exploration of dependencies between different features brings new aspects to the creation of decision rules and recommendations. *Regression models* offer a set of techniques to predict and understand dependencies among datasets.

In statistical modelling, *regression analysis*[20] is a set of statistical processes for estimating the relationships between a dependent and one or more independent variables. To define a regression model, we can use the formula:

$$Y_i = f(X_i, \beta) + e_i \tag{2.10}$$

where

β is the unknown parameter,

X_i – independent variables, which are observed in data (*i* is defined by a row of data)- factors, which can impact the dependent variables,

Y_i – dependent variables – target variable (the parameter which is planned to be explained or predicted),

e_i – error terms.

The regression analysis is a quite powerful tool in the combination with other statistical and machine learning techniques because it provides information about factors, which affect the system.

[20] en.wikipedia.org/wiki/Regression_analysis

Regression trees

A *decision tree* creates regression (or classification models) in the form of a tree structure. It splits down a dataset into smaller subsets till the point when an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.

According to professor Saed Sayad[21], the main algorithm for building decision trees is named *ID3* by J. R. Quinlan which conducted a top-down, greedy search through the set of possible branches with no backtracking. The *ID3* algorithm can be used to build a decision tree for regression with standard deviation reduction, which is calculated for each node. It is important to note that instances in the decision tree should be homogenous (with a similar types of values). Developing a decision tree is about finding the attribute that returns the highest standard deviation reduction. In reality, certain thresholds can be defined for nodes. The sample of the decision tree regression model is presented below.

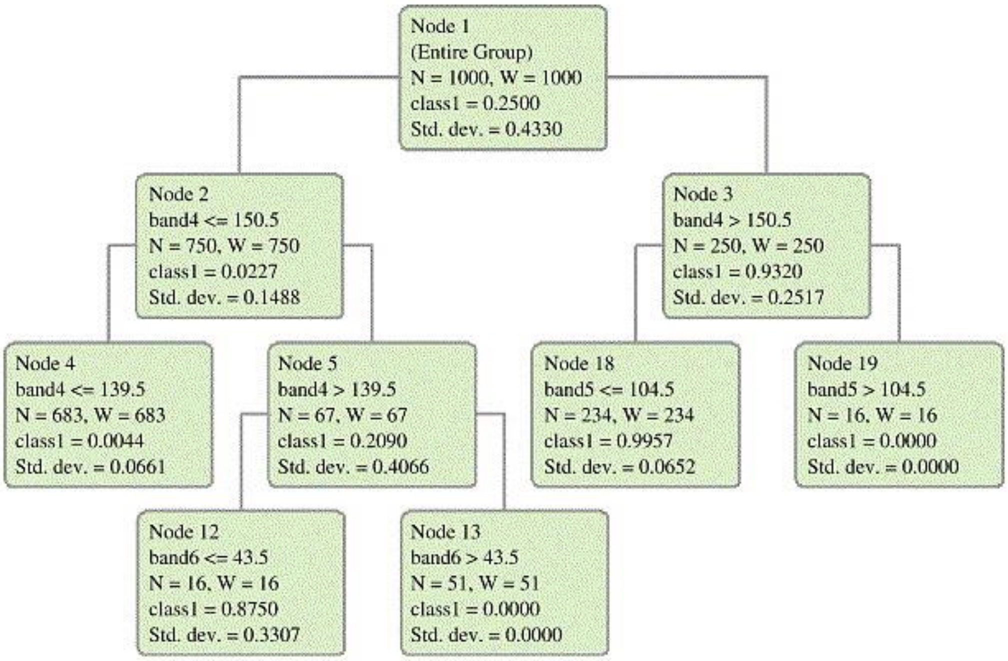


Figure 2.32: The sample of the decision tree regression model [72]

The advantage of the regression tree method is the possibility to explore a big set of features and their dependencies. Also, we can create decision rules and recommendations using context awareness, what is not possible with time-series analysis.

Disadvantage of regression models. According to FACT issue, described in the process mining section, even if the correlation between two variables equals 1, it does not mean their strong dependency. In fact, there can exist additional hidden variables which were not considered, and it impacts two analyzed ones. One should be careful with statistical numbers for nodes and have an understanding of the process. Also, the regression model is not connected to the process model. Similar to previous methods, it is data-centric and is not able to bring process aspects into the analysis. However, it can be fixed with context awareness. Thus, the method can be exploited as a part of another model to reinforce the results of the prediction.

[21] www.saedsayad.com/data_mining_map.htm

2.4.3 Transition systems

All models which were described before are not process-oriented and are more data-centric. They can obtain behavior of the process, but this disconnection requires additional research to be able to create a process map further. One of the possible process-centric predictive methods is a transition system.

A *transition system* is a triplet (*S*, *E*, *T*), where *S* is the state space (i.e., possible states of the process), *E* is the set of events labels (i.e. transition labels), and is the transition relation describing how the system can move from one state to another [61]. A transition $(s_1, e, s_2) \in T$ describes that the process can move from state s_1 *to* s_2 by an event labelled *e*. A transition has some initial state and a set of final states. The set of behaviors possible according to a transition system is given by all traces from the initial state to some final states. The naive algorithm for constructing a TS is straightforward: for every trace σ a new state should be created (if it does not exist yet). The algorithm can be studied in more detail in the paper [61].

For this study, a set of activities abstraction was chosen as the most suitable. It means that the mechanism optionally removes the order or frequency from the resulting trace. Only the fact that it occurs makes sense.

In order to use the transition system for making predictions, it is necessary to annotate it first. The system learns from the information collected for earlier process instances that visited the same state. This way states are annotated with a set of measurements that are used as a basis for predictions. The set contains:

- *remaining time* (the average time in a current state until completion of the case),
- *elapsed time* (the average time to reach a particular state from the initial state),
- *sojourn time* (the average duration of a current state) (see Figure 2.33).

It is important to note that in the event log we usually see activities and no states, therefore, state information is deduced from the activities executed before and after a given state.

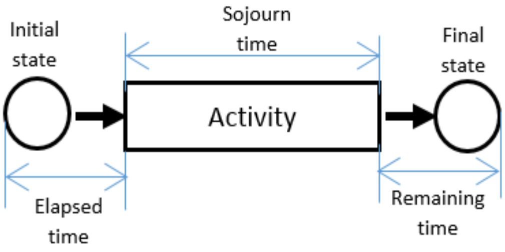


Figure 2.33: Time measurements for transition system

These measurements can be used for calculating additional time parameters for every state: average time, standard deviation, minimum and maximum remaining time until completion. The proposed bottleneck method in the third part relies on these metrics.

Eventually, the transition system can be obtained using the set of abstractions based on all activities. With the plugin FSM analyzer in PROM, the transition system might be also reinforced with information about times for every operation. The goal is to predict, at any point in time, the remaining handling time of a case. The sample of certain parts of the process – initial and finish states – is presented below.

A transition system can provide information about remaining time in every state and, at the same time, help with the understanding of control flow, delays detection and reasons for them. Thus, the process owner will be aware of exactly how every operation affects the whole process and will be able to make arrangements. Such analysis can lead to the creation of decision-making rules, and the determination of measures to improve the quantitative and qualitative dimensions of the process.

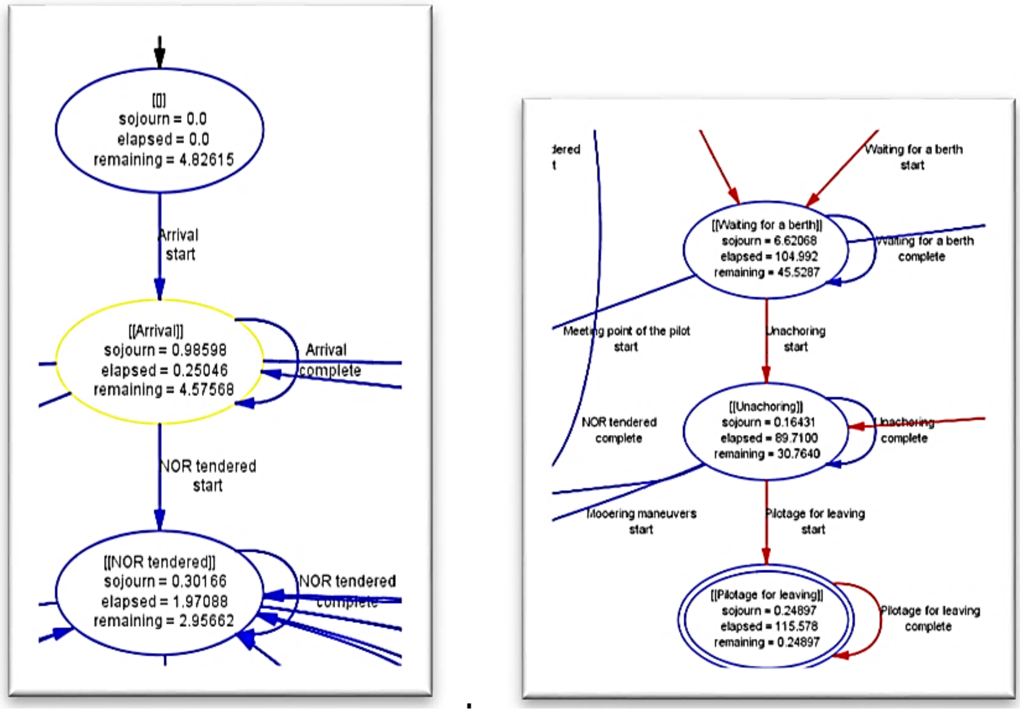


Figure 2.34: Annotated transition system for two states (on the left-initial state, on the right-final state) [64]

Disadvantages of transition system models. In contrast to statistical methods, the transition system is able to indicate deviations and causes for delays, as well as support multicriteria analysis for process instances. It is also important to note, that results can be improved substantially by increasing input data quality and using other abstractions of transition systems like sequence and multisets. However, the transition system predictive method has some disadvantages. The main problem is heightened sensitivity to input data. Duplicates, noise, incompleteness, and uncertainty in the event log can lead to meaningless results. Before applying the algorithm, the event log should be preprocessed to avoid potential mistakes. Furthermore, the final representation of the transition system has problems with complex control-flow constructs (see Appendix D). Overloaded with many different parameters, the process model loses transparency, usability and readability. In order to tackle this problem, a two-step approach (combination transition system method with other process mining techniques) [62].

2.4.4 Evaluating methods

After developing the predictive model based on historical data, one should know its performing metrics to use it in future and to compare it to other models. To evaluate the results of predictive models data are usually divided into two groups: the training set and test set. The first one we need to build the model and the second is used to asset future performance. Most of the time, only a limited amount of data is available, so the unbiased estimate can be achieved by using cross-validation. This resampling method takes different portions of the data set to test and train the model. Since time series have ordered structure and dependencies, the cross-validation method here is not so considerable. Data can be separated manually taking last year/quarter/month (depending on the dataset and the target variable) as a test dataset. After splitting the data, test data can be compared to the results of the predictive model to analyze and find the best fitting one. As long as, many target variables are numerical, for model comparison next metrics can be useful:

- *AIC (Akaike Information Criterion)* – a function of both the fit, the sum of squared residuals, and the number of parameters.
- *MSE (Mean Square Error)* – the mean/average of the square of all of the errors,
- *RMSE (Root-mean-square deviation)* – the standard deviation of the prediction errors. The values between 0.2-0.5 can be evaluated as well-predicted results.
- *MAE (Mean absolute error)* – a measure of errors between paired observations expressing the same phenomenon.
- *MAPE (Mean absolute percentage error)* – is the sum of the individual absolute errors divided by the demand (each period separately). The values lower than 20% can be evaluated as well-predicted results.

In this thesis, the last four metrics will be used to evaluate the model in the Application part.

Summary of the section

Predictive modelling is a statistical approach that explores data patterns to predict future events or outcomes. These outcomes can be used then for recommendation and decision support systems or process maps.

In this part of the work, I have described three predictive model types: time series models, regression models and transition system models. All of them are widely used in different fields and each has its own benefits in comparison to others. To summaries all their strengths and weaknesses, the following table is created.

Table 2.3: The comparison of predictive models

Predictive model	Advantages	Disadvantages
Time series model	• Easy to apply to the different systems • A lot of samples and use-cases • Algorithms are understandable • Few parameters to estimate	• Ability to model only linear relationships • Low accuracy in the longer-term prediction • Statistical and domain knowledge needed • Sensitive to process changes • Not multicriteria, no connection to the process model
Regression model	• Can empower other models • Multicriteria (possibility to use context-awareness) • Model also non-linear relationships • Sustainable to changes	• Strong correlation between variables can be incorrect (hidden parameter) • Difficult interpretability • Requires a lot of calculations in case of a big amount of parameters • Data-centric, not related to the process model
Transition system model	• Process connected, shows connections in the process and delays • Multicriteria (possibility to use context-awareness)	• Difficult interpretability in case of complex *spaghetti* processes • Need to build the process model first

	• Sustainable to changes since considering different variants • Can be used for the process map	• Very sensitive to the input data and thresholds

In the Chapter 3, there will be a proposed two-step predictive model, which considers context awareness as an input parameter. The transition system impacted the development of the bottleneck detection method. All these three predictive models will be applied to the logistics process in the application part and evaluated using the metrics described above.

Summary of Chapter 2

This chapter covers recent research on the main parts of this thesis: Process Mining, Bottleneck detection analysis, Predictive models, and Domain knowledge. In the beginning, we defined the position of process mining in the data science field, considering also other fields which will be discussed in the following sections – statistics, machine learning and domain knowledge. Each section has a summary with conclusions about gaps in the research, challenges, weak points and advantages of described techniques with a hint for improvements. All this information will be used in the Chapter 3 and will help to understand the proposed methods.

Process modelling is necessary in order to display the entire process from beginning to end with various levels of detail and the ability to enrich the model with additional attributes in the future. The process mining approach has a lead role and differs from others by having flexible and powerful tools. It provides an opportunity to get a descriptive process model based on raw data from different sources like documentation, ERP systems, Intranet, etc. The descriptive type of model shows the current process and is made by algorithms based on raw data. The obtained model is used for optimization, improvement, and enhancement of the real process and can be the basis for recommendation or decision-making systems as well as for process maps. In this thesis, I focus more on Discover and Enhancement types of process mining. Once modelled behaviour and real behaviour are aligned it gives the opportunity to include other perspectives of process mining, what can be beneficial in terms of online prediction, automated process improvement, resource recommendation, etc. The current work aims mostly to control-flow, case, and time perspectives.

As a fast-growing area, process mining getting new challenges for academics and analysts. Solving issues brings up more perspectives and a deeper view of the process nature, which creates more challenges. The aim of the work is challenge 6 *Combining process mining with other types of analysis, considering FACT problematic and feature selection* – context awareness.

In turn, obtained process model brings follows benefits:

- opportunity to look at the real descriptive process model with *highways* in the process;
- performance analysis can show bottlenecks in the process and can help to define their causes according to the process model;
- process model can be used for audit (to compare normative and real models, to follow rules of standardization), for training new personnel;
- analysis of additional process characteristics can create decision rules, make predictions and recommendations;
- resource analysis defines overloaded resources. Can be used for the redistribution of work.

Moreover, there is a limited number of studies regarding process mining in logistics and only a few cases in logistics applications. Therefore, this research objective is to discuss the application of process mining, its advantages, as well as the challenge in the logistics domain. Maritime port systems represent a major challenge both from a theoretical and practical point of view and require the development of innovative tools for the analysis and synthesis of such systems. Despite the fact, that logistics processes should be highly structured, in reality, they constitute spaghetti type. There will be offered the framework to work with such processes by help of context-aware information.

The bottleneck detection is a crucial part of the continuous improvement of processes. The bottleneck in the process can limit working hours, information flow, materials, and products. Thus, the biggest bottleneck in the process trace significantly impacts the capacity and throughput of the whole system. Two concepts – theory of constraints and lean management – focus on detection and reducing possible bottlenecks/weak points by reconsidering process management. It leads not just to improving KPIs, but also to knowing risks and threat points in the system.

The common methods of bottleneck detection – Statistical, Bottleneck walk and Bottleneck analysis with process mining methods – were discussed in the current section. Despite their advantages, the flexible and appropriate method of bottleneck detection stays relevant. We described the drawbacks of each method. Most of the time present bottleneck detection methods are data-centric, not aligned with the process model, and do not consider loops and concurrencies. However, this situation can be improved by mixing the advantages of each method and using context awareness.

Finally, the predictive module presents widely used predictive models – time series, regressive and transition system models – and their characteristics. All described predictive models will be applied to the logistics process in the application part and evaluated using the metrics highlighted in the section.

Combining knowledge of logistics processes and modern methods of analytics - process mining, data mining, machine learning and statistics – can undoubtedly help improve both fields and adopt algorithms to a wider range of process types. The results of its use directly depend on the settings and understanding of the analyzed process. A prototype decision support system or process map can be developed for different participants involved in the interconnected logistics processes.

In the next chapter, we will discuss how to deal with spaghetti models by context awareness in terms of different types of process mining. And, after, use the obtained process model to enhance it with the results of another analysis. It will also present how mixing the advantages of bottleneck detection can be improved with the help of the process model. Moreover, the two-step predictive model, which considers context awareness as an input parameter, will be proposed.

Chapter 3

Methodology of applying process mining in logistics for predictive analysis

The main aim of this study is to extract knowledge and models from real-world raw event logs of interconnected logistics processes and to highlight problems to further improve the processes, particularly for further process map development.

As was mentioned in the previous part, additional research is needed for combining PM with other data mining and statistical methods as well as applying their techniques to logistics processes. This type of process remains to be a black box for analysts as well as for business/process owners. There are many simulations and normative models to understand what the logistics process is, but most of them do not aim on reflecting the real/descriptive process model. In addition, logistics processes were supposed to have a structured type. However, in the real world, they are represented by so-called *spaghetti* models due to their complexity and the many parts involved.

In this part of the work, we can identify several interconnected promising approaches to better analyze and synthesize complex processes with their key aspects and attributes. These approaches have not been applied rigorously to logistics processes so far. The main difference from other studies is a focus on context-awareness of the process and having it as an input parameters for the main steps of the research. To have a quick overview, the simplified scheme for the research is shown below.

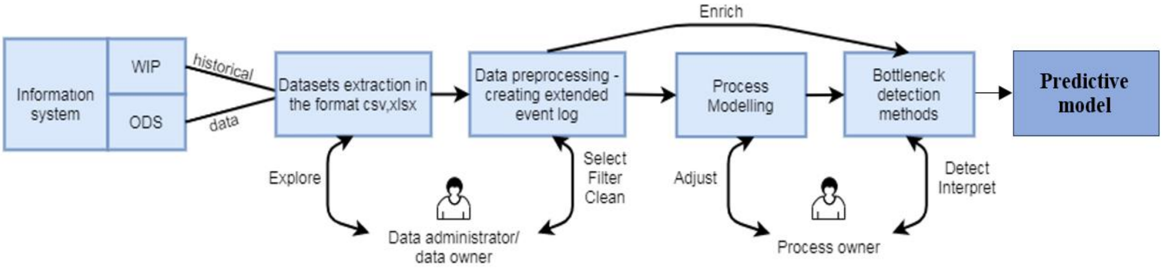

Enrich
Information system
WIP
ODS
historical
data
Datasets extraction in the format csv,xlsx
Data preprocessing - creating extended event log
Process Modelling
Bottleneck detection methods
Predictive model
Explore
Select
Filter
Clean
Data administrator/ data owner
Adjust
Process owner
Detect
Interpret

Figure 3.1: Simplified roadmap of the methodology part. The figure is taken from [68] with an extension for predictive model module

The methodology consists of five distinctive phases:

- logistics event log extraction,
- data preprocessing,
- process modelling for complex non-structured *spaghetti* processes,
- bottleneck detection analysis
- and a two-step predictive model, including results of all previous steps.

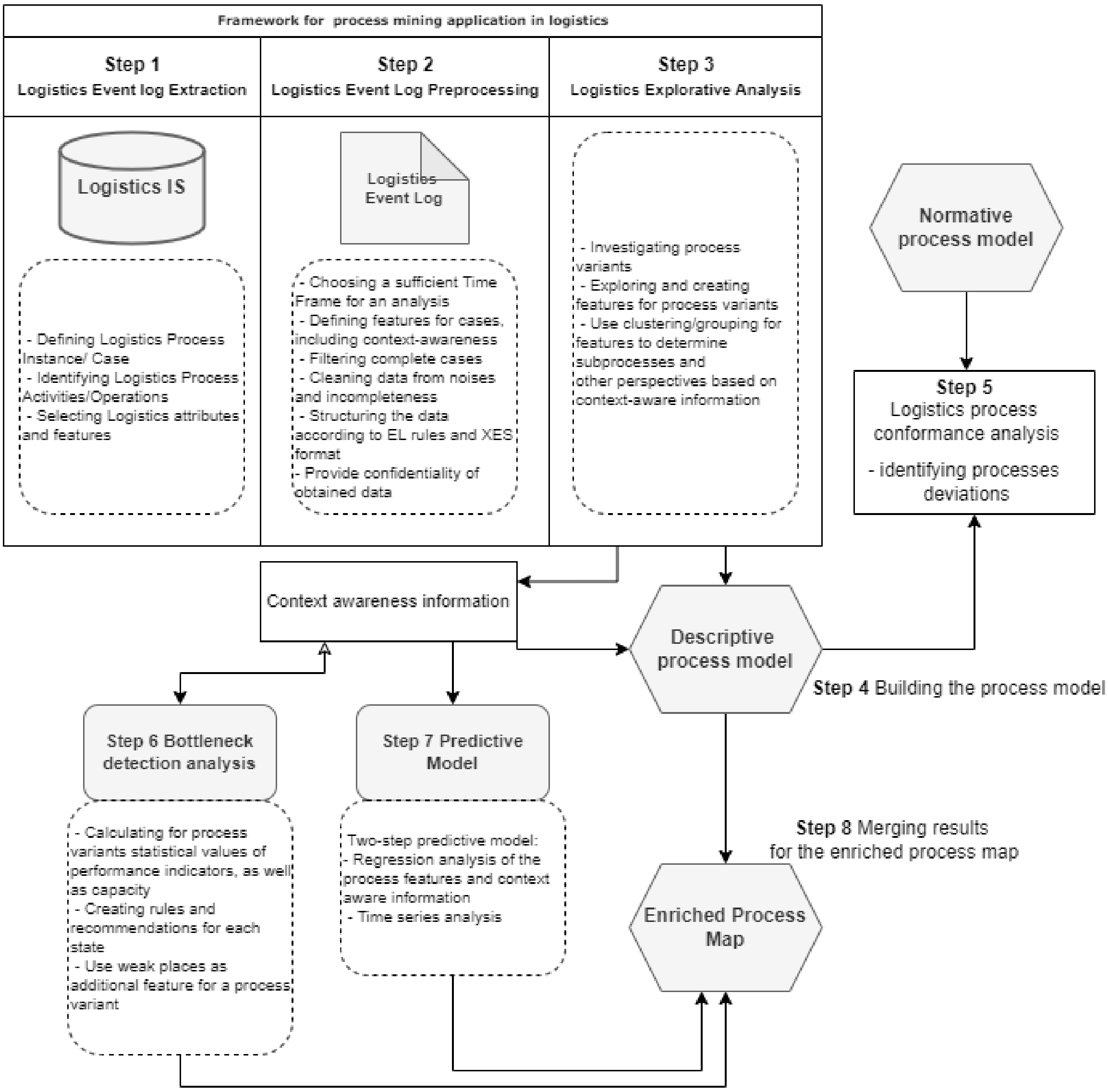
All steps are interconnected and affect the results of the following ones. Extracting data from the information system (IS) combine datasets from *work-in-progress* (WIP) and *operational data store* (ODS) parts depending on the structure of IS in the company, analysis and required results. ODS tables provide storage for data needed for internal and external reporting. WIP refers usually to the raw materials, labor, and overhead costs incurred for products that are at various stages of the logistics process. Considering Figure 3.1, three parts are impacted by context awareness:

- process modelling,
- bottleneck detection methods
- and predictive model.

Eventually, it is highly recommended to adjust all obtained results and models with data and process owners to avoid wrong interpretations of data or to bring new aspects of knowledge to them.

3.1 The main framework for a process mining application in logistics

In order to apply process mining to the logistics processes, we offer the framework presented in Figure 3.2.



Framework for process mining application in logistics
Step 1
Logistics Event log Extraction
Step 2
Logistics Event Log Preprocessing
Step 3
Logistics Explorative Analysis
Logistics IS
Logistics Event Log
- Defining Logistics Process Instance/ Case
- Identifying Logistics Process Activities/Operations
- Selecting Logistics attributes and features
- Choosing a sufficient Time Frame for an analysis
- Defining features for cases, including context-awareness
- Filtering complete cases
- Cleaning data from noises and incompleteness
- Structuring the data according to EL rules and XES format
- Provide confidentiality of obtained data
- Investigating process variants
- Exploring and creating features for process variants
- Use clustering/grouping for features to determine subprocesses and other perspectives based on context-aware information
Normative process model
Step 5
Logistics process conformance analysis
- identifying processes deviations
Context awareness information
Descriptive process model
Step 4 Building the process model
Step 6 Bottleneck detection analysis
Step 7 Predictive Model
- Calculating for process variants statistical values of performance indicators, as well as capacity
- Creating rules and recommendations for each state
- Use weak places as additional feature for a process variant
Two-step predictive model:
- Regression analysis of the process features and context aware information
- Time series analysis
Step 8 Merging results for the enriched process map
Enriched Process Map

Figure 3.2: The main framework for process mining application to logistics processes

Step 1. I consider that IS has already collected and stores all data from different systems needed for analysis. Four required parameters of the event log (case id, activities, timestamps, and additional parameters) have to be formed already at this point. Dataset's extraction implies exploring and understanding the data structure and includes:

- *Defining logistics process instance – Case ID.*

This step defines which part of the logistics process will be analyzed and shown. As long as we focus mainly on the ship handling process, what is part of the loading/unloading operations of the seaport, ship ID will be taken as a case.

- *Identifying process activities* (operations, steps, resources, etc.) *and their timestamps.*
- *Defining additional logistics process attributes and features* (costs, teams, weather conditions, berth number, weekdays, etc.).

The more parameters we have, the higher level of flexibility can be reached in the following steps. Here, we already may collect context-aware information about the system, its updates, and possible connections with other data sources. In the case of big data, this step can also consider regrouping or clustering data to reduce diversity.

To explore other possible scenarios and perspectives of the logistics processes, Table 3.1 presents different cases. For example, when using documents as a case, the movement of cargo through the seaport infrastructure can be represented as a path (set of nodes), connecting elements of the intraport cargo-transferring chain. During the passage of the track, the instalment of cargo (token) passes through the different elements of the cargo-transferring process (places) by performing cargo handling operations (running of the tokens along arches that connect nodes). Any activity can be represented in the form of documents transmitted between their editors on a certain scheme in accordance with the prescribed rules. Alternatively, the process model can be reached with documents (where it is necessary) directly.

Table 3.1: Scenarios for maritime process management

Case ID (name/number)	Process examples	Activities / Operation examples	Timestamps	Additional attributes
Ship	Ship handling process at the terminal	Ship arrival, mooring, loading, unloading, piloting, departure	Start and end time for each operation	Type of cargo, amount of cargo, type of ship, operator, weather condition, berth, weekdays, costs, country of consignee/sender. Can be used for further research, e.g. defining parameters, which affect the whole process and its performance indicators the most.
Container (other types of cargo/products)	Following the process of container (or other types of cargo) shipment from sender to the consignee	Operations with containers/cargo at terminals, warehouses, forwarders, railway or aircraft transportation		
Warehouse	Process of storage and movement of cargo	Cargo arrived, cargo was replaced…		
Resources workload (operators, cranes, equipment)	Resource analysis, Network analysis	Messages between workers		
Document	Document flow, administration processes	Document signed, the document was sent to…		

Step 2. This step is crucial and highly impacts all results of the analysis. In the previous chapter, we already mentioned raw data issues. Without preprocessing, even the most powerful algorithms will lead to the wrong not-interpretable models. The algorithms also are not able to work with missing values and just skip them in calculations what leads to losing important processes and statistical information. Therefore, an event log should be prepared for the further steps accordingly:

- *Preprocessing data based on the time parameter.*

Choosing an appropriate time frame, calculating end time stamps (if they are absent), process the level of granularity in times. Depending on the runtime of a single process case, it can be better to consider a one-year interval. For defining the time frame, we use the method which was described in the work [63]:

$$timeframe = expected * 4 * 5 \tag{3.1}$$

The 4 value ensures that at least 4 cases were started and completed after each other. The 5 accounts to the occasional long-running cases (according to the Pareto principle) and ensures that cases that take up to five times longer are included in the extracted time window.

- *Filtering of completed cases.*

It means we should define the first and last activities for each process and subprocess. It will cut off incomplete cases and, therefore, the number of variants. Also, it can happen that one wants to consider specifically incomplete cases and understand the reason. So, the implementation of these steps highly depends on the goal of the analyses and case.

- *Cleaning data from noises, incompleteness, and outliers.* Solve uncertainty.

Missing values should be replaced with mean/median/mode values or according to production rules and be deleted just as the last resort. The method for outliers' detection was described in our work [54]. Usually, in the real world, the data distribution is not normative and has an abnormal type with left shift. For this type of distribution, we use thresholds which are defined as the interval

$$Q_2 \pm \sigma \tag{3.2}$$

where
Q_2 is the second quartile (median) and σ is the standard deviation.

Figure 3.3 shows the distribution of processes' time duration, and this approach helps to cover needed processes for an analysis better than with μ - expected value.

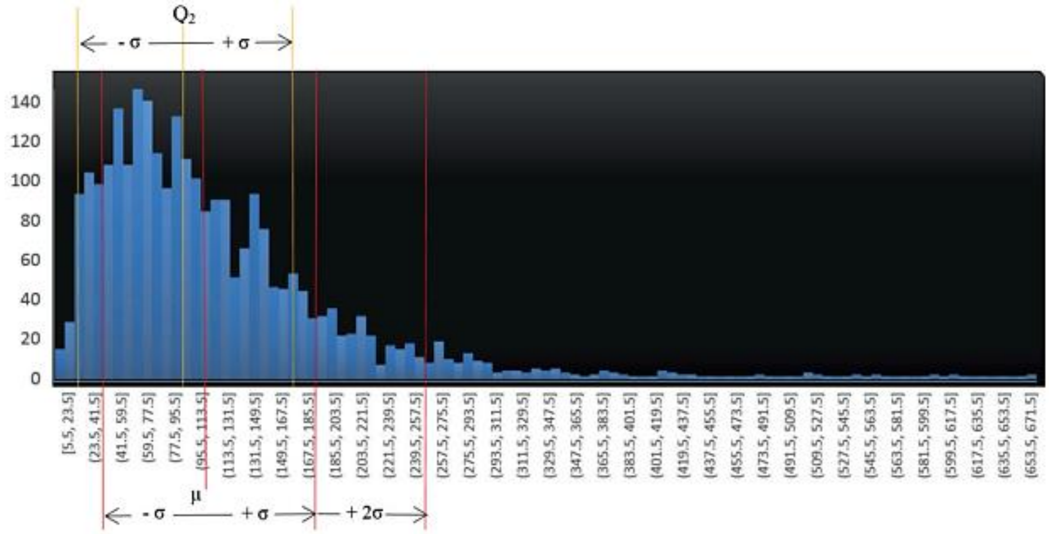


Figure 3.3: The defining thresholds for time duration of the processes [54]

Additionally, logistics data often come from different systems and do not have a unified standard. For instance, two similar operations *Weather conditions* and *Weather conditions (Other)* (or *NOR tendering-berth* and *NOR tendering-road*) have two different identifications. This problem could be tackled by the aggregation of operations according to experts' knowledge or based on an international standard. Meanwhile, we can create additional features to differ activities by data source or resources. For example, two different operations *Mooring with Operator*$_1$ and *Mooring with Operator*$_2$ can be combined into one operation *Mooring* and adding additional resource parameter *Operator*. Reducing and unifying the number of operations improves quality and representation of final process model.

Another issue of equal importance is uncertainty in the data. There are duplicates with different timestamps within a case, which may affect the results of algorithms. For example, we can find two the same operations in one case with start and end times (1:00; 2:00) and (1:00; 3:00) respectively. This uncertainty can be solved by help process owners or using the common time frame that covers both times (for our example, the operation has time (1:00; 3:00))

- *Defining features for cases, including context awareness.*

In the first step, we chose the set of possible influencing parameters of the process, here we specify them for the case.

- *Structuring the data according to EL rules and MXML/XES format.*

Existing process mining algorithms work with XES (eXtensible Event Stream) or MXML(Mining eXtensible Markup Language) formats. Nowadays, there are many converting tools for these formats, but the structure of the event log should respond to the rules which are described in the second chapter. For example, the event log should have cases; activities and timestamps, activities should be ordered, etc.

- *Provide confidentiality of obtained data.*

According to the FACT issue, even part of the process with names of instances can reveal the whole of the process and its parameters for competitors. So, names should be encrypted before building the model.

Step 3. After the event log was created, we should conduct a *Logistics explorative analysis* to understand cases and their parameters. At this step, it is possible to find new knowledge about the process and, thus, even go back to the previous step for corrections or specifications. This step is essential for context-aware information for process variants (a unique trace from the beginning to the end of the process or the certain order of activities) and it will be used as an input for proposed methods in this research.

- *Investigating process variants.*

Explore the structure of different types of processes, the order of activities, missing operations, check data dependencies in the dotted chart and so on.

- *Exploring and creating features for process variants*.

Here we can use all possible statistical measurements and general information for each process variant, which comes from the first and second steps. Moreover, additional features can be calculated and used as performance or productivity characteristics.

- *Use clustering/grouping for features to determine subprocesses and other perspectives based on context-aware information.*

In the first step, we collected possible features for the process and here they can be used to build hidden subprocesses, what improves in general the level of the detalization for the resulted process model.

Step 4. Process Modelling – building the process model

- *Discover the real process model with various algorithms* (alpha miner, fuzzy miner, inductive miner, etc.).

According to the specification of our domain area, we apply a fuzzy miner.

- *Use different resulted notations to avoid their limitations* (workflow, BPMN, DFG).

Taking into account the limitations of DFG, described in [24], we use DFG notation as the most presentable and interactive.

- *Explore the main process flow.*

Once the process model is built, we can explore the most frequent activities, resources, paths, relations between operations, loops and concurrencies, possible drill-down levels, and so on. Discovering the real descriptive model is essential for further analysis.

Step 5. Logistics process conformance analysis

This step is needed in auditioning and identifying process deviations. It connects the obtained descriptive process model with the normative one, which was created by process owners. There are some reasons describing why existing models do not reflect the real situation in logistics:

- *Subjectivity.* Depending on the roles and perspectives, everyone has a subjective picture of the process. Therefore, interviews, workshops and expert knowledge are not enough to discover the real process.
- *Partial view.* Processes are usually complex and include the participation of multiple people, teams, departments and so on. The challenge is that there is no single person, who performs the complete process and knows it from the beginning to the end.
- *Change.* Processes are changeable under influence of different causes (seasons, product changes, weather conditions, covid lockdown, etc.). The real process model should present relevant data, be flexible and actual at the time of analysis.
- *Invisibility.* Some cases can be unfollowed because of system breakdown or data issues. Also, there are some *black-box* subprocesses that might be omitted due to their trivial nature, but they can impact the results of the analysis later.

In this case, the comparison is useful to improve the quality of both models and helps to connect theory/guesses and reality. Nevertheless, this step is optional and is not considered in the application part.

Step 6. Bottleneck detection analysis.

- *Calculating for process variants statistical values of performance indicators*, as well as capacity.

Taking the specific process variant, we calculate statistical values for each position of the process, both path and activity. Selecting the flow time as a goal performance characteristic of the bottleneck analysis, the path will represent the time in queue (waiting time) and the activity will show elapsed time (active time). In the case of capacity as a goal performance characteristic, the values of the path– items/cases in the queue, and the activity values – item/cases are processed in the operation by the resource. It helps to see concurrent processed cases and the amount of waiting units that leads to the calculation of additional performance parameters.

- *Creating rules and recommendations for each state.*

Considering thresholds for bottleneck detection and the direction for possible bottleneck roots, recommendations can be developed with the help of the process owner, who knows explanations for specific issues.

- *Use weak places as an additional feature for a process variant.*

After we get information about weak places (also called threat/risk spots), as well as recommendations for them, we can use it as an additional feature for a process variant. Thus, this data can be taken as input for the predictive model further.

Step 7. Predictive Model

Finally, we have all source data to utilize them as predictors for the target variable. We offer to use a two-step hybrid predictive model that includes feature analysis from one side and timeline analysis from another side.

- *Regression analysis of the process features and context-aware information*
- *Time Series analysis*

Step 8. Merging results for the enriched process map.

The last step will be theoretically described in the work. However, the creation of the process map requires more collaboration with current projects and the involvement of process owners as well as software engineers. The scope of this thesis does not cover the development of the process map and suggests it for future research. Nevertheless, critical elements and methods of the process map are connected to the ones, explained in the research.

The framework is the core of the work and the main result of the research. It is adjustable and can be easily applied to different kinds of processes. It provides a new ordered way of process mining application and brings a full understanding of the current process as well as the nature of logistics processes. Knowledge obtained through the framework execution can be further used for process optimization, better structuring of work, decision support, recommendations, and planning. Eventually, it is highly recommended to adjust all obtained results and models with data and process owners to avoid wrong interpretations. Working with the framework helps data and process analysts to get a key understanding of the whole process, ask meaningful questions and propose particular ways for improvements. Steps 3, 4, 6, 7, 8 are discussed in the following sections.

3.2 Process Variants

Step 3 of the framework is devoted mainly to the explorative analysis of process variants and their features.

Process variant - a unique trace from the beginning to the end of the process or the certain order of process activities. It defines the type of the process. As a rule of thumb, we can formulate the Pareto principle[22] for process variants as follows: 20% of most frequent process variants cover 80% of all cases and the rest 80% of variants cover just 20% of cases. So, the first part shows *highways*, the most common traces and the second part shows unique/specific cases. Thus, we can say that behaviour of some most frequent process variants defines the behaviour of all system in general. Nevertheless, both groups are needed to be explored depending on the goals of the research. The sample of process variants explorative analysis is presented in the figure below.

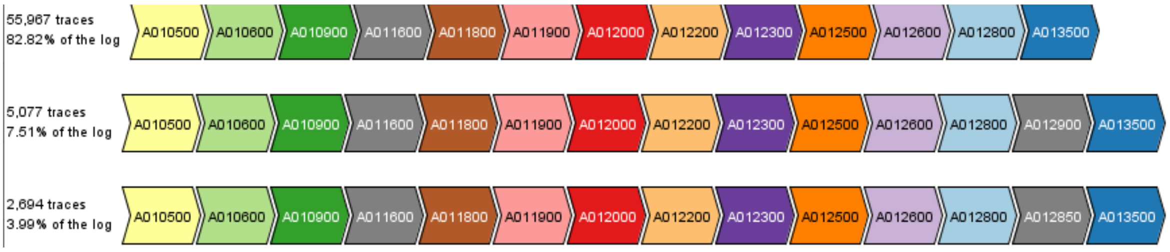


Figure 3.4: The sample of process variants. The figure was created by PROM

Figure 3.4 shows three process variants and we can see that the first trace (activities' order) was taken by 55967 different cases. It means that this trace describes the process for almost 83% of all analyzed cases from the event log.

We can define a set of features as:

$$F = \{f^0, f^1, f^2, \dots, f^n\} \tag{3.3}$$

[22] www.en.wikipedia.org/wiki/Pareto_principle

where

f^0 – is a *target feature* – the feature we want to predict. Depending on the analysis, this feature can be changed and set as another performance parameter. For instance, one time we focus on the time flow and consider duration time in this feature, another time – the goal is to define the resource which will be in charge or the number of bottlenecks and places where they occur most of the time.

$S = \{1,2,..,m\}$ – is a *set of situations*. In the scope of this thesis, a case ID and a process variant can be defined as the situation.

f_s^i – is the *value of the feature i* $(0 \leq i \leq n)$ for situation *s* $(1 \leq s \leq m)$.

Therefore, the feature matrix will be presented as follows:

Table 3.2: Matrix of features

s	f^0	f^1	f^2	…	f^n
1	f_1^0	f_1^1	f_1^2		f_1^n
2	f_2^0	f_2^1	f_2^2		f_2^n
…					
m	f_m^0	f_m^1	f_m^2		f_m^n

This feature selection part was already pointed out in the first (the situation is the whole process) and second (the situation is the specific case) steps. When the situation is taken as case id (e.g., ship – for logistics processes), follows features can be considered:

Table 3.3: Matrix of features for situation=case

s = ship case	f^0= flow time	f^1 = start workday	f^2=process variant	…	f^n=country of origin
9997	2 days	Saturday	I		USA
9998	3.5 days	Monday	IV		Singapore
…					
10100	5 days	Tuesday	V		Norway

Here we see different sets of features which can impact the target value f^0. It is interesting to notice that one of the features belongs to the context-aware information and is generated after the process modelling step – process variant. In turn, this feature f^2 can be also represented as a set of features related to the specific process variant.

Table 3.4: Matrix of features for situation=variant

s = variant	f^0= average flow time	f^1 = resources	f^2=# of activities	f^3= risk places	…	f^n=common delay (in case of repairs)
I	2.8 days	John	10	Activity 4		3h
IV	3 days	Lukas	6	Activity 6		20min
…						
V	4.8 days	Jan	15	Activity 9		3 days

Thus, the matrix (see Table 3.4) presents the sample of context-aware information for different types of the process. These features can be obtained only after previous steps of analysis like process modelling, bottleneck detection, exploring subprocesses, etc. For example, Variant I has the next parameters: {2.8 days, John, 10, Activity 4, …, 3h}. Utilizing these parameters significantly impacts the quality of possible predictive models empowering even simple algorithms. Also, one of the additional context-aware parameters for the process variant can be similar process variants, which brings a new point for analysis of alternative process ways. For instance, if features of two variants are similar but the order or name of activities are different, these two variants can be considered as a group/cluster and used for recommending alternative trace (activity order) in the risk weak places of the process. This research is out of scope of this thesis, but more detailed can be found in the work [31].

3.3 Approach to process *spaghetti* models with context-awareness

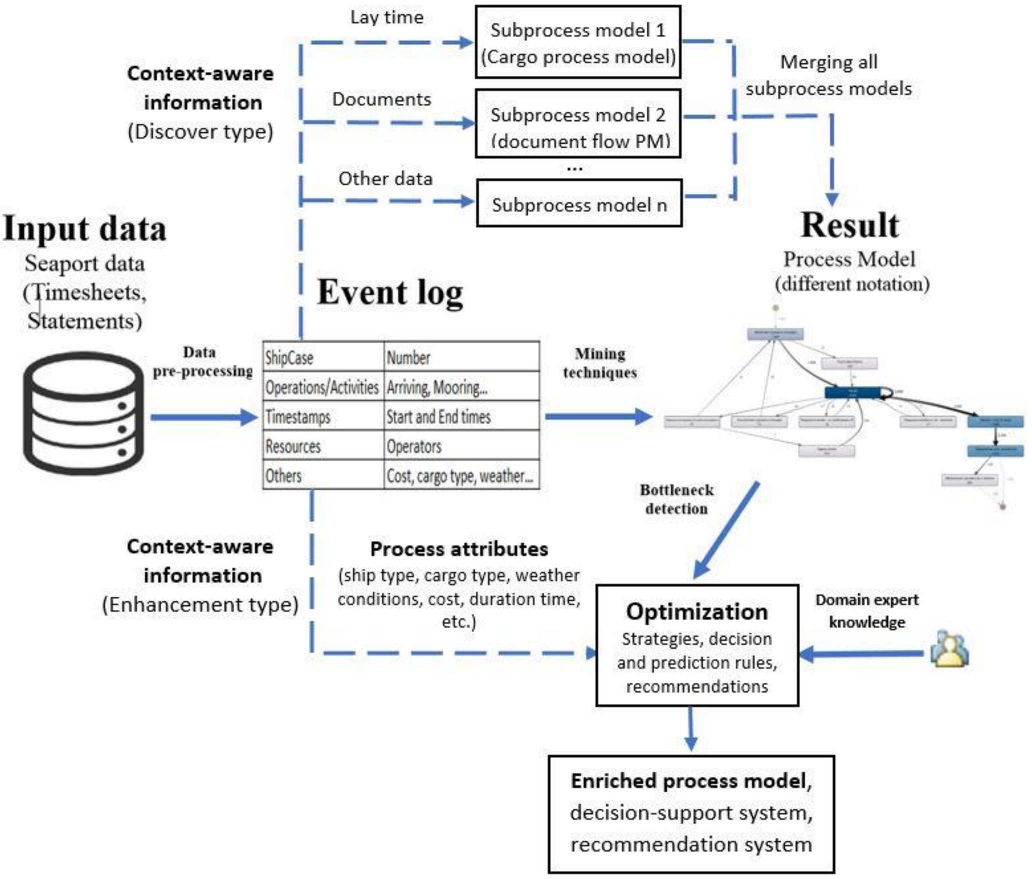
As we already know that logistics processes are presented by complex not-structured process models – *spaghetti* models. It happens because of many participants, hidden activities and perspectives included in the process. The incomprehensible and unreadable model cannot be used for representing real process flow to stakeholders or for enhancement with additional process attributes. There are methods how to work with these kinds of unreadable and useless models:

- filtering processes,
- drilling up to get the top view of the main process variant,
- changing coefficients for process modelling algorithms.

All these steps result in losing crucial processes and their parameters along with different perspectives. Our approach can be described by *divide and conquer* algorithm. We do not need to skip the information, but instead, we will add more information for the input. It allows to create subprocess models, which can be used after to see the main process from different perspectives, as well as drill down to the lower levels revealing hidden unknown operations.

We offer to use context information according to two main types of process mining – discovery and enhancement. As we already mentioned in the second chapter, it can refer to a person, place, time, frequency of events, operators, and other attributes of the process, case or process variant.

For this specific type of process model, we propose a scheme for logistics processes (see Figure 3.5) which can help to avoid complexity in process models. Two dotted lines show possible solutions for improving process mining techniques in terms of logistics processes. In the application part we build common process model and one of subprocess models.



ShipCase	Number
Operations/Activities	Arriving, Mooring...
Timestamps	Start and End times
Resources	Operators
Others	Cost, cargo type, weather...

Figure 3.5: The scheme of processing *spaghetti* complex process models in terms of logistics processes (extension of the main framework) [67]

As we can see in Figure 3.5, steps 1 and 2 are briefly mentioned and the primary focus here is on Step 3. We consider here using context awareness in terms of two types of process mining: discovery and enhancement types.

Discovery type. PM techniques are flexible but strongly depend on input data. Therefore, context information can be used to detect sub-processes, find relationships among process instances, and enhance the result of discovered models from various perspectives. Then, all the sub-process models can be used for making different perspectives of the general process model and avoiding *spaghetti*. Moreover, context information can provide a specific point of view for a certain objective of analysis.

Thus, using $F_{process}$ *(s=process)* we obtain different perspectives of the process (document flow, warehousing connections, resource connections, etc.). It can be information about documents type to discover the document-flow process or transitional data. These features might be grouped as it is shown in Table 3.5, to obtain different perspectives of the process. The more logistics features are presented in Appendix B.

$F_{activity}$ *(s=activity)*, in turn, brings new levels of subprocesses dividing all operations into groups. It might be shown as a cluster with the opportunity to drill down. In this way, all activities will be hidden (with the opportunity to reveal them), but not skipped. At the same time, the process model is not overloaded with all possible activities at once. The main thing here is to find the feature which define the subprocess. For example, in our case, we propose to use such additional parameter as laytime (i.e., the amount of time specified in a charter party for a vessel's loading and unloading) for building the cargo handling subprocess, since just these activities have this additional time parameter.

Table 3.5: Feature selected for grouping for complex models [39]

Transaction info	Master Data	Cargo info	Location info
- Price per unit - Quantity of goods - Total value of goods - Currency - The purpose for export - Payment terms (days) - Delivery terms (days) - Freight charge terms	- VAT # - Full name - Address - Contact info - Code	- Commodity codes - Weight - Number of packages - County of origin - Country of departure	- Vessel ID - Positioning on vessel - Trailer ID

Enhancement type. For this type of PM, we can combine context-aware information with different statistics, data mining and machine learning techniques. Logistics process activities have the multitude of information available behind the observed process. Type of cargo, type of vessel, weather conditions, related operators, traffic jams, number of the berth, etc. – all these attributes have inner dependencies, which can be utilized for making decision rules, recommendation rules, and predictions [64]. Here, not just $F_{process}$ and $F_{activity}$ can be utilized, but also $F_{variant}$. $F_{variant}$ have already information about the trace and brings new calculated features for the predictive model. Also, once we recognize the process variant for the case, all its parameters can be used for prediction without the predictive model. It gives the opportunity to foresee the next activity, resources, estimated end time, or possible weak places on the trace. Also, performance and productivity characteristics of the process can be obtained, if they were calculated in advance (e.g., activities/hour, the capacity of the trace, possible rework/loops, number of units per hour, etc.

After filtering methods applying context awareness is promising approach to decrease variability of the processes. Once, we use a cluster of activities it helps to reduce numbers of process variants, that in turn leads to more structured final descriptive process model. In the next parts we focus more on enhancement type of process mining, especially taking the context awareness as an input parameter for bottleneck detection and predictive methods.

3.4 Bottleneck detection with the process model and confidential intervals

The bottleneck in the process can limit working hours, information flow, materials, and products. Thus, the biggest bottleneck in the process trace significantly impacts the capacity and throughput of the whole system. We already mentioned before that the capacities of the process trace before and after the bottleneck are not important, the throughput of the process will be defined by the capacity of the bottleneck itself. Talking about capacity, we take not just units on the trace but also time flow.

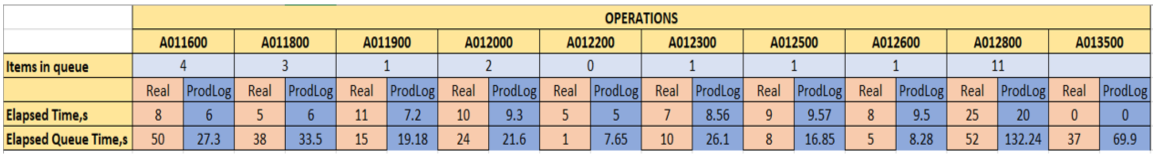
This section aims on finding the most time-consuming places in the process and the reasons caused them. We propose two bottleneck detection methods:

1. *TimeLag method.* It is effective in terms of theory of constraints. It can quickly define the biggest constraint on the trace, but reasons are still not traceable.
2. *Confidence interval method.* It is more detailed, connected to the process variants and helps with not just bottleneck detection but also guessing the roots for a delay.

Using the TimeLag method and the confidence interval method, the workflow will identify the bottlenecks from the dataset and add the root cause for the bottlenecks based on the operation and delay incurred by downstream or upstream process. The underlying process model makes the detection more transparent, accurate and valuable.

To apply both methods, it is necessary to obtain statistical values which were made in Step 2 and 3 of the framework. Also, we will need measurements of the transition system (see Figure 2.33): elapsed time, sojourn time and remaining time. Otherwise, we will not discuss the remaining time and focus on the two first ones. To connect this terminology to the information system we worked elapsed time in the transition system is named elapsed queue time and sojourn time from the transition system is elapsed time in the work. Automatically obtained descriptive data from the historical dataset and confirmed with the process owner can look like in the Figure 3.6. They can be also compared to the real manual observations to notice the differences. The Figure 3.6 contains next parameters:

- Items in queue (buffer) – cases/unit capacity that can be placed before the operation on the trace.
- Elapsed time (ET) – time is needed for the operation to be completed.
- Elapsed_queue time (EQT)– time, during which the case waits in the queue before starting the operation.

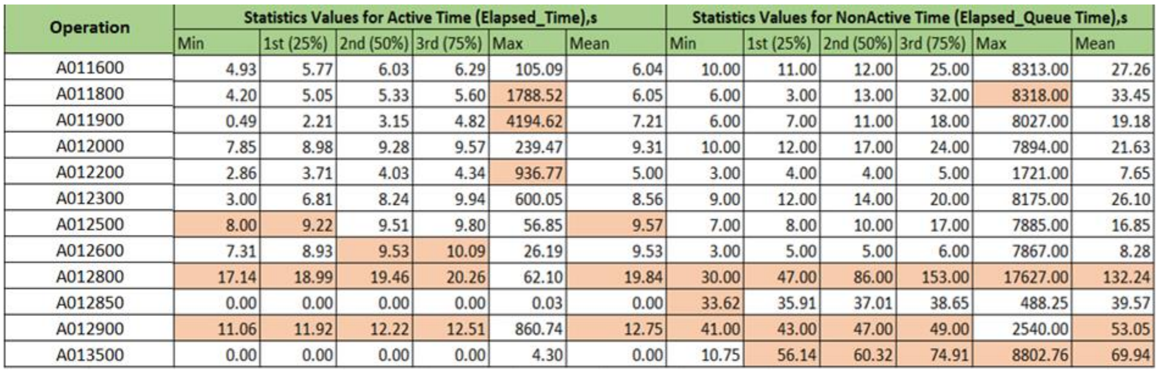


	OPERATIONS																			
	A011600		A011800		A011900		A012000		A012200		A012300		A012500		A012600		A012800		A013500	
Items in queue	4		3		1		2		0		1		1		1		11			
	Real	ProdLog	Real	ProdLog	Real	ProdLog	Real	ProdLog	Real	ProdLog	Real	ProdLog	Real	ProdLog	Real	ProdLog	Real	ProdLog	Real	ProdLog
Elapsed Time,s	8	6	5	6	11	7.2	10	9.3	5	5	7	8.56	9	9.57	8	9.5	25	20	0	0
Elapsed Queue Time,s	50	27.3	38	33.5	15	19.18	24	21.6	1	7.65	10	26.1	8	16.85	5	8.28	52	132.24	37	69.9

Figure 3.6: Descriptive parameters for the process trace [68]

For example, for operation A011800 – 3 items can be in queue before this operation, the real elapsed time is 5s, the mean elapsed time from event log is 6s, the real time for a SFC in queue before this operation (on the way from A011600 to A011800) is 38s and the mean elapsed time obtained from event log is 33.5s.

Figure 3.7 shows descriptive statistical characteristics like min/max values, data quartiles (three values that split sorted data into four parts, each with an equal number of observations – 25, 50, 75% of data set size), and mean (average) values which were calculated before. The three longest operations are highlighted.



Operation	Statistics Values for Active Time (Elapsed_Time),s						Statistics Values for NonActive Time (Elapsed_Queue Time),s					
	Min	1st (25%)	2nd (50%)	3rd (75%)	Max	Mean	Min	1st (25%)	2nd (50%)	3rd (75%)	Max	Mean
A011600	4.93	5.77	6.03	6.29	105.09	6.04	10.00	11.00	12.00	25.00	8313.00	27.26
A011800	4.20	5.05	5.33	5.60	1788.52	6.05	6.00	3.00	13.00	32.00	8318.00	33.45
A011900	0.49	2.21	3.15	4.82	4194.62	7.21	6.00	7.00	11.00	18.00	8027.00	19.18
A012000	7.85	8.98	9.28	9.57	239.47	9.31	10.00	12.00	17.00	24.00	7894.00	21.63
A012200	2.86	3.71	4.03	4.34	936.77	5.00	3.00	4.00	4.00	5.00	1721.00	7.65
A012300	3.00	6.81	8.24	9.94	600.05	8.56	9.00	12.00	14.00	20.00	8175.00	26.10
A012500	8.00	9.22	9.51	9.80	56.85	9.57	7.00	8.00	10.00	17.00	7885.00	16.85
A012600	7.31	8.93	9.53	10.09	26.19	9.53	3.00	5.00	5.00	6.00	7867.00	8.28
A012800	17.14	18.99	19.46	20.26	62.10	19.84	30.00	47.00	86.00	153.00	17627.00	132.24
A012850	0.00	0.00	0.00	0.00	0.03	0.00	33.62	35.91	37.01	38.65	488.25	39.57
A012900	11.06	11.92	12.22	12.51	860.74	12.75	41.00	43.00	47.00	49.00	2540.00	53.05
A013500	0.00	0.00	0.00	0.00	4.30	0.00	10.75	56.14	60.32	74.91	8802.76	69.94

Figure 3.7: The descriptive statistical characteristics for each operation [68]

Figure 3.7 will be useful for the second bottleneck detection method.

TimeLag method

TimeLag method has as an input statistical information from event log and process variants and analyse time for a case. The previous and the next case from the operation where the bottleneck happened are observed, and the method searches where the delay happened in first place and highlight that operation/resource/activity which played a main role in creating that ripple effect (see Figure 3.8).

It helps in the investigating the operation which caused the delay after segregating the initial bottleneck and finding the previous and next case for that incident operation. The inputs can be the case or the operation of interest which results in the bottleneck analysis flow. The method has following structure:

1. Identifying an operation of interest to calculate the root cause of bottlenecks ensure they are clustered and sorted by datetime properly, find the previous and successive case/units, identify the big transfer times and isolate them for investigation.
2. The isolated time span is taken, and the previous 50 case/units processed, and the next 50 items processed during the datetime period is taken.
3. During the isolated time range, the operations which are running indicates they were functioning good during that time, and the missing operations indicates that it was the first operation to break and resulting in bottleneck which resulted in the ripple effect.

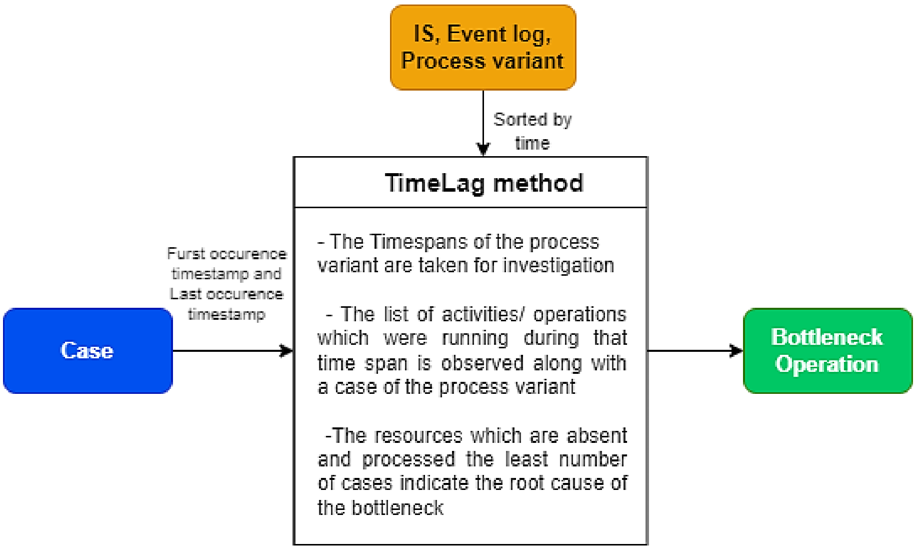

IS, Event log, Process variant
Sorted by time
TimeLag method
- The Timespans of the process variant are taken for investigation
- The list of activities/ operations which were running during that time span is observed along with a case of the process variant
-The resources which are absent and processed the least number of cases indicate the root cause of the bottleneck
Furst occurence timestamp and Last occurence timestamp
Case
Bottleneck Operation

Figure 3.8: The TimeLag bottleneck detection method

According to the theory of constraints, it is enough to get the biggest bottleneck of the system to start improvements. However, if the process has a sequence of bottlenecks, the method should be readjusted. Also, the TimeLag method is not able to find the roots and directions for bottlenecks, what is not efficient for process model. We can apply it for the process variant once the bottleneck place should be quickly calculated. The following method is taken as a more effective to find root causes and more suitable for process maps.

Confidence interval method

Confidence interval method combines advantages of statistical, bottleneck walk and transition systems methods. Taking process variant as an input and descriptive statistics from historical data we can present the method in Figure 3.9. To implement this method, one should follow next steps:

- Calculate median elapsed queue times for all operations/activities. Elapsed time considered just in cases when we want to check resource breakdowns as well.

In the example here, we take do not consider elapsed time.

- For each case – real times and previous ones are compared and give status *slower* or *faster* based on the result. According to expert's recommendations, it is preferable to add additional 6 - 7s to the confidence interval to avoid sensitive calculations. For example, if median EQT for operation A12800 is 86s accordingly, then the confidence interval might be less than 25s (low threshold is not important, since just longer time is considered as a bottleneck) (see Figure 3.7).
- For all operations with status *slower*, difference *Queue time – Median queue time* should be compared between previous and next operation. It will show direction of the bottleneck root – upstream or downstream.

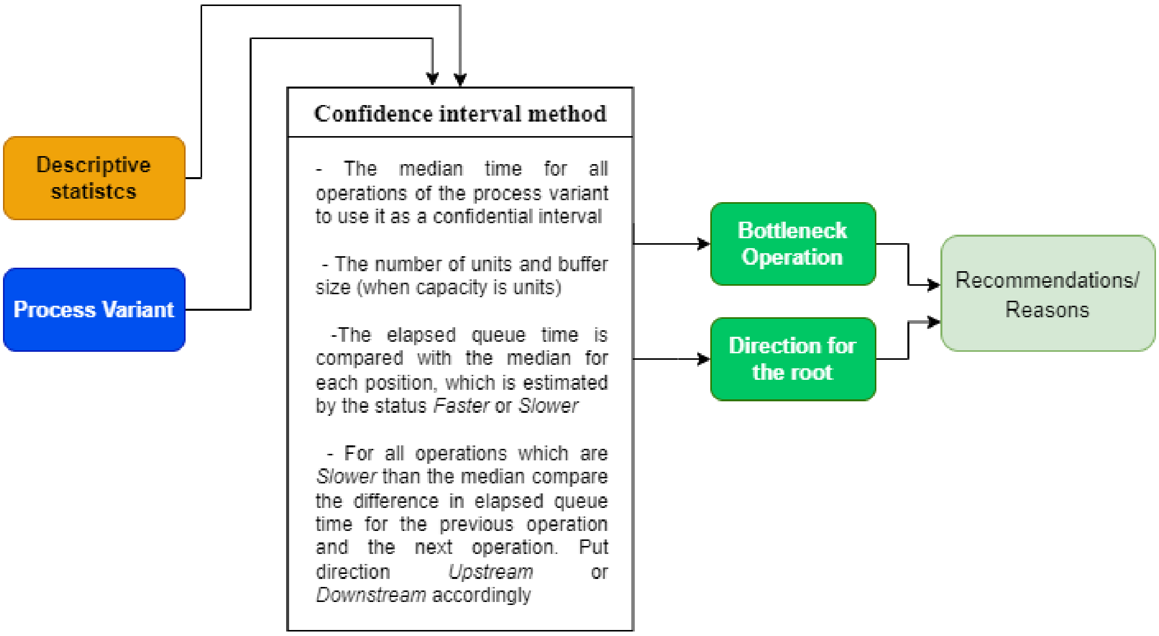

Descriptive statistcs
Process Variant
Confidence interval method
- The median time for all operations of the process variant to use it as a confidential interval
- The number of units and buffer size (when capacity is units)
-The elapsed queue time is compared with the median for each position, which is estimated by the status *Faster* or *Slower*
- For all operations which are *Slower* than the median compare the difference in elapsed queue time for the previous operation and the next operation. Put direction *Upstream* or *Downstream* accordingly
Bottleneck Operation
Direction for the root
Recommendations/ Reasons

Figure 3.9: Confidence interval bottleneck detection method

Whenever a time has a delay that means either one of the three options:

- Either the resource stops working (repairs, breakdown, absence)
- There are no resources in the queue to be processed (Upstream starvation – lack of materials)
- There is a queue in the next resource (bottleneck in Downstream – impossibility of shipment), which results in inability to process the next case in the trace.

The algorithm for Confidence interval bottleneck detection method includes two parts and presented below.

Algorithm 1: Bottleneck detection through the process

```
// the list of EQT median values for each operation in the process based
 // on historical data, can be chosen upper quartile $Q_3$, mean value or the value
 // defined be expert group

Median = [value for 1st operation,…value for nth operation]
diff= []
result= []
for i = 1 to N do
```

```
// to make calculations not too sensitive, it is recommended to add 8-9 seconds,

// depending on the operation type
    diff = Median [i] – EQT [i]
    if diff < 0 then
      result = “Slower”
    else
      result = “No bottleneck”
```

Therefore, minimum value of difference between median time and EQT – *min(diff)* – shows the longest bottleneck, *diff.index(min(diff))* – shows the index of the operation with the longest waiting time, and *index.result = “Slower”* presents all operations with bottlenecks through the process.

Algorithm 2: Bottleneck direction

```
if result = “Slower” then
    if diff[i] < diff[i-1] then
      direct = “Downstream”
    else
      direct = “Upstream”
else
    result = “No bottleneck”
```

For the last operation in the process if *result = “Slower”*, then status automatically will be *Bottleneck-shipping reason*. Usually, the process ends with transportation to another location. In order to get third possible reason (breakdown of resource), one should compare ET with the inspect time interval. As a rule, all machine resources are connected to the information management system and detection of bottleneck happens immediately.

Figure 3.10 shows an example of applying the method for a specific process variant.

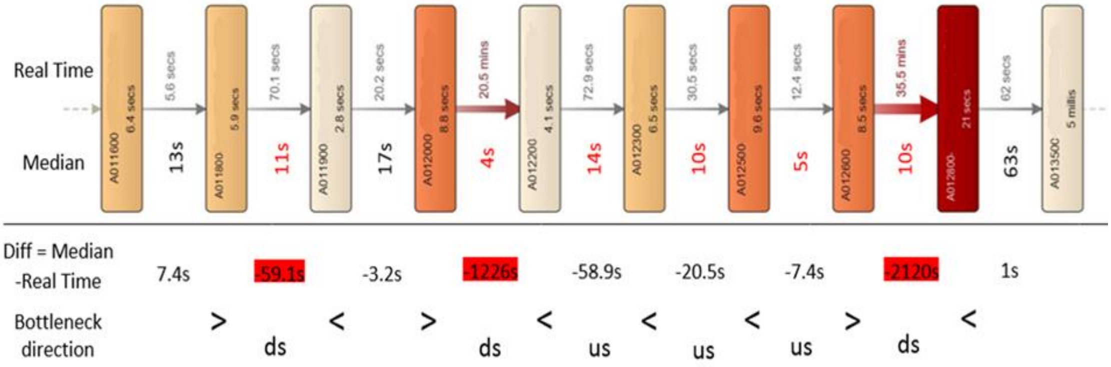


Figure 3.10: Visualization of applying the confidential interval bottleneck method to the one specific process variant [68]

Based on Figure 3.10 we can make next assumptions:

- EQT = <5.61, 70.12, 20.23, 12304, 72.95, 30.56, 12.47, 2130.8, 62.9> - elapsed queue time for each operation.
- Median = <13.1, 11.2, 17.3, 4.4, 14.5, 10.6, 5.7, 10.8, 63.9> - median time for EQT was taken from the Figure 3.7 with statistics of historical data and shows common time for general amounts of cases in the certain position of the process. These times have to be

discussed with experts and process owners. It is important to notice here that for more realistic median values the main process variant (sequencies of operations) should be taken and median values for operations A012800 and A013500 are different than in Figure 3.7, since descriptive statistics consider values for all activities and in the analyzed process variant we do not have loop/ rework. The term confidential interval was used to define correct boundaries for each part of the process. Low thresholds are not so important and will be omitted in calculations, because they show faster cases, which are not necessary for the analysis. High threshold, in a turn, determines the sensitivity of algorithms and the accuracy of results. For example, delay for index = 3 is 3.2s and was found as not crucial.

- Diff = <7.41, -59.12, -3.23, -12264, -58.95, -20.56, -7.47, -21208, 19> – the difference between median and real time.

Status *Bottleneck* is defined for all values in the set *Diff* that are lower than 0. Checking three longest bottlenecks presents next values (index:diff) = <8:-2120; 4:-1226; 2:-59.1>. These vulnerabilities were discussed with the process owner later and their influence on the whole process was proved.

Further, using the Algorithm 2, directions for all bottlenecks are identified (ds-downstream, us-upstream, nb-no bottleneck): direct = <nb1, ds2, nb3, ds4, us5, us6, us7, ds8, nb9 >. Indexes link to operations' numbers.

The longest bottlenecks are examined. They can be detected after comparison of differences between Median and Real time for each operation. The last operation of the process A013500 cannot be compared with the next operation, so if elapsed queue time is higher than median, the bottleneck can be marked at that point. The confidence intervals and reasons for bottlenecks must be discussed with process experts. Elapsed times can be also included in calculations for further analysis. It's important to notice that one delay has riffle effect and other items ate stuck in the line. So, the bottleneck detection for one specific case can prevent from big lines and delays.

After estimation root causes directions for bottlenecks, it's possible to make suggestions about real reasons for delays. Table below shows the list of possible options to create the reasons for each operation in the production line. Based on the algorithm and the list, one can make production process rules and offer some probable reasons for delays. Also, this recommendation table can be enriched with an information about breakdowns in the production line, that leads to making bottleneck reasons more accurate.

Table 3.6: Recommendations for combination (Operation+Direction)

Operation	Direction	Reasons
A011600	Downstream	Operation A011600 stopped working due to maintains or queue in downstream
A011600	Upstream	Materials are not available for this operation
A011800	Downstream	Operation A011800 stopped working due to unavailable shipment
A011800	Upstream	Materials are not available for this operation/ waiting for materials from another work center
…	…	…

The confidential interval method allows to detect bottlenecks in the process variants and define set of recommendations/reasons for each operation based on bottleneck direction. These parameters can further be used as an input for predictive model. The confidential interval is more robust and advanced method that TimeLag, it is also able to detect directions for bottlenecks. In comparison to Bottleneck walk method, the Confidential Interval method does not require manual work and all calculations can be extracted from event log and process variant variables.

3.5 Two-step predictive model

Predictive model $pred(f^1, f^2, \ldots, f^n)$ aims to predict the target variable f^0 regarding to other variables – predictors. I propose to use the two-step predictive model: random regression forest (ensemble of decision trees) – cover the need to analyze case attributes, context awareness and features of the process, and help to find correlations and dependencies between features; and SARIMA model – brings seasonality, trends, time distribution, prime time analysis into the prediction. The Figure 3.11 shows the algorithm for the predictive model.

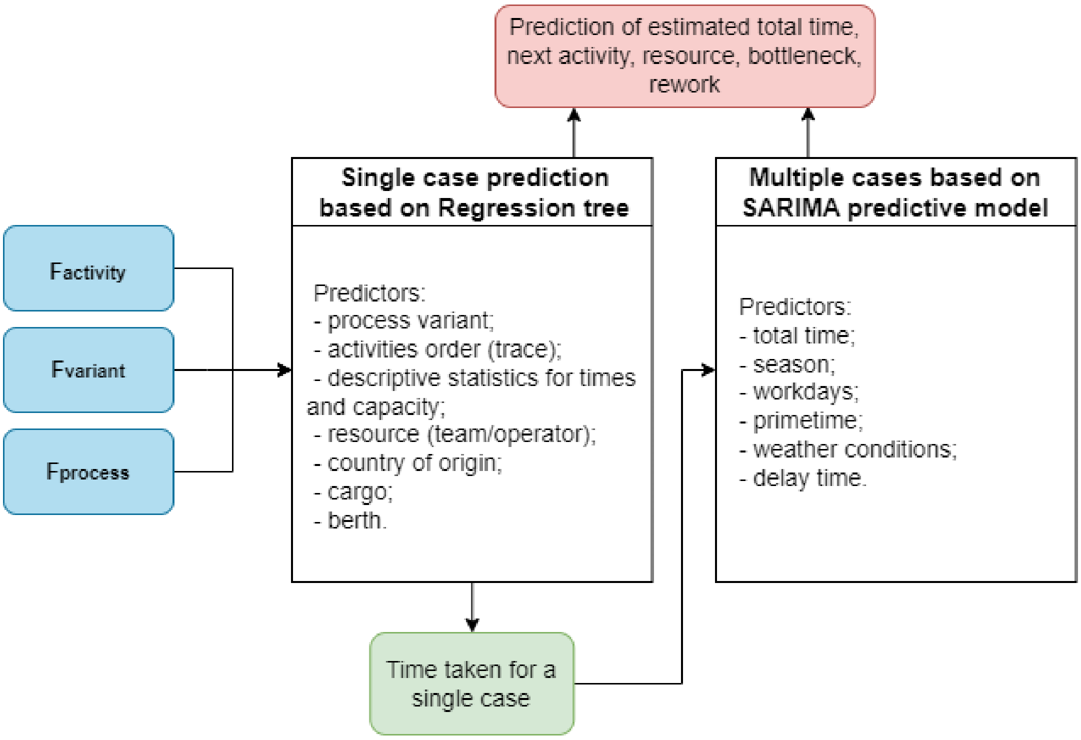

Prediction of estimated total time, next activity, resource, bottleneck, rework
Single case prediction based on Regression tree
Multiple cases based on SARIMA predictive model
Factivity
Fvariant
Fprocess
Predictors:
- process variant;
- activities order (trace);
- descriptive statistics for times and capacity;
- resource (team/operator);
- country of origin;
- cargo;
- berth.
Predictors:
- total time;
- season;
- workdays;
- primetime;
- weather conditions;
- delay time.
Time taken for a single case

Figure 3.11: Two step predictive model for context awareness

As an input we have three sets of features for activity, process and process variant. Then, first step is analyzing dependencies between features and define the flow time for a single case. After, second step brings time perspective from historical statistical data and calculate seasonality as well as trends in data. Depending on the target variable of the prediction we can use both models. The approach is able to predict along with times, information about next activity, resource, possible bottleneck or reworks on the trace.

It is important to notice that Process variant at first step includes its own parameter, obtained with process mining techniques and bottleneck detection method. $F_{variant}$ {activity order, bottleneck spots, probability, lead time, operations/hour, rework/loops}. In case of an operational support, the decision tree model can be applied to define the process variant based on activities order.

For instance, if we have 5 process variants: I = {abef}; II = {abdef}; III = {abcef}; IV = {abcdef}; and V = {abf}, then the decision tree to define the process variant at each operation will look like:

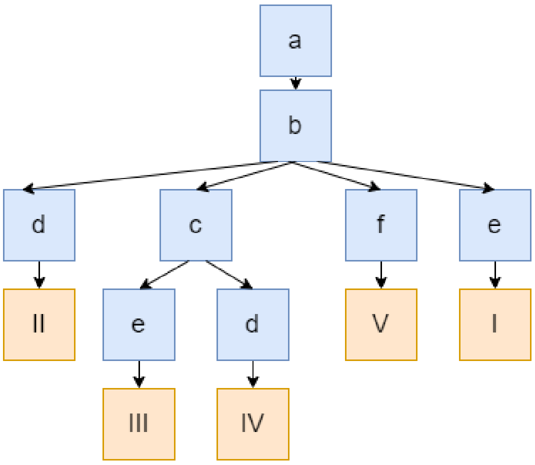


Figure 3.12: Defining process variant with decision tree model

As we can see in the Figure 3.12, there is no need to follow all traces, but just important decision nodes (b, c).

To evaluate the two-step predictive model we use MSE, which is calculated as:

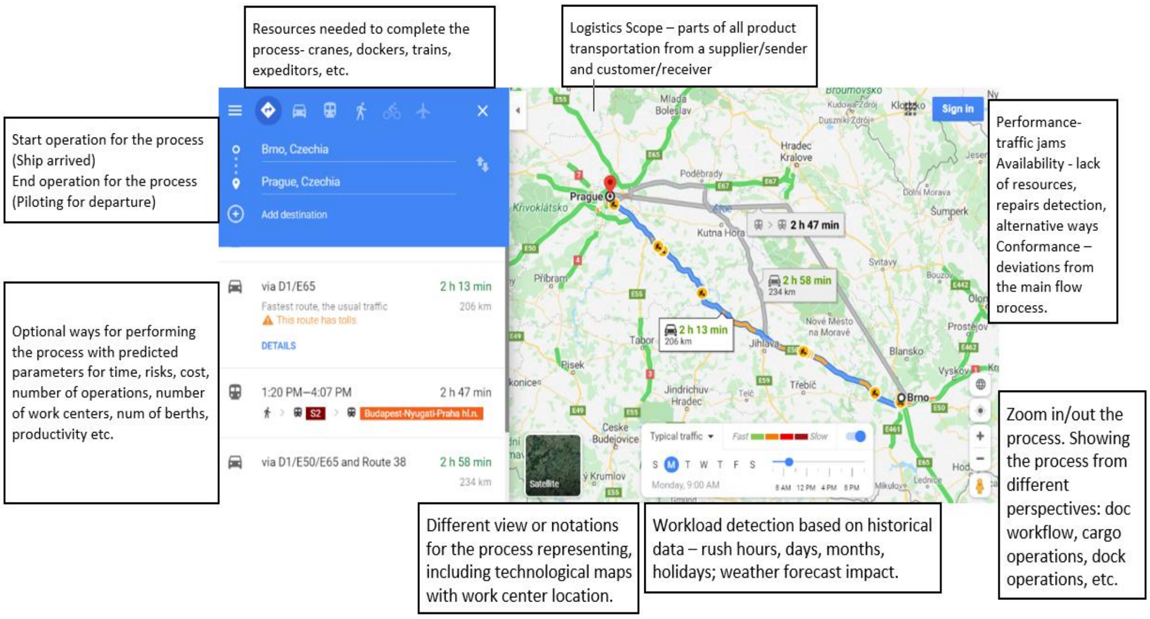
$$error(pred, S) = \sum_{s\epsilon S} |f_s^0 - pred(f_s^1, f_s^2, \ldots, f_s^n| / |S| \tag{3.4}$$

We try to enrich the predictive model with additional information to make predictions about the main performance parameters – risks, costs, times. The multiplicity, dynamism and interdependence of the process factors influencing performance parameters and do not allow the forecast to be created by traditional linear methods with a sufficient degree of reliability. The goal of predictive model is to minimize the error. In contrast with models, which were described in the Chapter 2, this model is aligned with the process model, support multicriteria, resistible for changes, forecast numeric and categorical target variables and is able to make long-term forecast.

3.6 Process Map

Eventually, the main goal of the framework is to obtain an interactive enriched process model with different perspectives in terms of a process map. *Process map* is the map that reflects a real process model (descriptive model) with the possibility of extension (scaling process), determination bottlenecks (traffic jam), detecting of deviations for operational response, representation of different perspectives (control-flow, resources, performance). Also, the map can be an alternative type of recommendation and decision-support systems. Parameters of process map can be used to create decision-making rules, define the procedure in a case of threats, to determinate the thresholds for main performance characteristics and to improve the quantitative and qualitative dimensions of the process.

The scheme of a process map is depicted accordingly to the existing navigation system – google maps – in the figure below. If we consider that all roads represent processes traces, traffic jams are bottlenecks, prediction parameters are estimated by performance characteristics for the situation, then the process map should look like in the Figure 3.13.



Resources needed to complete the process- cranes, dockers, trains, expeditors, etc.

Logistics Scope – parts of all product transportation from a supplier/sender and customer/receiver

Start operation for the process (Ship arrived)
End operation for the process (Piloting for departure)

Performance- traffic jams
Availability - lack of resources, repairs detection, alternative ways
Conformance – deviations from the main flow process.

Optional ways for performing the process with predicted parameters for time, risks, cost, number of operations, number of work centers, num of berths, productivity etc.

Zoom in/out the process. Showing the process from different perspectives: doc workflow, cargo operations, dock operations, etc.

Different view or notations for the process representing, including technological maps with work center location.

Workload detection based on historical data – rush hours, days, months, holidays; weather forecast impact.

Figure 3.13: Process map - navigation for logistics processes

The process map should include next parts:

- Process model with the simulation of cases going through the process and presented in different notations, which can be changed according to the user's knowledge. It can also be merged with technological maps and work center locations. This defines the focus of analysis and all other parameters.
- Different perspectives of the process: document flow, cargo flow, dock operations, resource interconnections defined by context awareness.
- Zoom in/out for the process to reveal hidden spots and discover the process in the detail.
- Input parameters – start and end operations to define the scope of the process and filter out unnecessary tracks.
- Resources needed to complete the process – operators, machines, cranes, trucks, etc…
- Workload detection based on historical data to define rush hours during the day, most overloaded with *traffic* days, months, and seasons. Also, weather forecast impact.
- Alternative ways to complete the process with recommendations for choosing resources and path, as well as predicted parameters for time, cost, risks/bottlenecks, number of operations, etc.
- Road service here can be considered as a resource, for example waiting in the warehouse or raid, when the process time flow is suspended.

The same time, process map should have an interactive mode with updated parameters for each case (operational support and online data stream):

- Each case should have speed (calculated performance) and estimated parameters till the process completion (time till the end, possible bottlenecks on the way, the estimated time when it the process might be completed, etc.).
- Trajectory – choice of the main trace or alternatives ones (if it is available, because not in each point of the process we can choose another option) with recommendation parameters about KPI, prediction of the next activity.
- Repairs on the process trace, weak/threat points, calculation of estimated flow time according to delay time and capacity on the road. These points also represent *complex off-*

road as well as troubles in the process (accidents, weather conditions).
- Calculated cost of the trace (optional).

The core of this map is already proposed in the main framework – the process model with zoom-in/out and different perspectives (framework to process *spaghetti* process models with context awareness), traffic jam and their causes (bottleneck detection method with process variants) and estimation of travel time (predictive models with additional calculated parameters from process mining explorative analysis).

The following recommendations can be carried out in future to maximize the functionality of the process map:

- Integrating results of the bottleneck and predictive model to the process mining workflow simulation, emulating the process animation.
- Adding more variants and specific process flows as well as resource analysis to the simulation model.
- Following the production process including manual resources (so-called hidden processes, information about which is not digitalized so far).
- Creating a heatmap (performance) during day/week/year to define *prime time* in the process.
- Discussing additional metrics, methods and results with process experts to approach the real situation and to consider cases, in which data are not reflected in the data sets.

Benefits of using process maps:

- Process maps make the process transparent for customers, stakeholders, authorities, and other participants of the process. Everybody can see the position of their part and impact of it in the global picture. Process transparency can improve our understanding and create the knowledgebase about the process's nature.
- Process maps helps in communication between different participants of the process, select important perspectives and focus on specific parameters.
- Performance analysis does not only detect bottlenecks within the process but also determines the causes of their occurrence. Predicted values help to make decisions based on algorithms and calculated information and be less human-centric. It helps to decrease waiting times and the intensity of the exploitation of available port resources and improve planning.

Summary of Chapter 3

In this chapter, we discussed and presented the core of the thesis – *methodology*. At the beginning, I proposed the framework to apply process mining to logistics processes. All steps are properly described and the Step 2 for data pre-processing shows efficient ways to deal with data issues. Then, the framework is extended with the method of dealing with complex unstructured process models – *spaghetti* models – by using context awareness. After filtering methods applying context awareness is promising approach to decrease variability of the processes. Once, we use a cluster of activities it helps to reduce numbers of process variants, that in turn leads to more structured final descriptive process model.

Here, we point two possible ways of applying context awareness in terms of process mining types – discovery and enhancement. $F_{process}$ and $F_{activities}$ were taken to deal with *spaghetti* models by building subprocesses with different perspectives and abstract levels (e.g. getting just cargo operations or building the subprocess for master data only). In turn, $F_{variant}$ impacts mostly enhancement type of process mining. Bottleneck analysis, from one side, and predictive analysis, on the other hand, are enforced with context-aware information, especially with these additional objective process attributes.

A new-found bottleneck detection method helps to highlight the risk/weak or threat places of the model. The method also is connected with a process model by taking a process variant as a trace for capacities (time or cases) detection. The advantage of method is that it is automated and oriented on the descriptive real process model with hidden spots. Confidential Interval method allows to detect bottlenecks in the process variants and define set of recommendations/reasons for each operation based on bottleneck direction. Knowing process trace, bottleneck spot and root direction (upstream, downstream) make it possible to create recommendations for bottleneck reasons. These bottleneck places as well as recommendations for them can be used for the predictive model as an extra attribute of process variant and impact the prediction of key parameters. Confidential interval is more robust and advanced method that TimeLag, it is also able to detect directions for bottlenecks. In comparison to Bottleneck walk methos, Confidential Interval method does not require manual work and all calculations can be extracted from event log and process variant variables.

In Step 7, there is another sight on the hybrid two step predictive model with additional information to make predictions about the main performance parameters – risks, costs, times. The multiplicity, dynamism and interdependence of the process factors influencing performance parameters and do not allow the forecast to be created by traditional linear methods with a sufficient degree of reliability. I propose to use the two-step predictive model: regression tree – cover the need to analyze case attributes, context awareness and features of the process, and help to find correlations and dependencies between features; and SARIMA model – brings seasonality, trends, time distribution, prime time analysis into the prediction. In contrast with models, which were described in the Chapter 2, this model is aligned with the process model, support multicriteria, resistible for changes, forecast numeric and categorical target variables and is able to make long-term forecast.

The last step is theoretically described in the work since the creation of the process map requires more collaboration with current projects and the involvement of process owners as well as software engineers. The scope of this thesis does not cover the development of the process map and suggests it for future research. The core of this map is already proposed in the main framework – the process model with zoom-in/out and different perspectives (framework to process *spaghetti* process models with context awareness), traffic jam and their causes (bottleneck detection method with process variants) and estimation of travel time (predictive models with additional calculated parameters from process mining explorative analysis). Also, the map can be an alternative type of recommendation and decision-support systems.

The framework is fully adjustable and can be easily applied to different kinds of processes. It provides a new ordered way of process mining application and brings a full understanding of the current process as well as the nature of logistics processes. Knowledge obtained through the framework execution can be further used for process optimization, better structuring of work, decision support, recommendations, and planning. Eventually, it is highly recommended to adjust all obtained results and models with data and process owners to avoid wrong interpretations. Working with the framework helps data and process analysts to get a key understanding of the whole process, ask meaningful questions and propose particular ways for improvements.

Chapter 4

Application – Case study

Since I have graduated from the maritime university and was working with international projects related to seaport processes optimization, the case for the work was taken in seaport logistics area, specifically a *ship-handling process in the harbor*. This choice helped to combine expert knowledge with PM analytical tools withing data processing and process modelling steps.

Ports are considered as potential logistics centers and important transport hub in supply chain [65]. Logistics intelligence covers the set of techniques that seek to improve logistical operations with their abilities to reduce the uncertainties and risks in logistics. Being aware of the actual logistic processes and deviations from the planned processes in a dynamic environment is essential for companies in order to gain additional flexibility and improve performance characteristics. The domain is complex and multivariable, involving a lot of different process participants. Even if there is a persuasion that logistics processes are well-organized and structured, the reality shows that it is highly human-centric, partially unknown process, not transparent process. Most of decisions, which can imply the process and its performance parameters globally, are made manually. Advanced research using descriptive models is needed to extend and improve the behavioral validity of the port system analysis, making it amenable for real-life applications

Logistics process was chosen for an application case of this work due to recent preconditions, which make it possible to apply and examine prominent algorithms of process mining in a new field. Moreover, there are just a few works related to a comprehensive methodology for applying process mining logistics to provide intelligence support for logisticians. And as we know, process mining techniques require to be applied to new process areas to get better knowledgebase.

Throughout this chapter, we focus on the loading/reloading process, which can be called the *ship-handling process*. In this research, we are extracting data from timesheets to create the process model. I present the way of forming an event log for logistics processes and context-aware information to handle complex process models. Also, we apply the proposed framework from the third chapter and see how it impacts the results. Eventually, we compare predictive models which we used for the dataset and make summary.

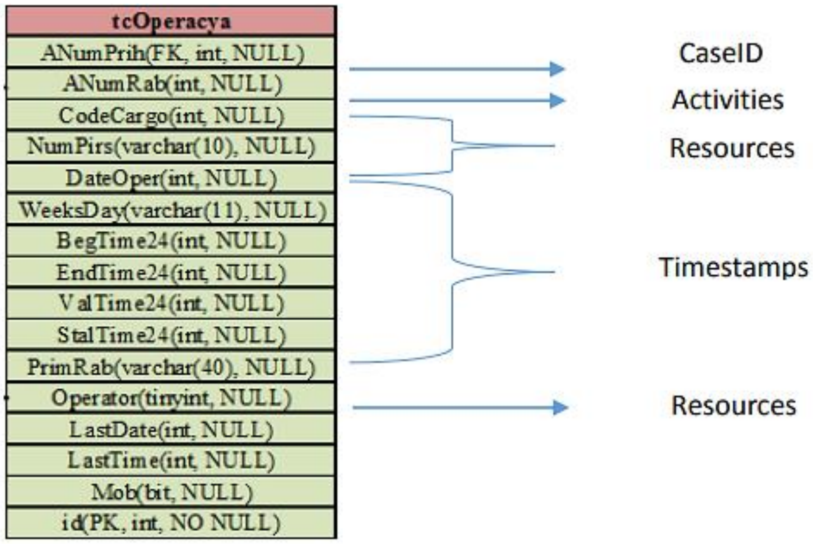
4.1 Object description

For the application of the framework from previous chapter to logistics, I obtain the collaboration of a seaport – Europe’s third largest port operator in terms of cargo turnover. The cargo turnover totals of the seaport 105.2 million tons per year, 84 million of them falls to the share of oil terminal. The port deals with both containers, liquid and bulk cargoes including oil, grain, steel, coat, chemical fertilizer, ore, automobiles, and so on. I focus on *oil terminal*. The core business for the oil terminals concerns the inward and the outgoing oil handling. During the inward handling process, liquid cargo transported to the harbor by ships, and then discharged to the port storage area or carried away from the harbor by trucks, vessels, or trains. During the outgoing cargo handling process, liquid cargo transported from inland are put in the storage area of the port and then carried away by ships. Both the inward and outgoing cargo handling process contain various types of logistics activities [14].

Thus, the data used in the work for process modelling is seaport the loading/unloading part of the logistics process, specifically ship-handling process at the oil terminal for the port.

4.2 Data preprocessing

The analyzed data were obtained from timesheets documents (see Appendix C1) that comprise necessary information about full ship handling operations at an oil terminal for subsequent financial settlement between ship and a port. Information from this document is digitalized and stored in databases as it is presented in Figure 4.1.



tcOperacya
ANumPrih(FK, int, NULL)
ANumRab(int, NULL)
CodeCargo(int, NULL)
NumPirs(varchar(10), NULL)
DateOper(int, NULL)
WeeksDay(varchar(11), NULL)
BegTime24(int, NULL)
EndTime24(int, NULL)
ValTime24(int, NULL)
StalTime24(int, NULL)
PrimRab(varchar(40), NULL)
Operator(tinyint, NULL)
LastDate(int, NULL)
LastTime(int, NULL)
Mob(bit, NULL)
id(PK, int, NO NULL)

Figure 4.1: The source for dataset stored in the database. There is a snapshot from database on the left side with description in terms of PM parameters on the right side

According to the framework from the Chapter 3, the data was pre-processed in the following way:

- The timeframe is calculated by Formula 3.1. The Figure 3.3 shows that usually time duration for most cases has interval [149.5h; 167.5h]. So, we take max expected time and multiply it by 4 and 5. Timeframe equals 3340hours = 139days = 5months. This timeframe is enough for a process modelling. Since we want to apply also predictive model with time series analysis, we can take one year to see impact of different seasons.
- Ship cases with start activity *Ship arrival* and end activity *Pilotage for leaving* were defined as a process with a normal behavior. Part of processes started with the operation *Nor-tendered*, which by words of process owner cannot be before *Ship arrival*. Such cases can be analyzed separately or filtered out.
- All ship cases that have duration time not belonging to the confident interval {the median value ± standard deviation} = [23.5h; 257.5h] were referred to data as a noise and outliers. In this way, incompleteness and partially noise problems were solved.
- Raw data was obtained from different information systems and collected to database. Documents do not have a unified standard for activities names. For instance, two similar operations *Weather conditions* and *Weather conditions (Other)* (or *NOR tendering-berth* and *NOR tendering-road*) have two different identifications. This problem was solved by aggregation of operations according to experts' knowledge and new parameter as *Method* was added.
- Duplicates with different timestamps within a case were merged. For example, we found two the same operations in one case with start and end times (1:00; 2:00) and (1:00; 3:00) respectively. It was solved by using the common time frame that covers both times (for our example, the operation has time (1:00; 3:00)).

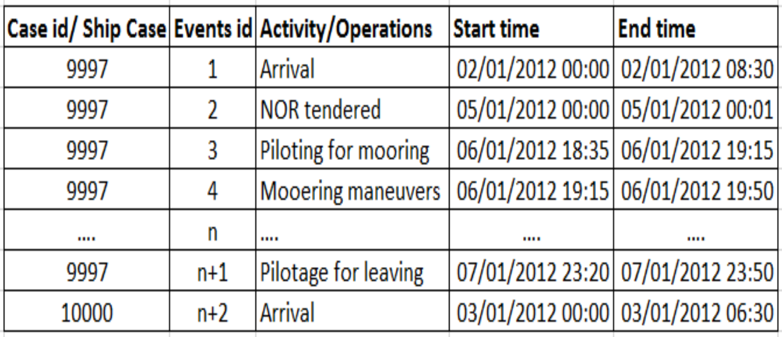
- Also, the specific of the system is to collect times in standard format. So all-time values were converted to certain common format.

Event log.
An event log usually contains four important parameters:

- case id,
- activities,
- and timestamps (start and end time for the activity).

For further analysis, an event log can be enriched with an information about process and activity attributes such as costs, operators, ship sizes, weather conditions, etc. In our case, an event log (see Table 4.1) is created based on the information from port documents – timesheets/statements.

Table 4.1: Event log sample extracted from preprocessed data of timesheets



Case id/ Ship Case	Events id	Activity/Operations	Start time	End time
9997	1	Arrival	02/01/2012 00:00	02/01/2012 08:30
9997	2	NOR tendered	05/01/2012 00:00	05/01/2012 00:01
9997	3	Piloting for mooring	06/01/2012 18:35	06/01/2012 19:15
9997	4	Mooering maneuvers	06/01/2012 19:15	06/01/2012 19:50
....	n	....	....	....
9997	n+1	Pilotage for leaving	07/01/2012 23:20	07/01/2012 23:50
10000	n+2	Arrival	03/01/2012 00:00	03/01/2012 06:30

The documents gave us a detailed chronological description of the activities of the vessel during the stay in a port: arriving, taking the sea pilot on-board, hailing in, mooring, preparing for the loading and unloading operations, the actual loading and unloading operations, unmooring, departure. For the ship-handling process, every ship case at the terminal is referred to a case. While the operations are activities and timestamps are start and end times. Process mining techniques simulate process models, extracting necessary information from event logs. For ship-handling process, every ship case at the terminal is case ID, operations, which were made with the ship, are activities, and timestamps are start and end times of each operation. For further analysis, an event log will be enriched with an information about resources, a berth number, ship sizes, weather conditions and so forth. Process mining techniques as well as bottleneck and predictive methods can be applied.

4.3 Process flow modelling

To provide insights into the process structure, a process models will be presented further. The event log is composed of 2659 ship cases, 62245 events and 135 different operations/activities for an year. I took data for one year to discover the process model. Eventually, there were 324 cases and 6,405 events which means each ship case has on average approximately 20 events. We chose the Fuzzy miner as a suitable algorithm for discovering the logistics process model. The notation for model visualization is DFG – an oriented graph where each rectangle refers to the activity, arches show the relations between them, start and end of the process are defined accordingly. Moreover, nodes of the diagram detect frequency of operations with colors: grey nodes are less frequent operations and blue ones are the most frequent.

Since the process is complex, we get a *spaghetti* process model at the first run (it is not presented here but looks like in the Figure 2.15. To show the framework concept of working with context awareness we compare different views of the process models at once. The figure 4.2 shows the sample of the process model visualizations and their changes. In Appendixes A and F, process models are presented in detail. Disco tool was used to build presented process models. The Figure 4.2 presents the models which were obtained based on different inputs. The process model on the right has more details and information about frequent and rare activities. To avoid unreadable view, algorithm and visualization adjustments were used. The process model in the middle shows the scheme after data-preprocessing step and visualization adjustments. Here we can see the cluster of cargo activities. At this point, we can apply $F_{activity}$ = *laytime*. According to the process explorational step, just cargo operations have this additional parameter. In this way, we can build the subprocess model just for cargo operations. The right process model presents cargo operations and starts with the operation *Connecting cargo loading arm* and finishes with *Signing cargo documentation*.

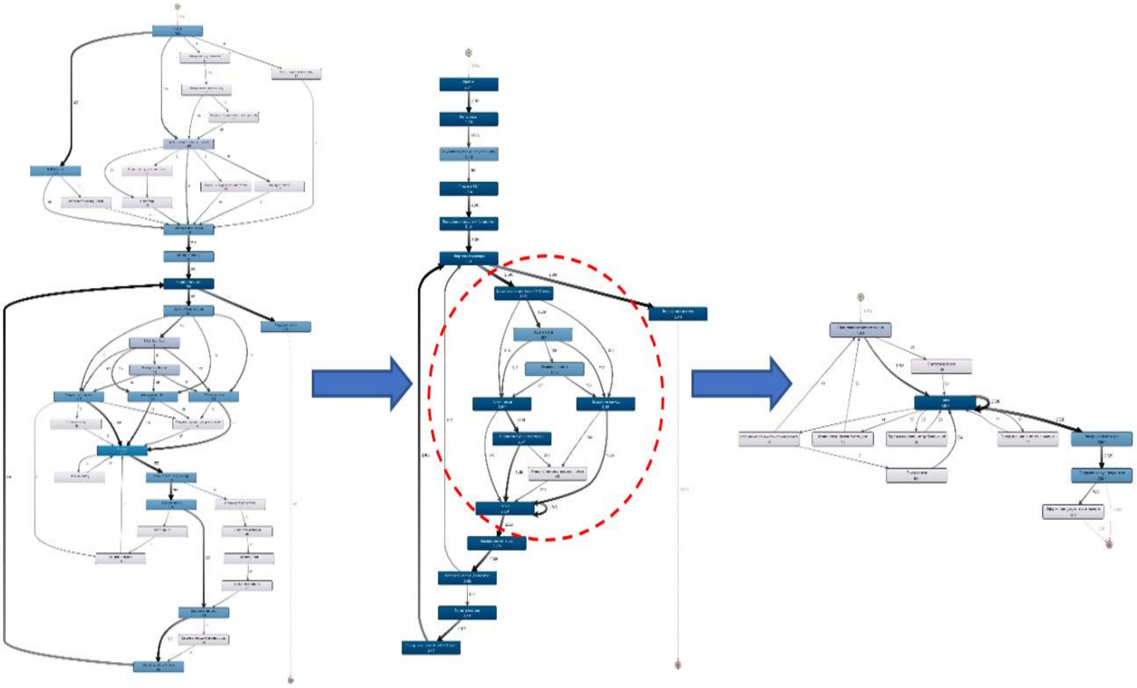


Figure 4.2: Process model comparison: left - model with algorithms and visualization settings, middle – with data preprocessing steps plus visualization settings; right – obtained subprocess with applying context-awareness.

Using the right process model from Figure 4.2 for cargo operations, the most frequent process variants for it can be defined. Since, following calculations are based on real data and will include some performance characteristics, I will replace operations' names with codes like A012800 and so on.

Process variants are presented in the Figure 4.3. We can see variability in the activities order. Also, some variants seem to present incomplete cargo processes. I made preprocessing for ship handling data, but cargo operations were recognized automatically. Some variants are named since they will be used in next steps of the framework.

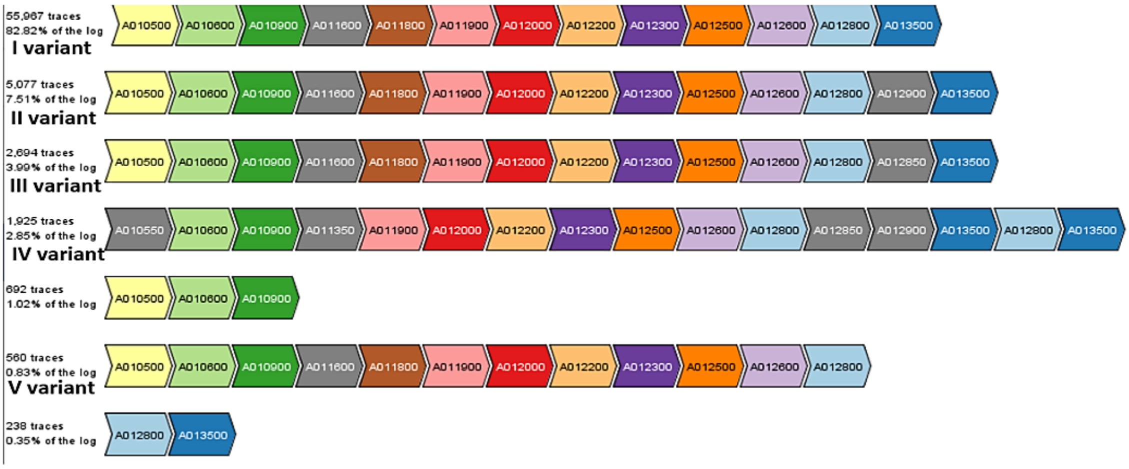

55,967 traces
82.82% of the log
I variant
A010500 A010600 A010900 A011600 A011800 A011900 A012000 A012200 A012300 A012500 A012600 A012800 A013500
5,077 traces
7.51% of the log
II variant
A010500 A010600 A010900 A011600 A011800 A011900 A012000 A012200 A012300 A012500 A012600 A012800 A012900 A013500
2,694 traces
3.99% of the log
III variant
A010500 A010600 A010900 A011600 A011800 A011900 A012000 A012200 A012300 A012500 A012600 A012800 A012850 A013500
1,925 traces
2.85% of the log
IV variant
A010550 A010600 A010900 A011350 A011900 A012000 A012200 A012300 A012500 A012600 A012800 A012850 A012900 A013500 A012800 A013500
692 traces
1.02% of the log
A010500 A010600 A010900
560 traces
0.83% of the log
V variant
A010500 A010600 A010900 A011600 A011800 A011900 A012000 A012200 A012300 A012500 A012600 A012800
238 traces
0.35% of the log
A012800 A013500

Figure 4.3: Process variants for cargo handling process model (was created with PROM tool)

Once, we explore process variants, we can use the most important of them, which cover higher number of all cases, to build last level of the model for certain process variants. The Figure 4.4 shows the process model for variants I-III. The Figure 4.4 represents the visualization and activity names are not important in this case. However, the loop in the process model is visible on all three notations.

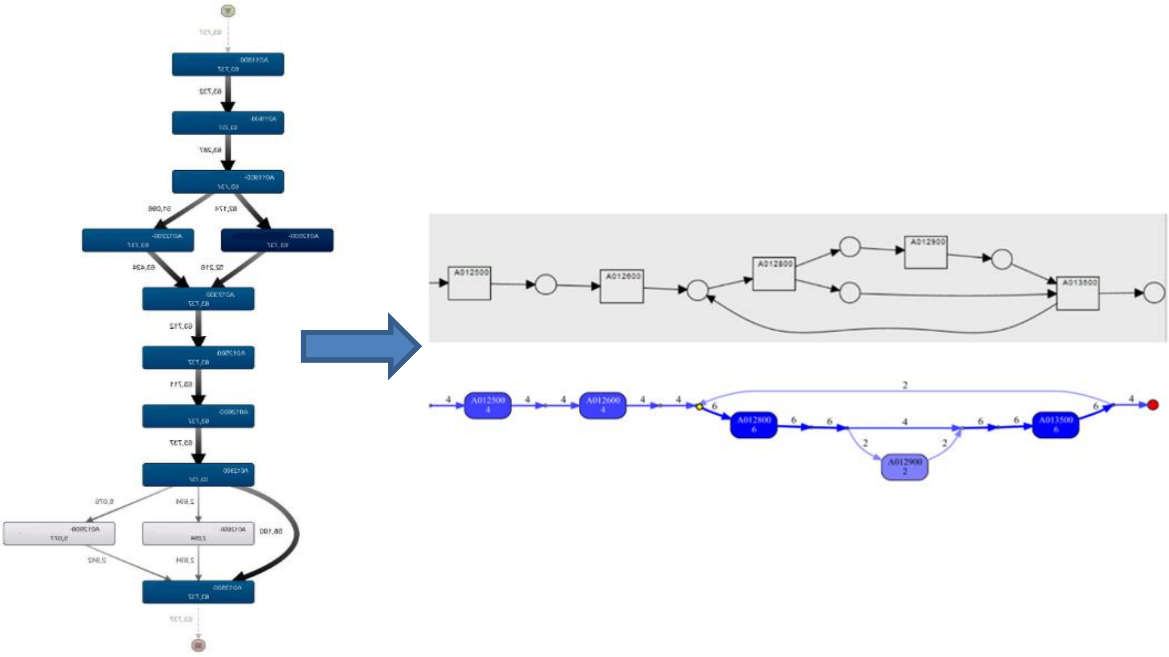


Figure 4.4: The process model built for specific process variants and its representation in different notations – DFG, petri net, inductive miner (left to right)

Choosing process variants as an input for process modelling, we simplify the visualization. Thus, it makes it possible to explore model with another notations like petri nets or inductive miner to use advantages of these notations as well and to avoid complexity. It is necessary to build and check the process model in various notations to avoid the limitlessness of each algorithm.

4.4 Process attribute analysis – process variants

After first three steps of the framework were completed and the process model was built, process variants better can be considered in detail. To visualize process variants, I use the Figure 4.5.

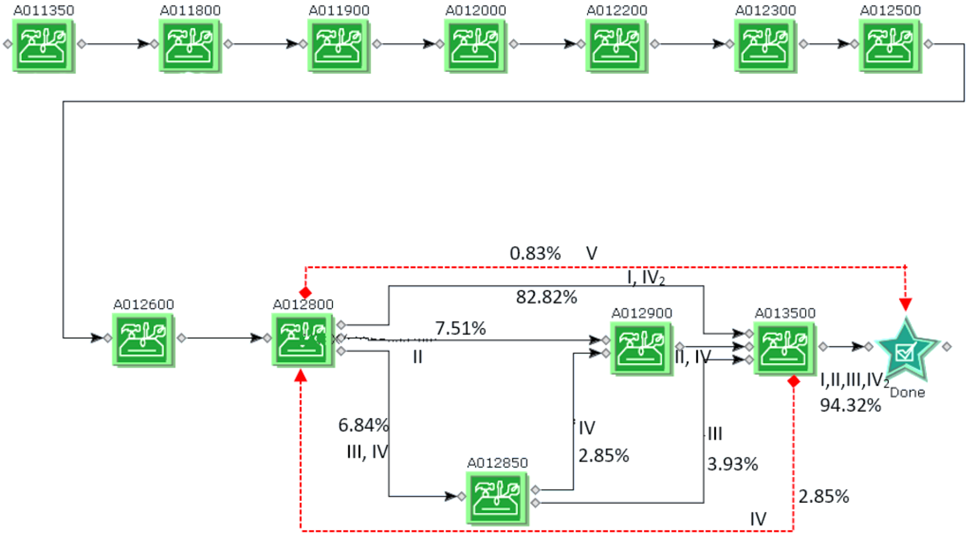


Figure 4.5: Visualization of process variants on the normative process model with new-observed paths from the descriptive model

The Figure 4.5 shows a normative process model for the cargo handling process. The descriptive model discovered by process mining reveals also two additional connections in the process as well as the loop. These relations are shown as dotted red arrows and can indicate specific cases or temporal changes in the processes as well as hidden unknown relations. The first decision point we see in the picture is in the operation A012800. There are four outputs and one input. Based on product configuration or other attributes of the process the further route is defined at this point. The roman numbers (I, II, etc.) represent the type of the process or process variants (see the Figure 4.5). To process loops and to keep correct calculations we should differ a number of *transit* throughout the operation. In our case, the loop is observed in variant IV. Therefore, the second run through the process adds to the process variant the index 2 – IV$_2$.

Each connection has an assigned probability which was taken from variants of historical data and Explorative analysis step. Exploring all possible process variants that happened in recent times gives the distribution for activities' relations. It can be higly beneficial in terms of the concept drift (sudden changes in the process behaviour caused by economic crisis, epedemics and so on). Once, changes are noticed, an updated explorative analysis for variants of the process can be repeated and quickly adopted to the new process variants with their attributes. Thus, based on the Figure 4.5 we can make the next assumptions:

- Operation A013500 happens after A012800 in 82.82 % of cases – Variant I.
- Operation A012900 happens after A012800 in 7.51 % of cases – Variant II.
- Operation A012850 happens after A012800 in 6.84 % of cases – Variant III + Variant I.
- Operation A013500 happens after A012800 in 82,82 % of cases – Variant I. Variant IV is not calculated in this probability, because it belongs to the second round and should be counted with another loop accordingly.

Similar conclusions can be done for other decision points in the process – the mode where we have more than one output - A012850, A013500. Probabilities help to predict the next activity in the process. Whereas each process variant has not just frequency distribution, but also various quantitative and qualitative parameters.

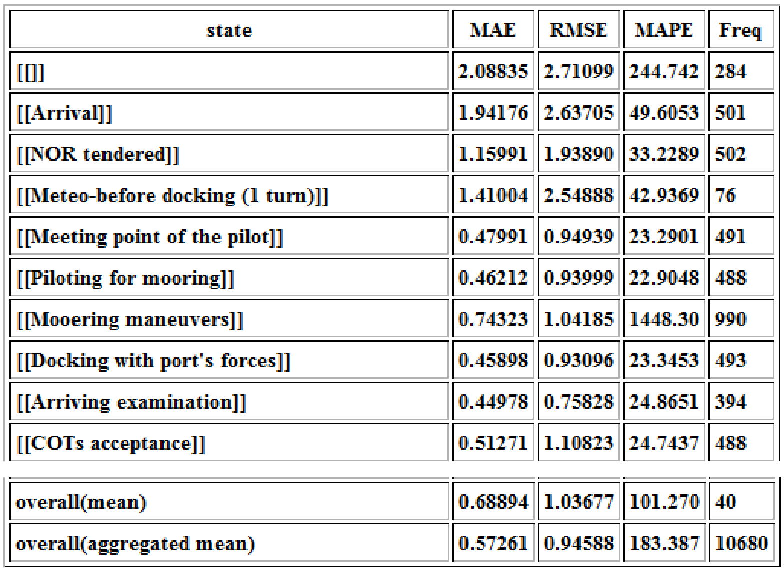
Bottleneck analysis has been described in the Chapter 3 and the used example was calculated based on the data set from the application. The Figure 3.10 shows the main process variant I and all these time parameters. The biggest bottleneck was found in EQT (time in queue) before the operation A012800. It can be taken as an additional feature of the process variant I as well as the recommendations from the Table 3.6. It is necessary to add, that finding bottleneck for one case represents the same bottlenecks for batch of cases depending on buffer size of the process line. If process line processes 14 units, then all of them will have the similar time to be completed and delays in the different parts of the process according to capacities (items in queue) between parts. However, roots of bottleneck for all these units are the same.

4.5 Predictive models

In this work, we try to enrich this model with additional information about context awareness to make predictions about one of the main performance parameters – *ship handling duration*. The multiplicity, dynamism and interdependence of the factors influencing this parameter, does not allow the forecast to be created by traditional methods with a sufficient degree of reliability. First, I describe parameters which are used by applying and building transition systems and time series models.

Transition systems

I used two data sets for construction and evaluation of the transition system. The training data set L1 contains all ship cases that were handled at oil terminal within first half of the year. L1 holds 303 cases, 12480 events and 68 different operations/activities. The test data set L2 has 236 cases and 76 operations/activities, occurred within first half of another year. The main operations for the ship handling process are common for both data sets, but each has its own operations relevant to ship cases specific. Eventually, the transition system was obtained using the set abstraction based on all activities. With the plugin FSM analyzer the TS was also reinforced with an information about times for every operation. The goal is to predict, at any point of time, remaining handling time of a case. The transition system and the annotated TS extracted from event log L1 are not intended to be readable (also with ignoring unnecessary details). For this reason, result is presented with certain parts of process – initial and finish states (see the Figure 2.34). For the initial state *[[]]*, the predicted remaining time until completion 4,82 days. The first activity is always *Arrival* and the second activity is *NOR tendered*. After these two steps, the process is less structured (see the Appendix A). The resulted transition system with states is presented in the Appendix D. To evaluate the quality of the predictions we use data set L_2 and FSM Evaluator [64]. Results are presented in the Figure 4.6.



state	MAE	RMSE	MAPE	Freq
[[]]	2.08835	2.71099	244.742	284
[[Arrival]]	1.94176	2.63705	49.6053	501
[[NOR tendered]]	1.15991	1.93890	33.2289	502
[[Meteo-before docking (1 turn)]]	1.41004	2.54888	42.9369	76
[[Meeting point of the pilot]]	0.47991	0.94939	23.2901	491
[[Piloting for mooring]]	0.46212	0.93999	22.9048	488
[[Mooering maneuvers]]	0.74323	1.04185	1448.30	990
[[Docking with port's forces]]	0.45898	0.93096	23.3453	493
[[Arriving examination]]	0.44978	0.75828	24.8651	394
[[COTs acceptance]]	0.51271	1.10823	24.7437	488
overall(mean)	0.68894	1.03677	101.270	40
overall(aggregated mean)	0.57261	0.94588	183.387	10680

Figure 4.6: The results of prediction for transition model. Calculations and the snapshot were made in PROM tool [64]

Applying statistical methods

For time series analysis I use the data set which has 2121 instances and two attributes: Started point of ship case handling and Duration (h) of *full handling* (not just cargo operations). It includes period from 2012 to 2015 years. For days, when there were some cases, the mean value was given. For example, if at 2.1.2012 four ship cases were handling and their total duration was 542 hours, then the value of this day equals 542/4=135.5 hours. The problem with missed data is also occurred. The missing values were replaced by the mean of their neighboring values. Thus, the frequency for timeseries analysis can be adjusted as daily, weekly or monthly. The one case takes an average of 83 hours (more than 3 days), so it is logical to determine weekly frequency. So, although the daily one can provide more detailed model, it was decided to compare results of models, based on time series with weekly frequency (see the Figure 2.29 TSw).

Three methods (single, double and triple exponential smoothing) are built for our TSw and are depicted on the Figures 4.7 and 4.8. The Figure 4.7 shows the visualization of different coefficients, so the most suitable coefficient to make the line close to the actual values is 0.3 for *single exponential smoothing* and (0.9;0.9) for *double exponential smoothing*.

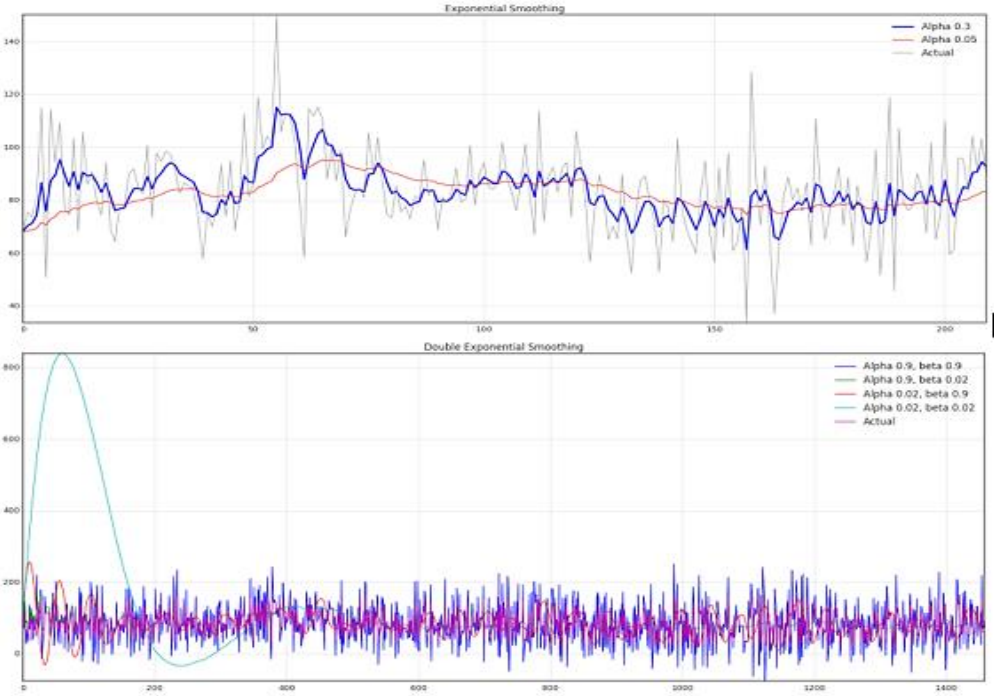


Figure 4.7: *Single and double* exponential smoothing [71]

Holt-Winters model: the best-found coefficients for this model are: α=0.0066, β=0.0, γ=0.0467. The Figure 4.8 demonstrates results of forecasting with confidential interval, with the Brutlag's algorithm of anomalies detection. Training and test data are separated by red line.

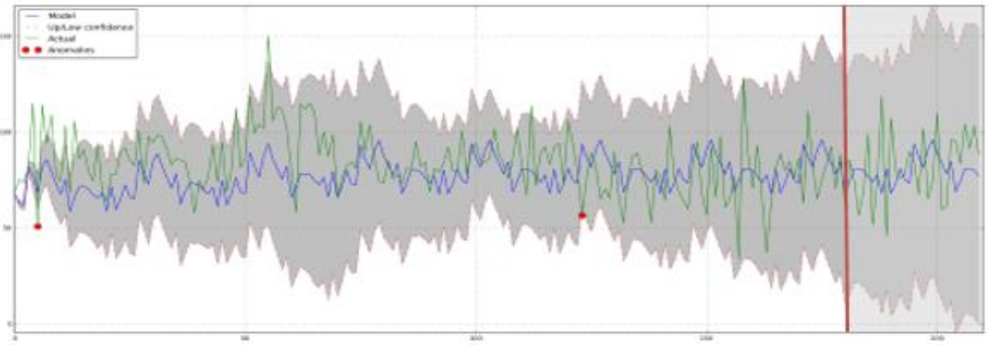


Figure 4.8: Triple exponential smoothing or Holt-Winters model [71]

ARMA

Finding appropriate values of p and q in the ARMA (p, q) model can be facilitated by using ACF and PACF plots shown in the Figure 2.31. Thus, from the Figure 2.31 it can be observed that both functions tend have zero values after lag 3. It means parameters p=q=3 and the model will look like ARMA(3,3) or ARIMA(3,0,3). The Figure 4.9 shows the actual time series, which is indicated by a blue line and forecast by a red dot line with the ARMA predictive model.

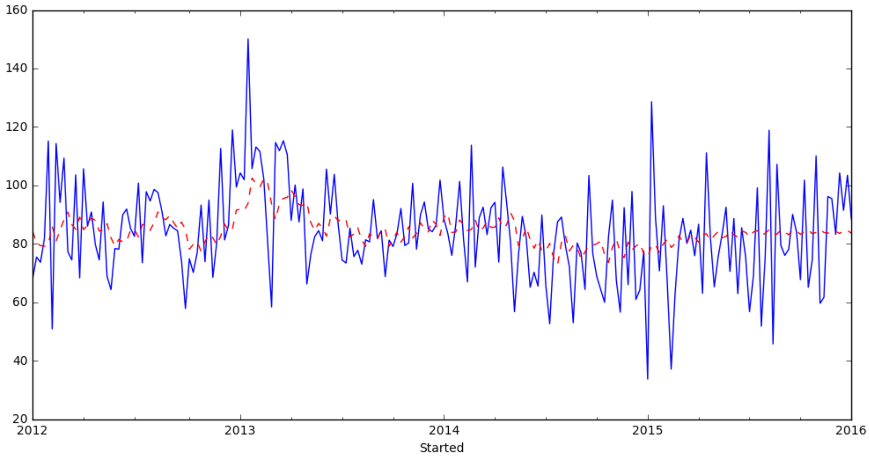


Figure 4.9: ARMA (3,3) predictive model [71]

SARIMA

To choose parameters for this model, additional investigations should be conducted. The grid search method might be used for searching of the best parameters. The frequency is monthly for the time series so per year s=12. To reduce the number of calculations (p,d,q) remains the same as in ARMA, i.e. (3,0,3). The best parameters are reported as SARIMA $(3,1,3)(1,1,0)_{12}$ and are corresponding to the lowest MSE value and to quite low AIC value. Also, the model SARIMA $(3,0,3)(1,1,3)_{12}$ has a good results and it will be added to comparison table (see the Table 4.2). It is necessary to note, that the best parameters cannot be found based just on AIC. There were iterations with fairly low AIC, but model could not be built. Moreover, it is important to run the model results diagnostics to assure that none of the assumptions made by the model was violated. The Figure 4.10 allows to investigate any unusual behavior, which in our case is not.

Forecasts from the ARIMA$(3,1,3)(1,1,0)_{12}$ model (which has the lowest MSE value on the test set) are shown in the Figure 4.11.

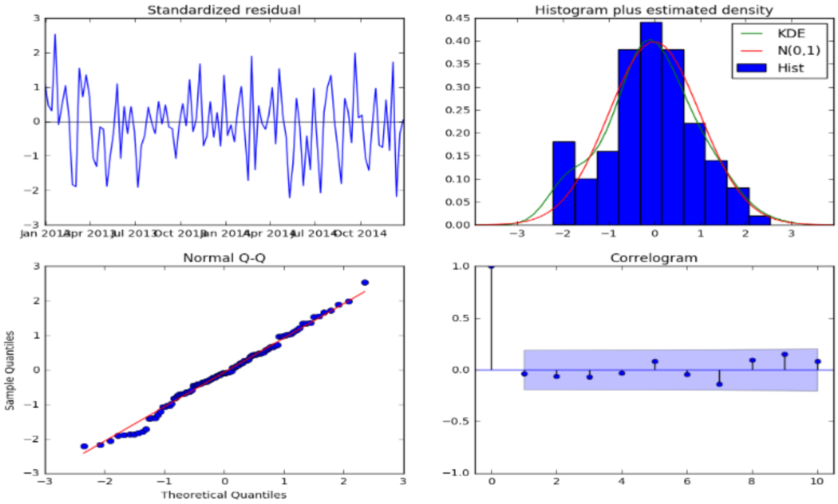


Figure 4.10: Diagnostics of predictive model results [71]

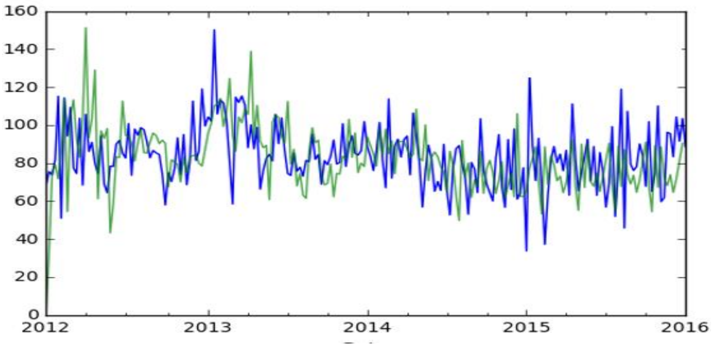


Figure 4.11: ARIMA(3,1,3)(1,1,0)$_{12}$ predictive model

For all the models, Mean Square Error (MSE) is calculated on the training and test data. The best results were achieved by ARMA model with weekly dataset frequency. Nevertheless, there are some other ways to improve prediction models and to obtain more accurate results, which are proposed in the framework.

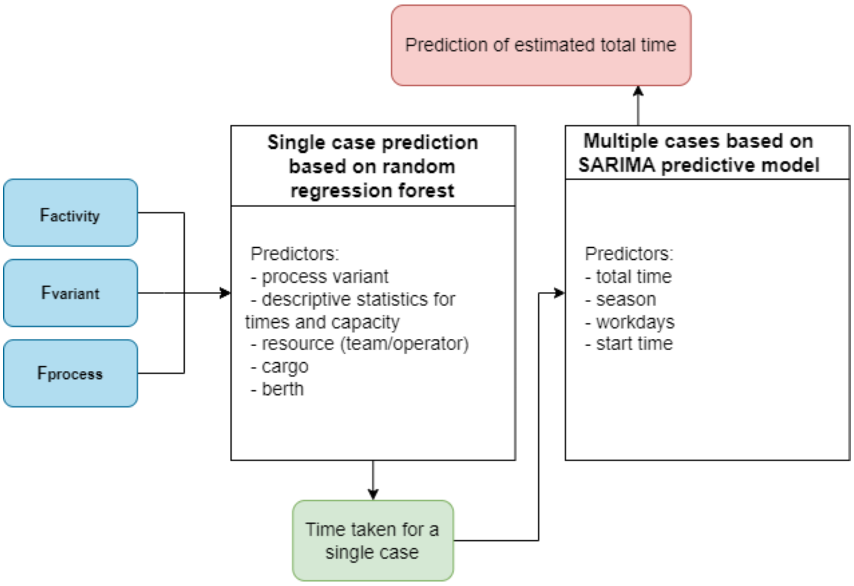
To compare results of all developed models, comparative table is presented in the Table 4.2. The most accurate model is ARMA(3,3). Although, it shows the lowest MSE with train data as well as with test data, it still needs furthermore investigation. To check whether this model can be further improved, other forecast errors should be examined.

Table 4.2: Evaluation of statistical predictive models

Model	MSE	
	Train data	Test data
Triple smoothing/ Holt-Winters (α=0.0066, β=0.0, γ=0.0467)	299.24	387.35
ARMA (3,3)	211.34	350.16
SARIMA (313) (110)$_{12}$	392.01	422.3
SARIMA (303) (113)$_{12}$	399.01	496.82

Two-step predictive model

Using the predictive model from the framework, prediction can be connected to the process variants attributes by first step- random regression forest (see Figure 4.12).


Prediction of estimated total time

Single case prediction based on random regression forest

Predictors:
- process variant
- descriptive statistics for times and capacity
- resource (team/operator)
- cargo
- berth

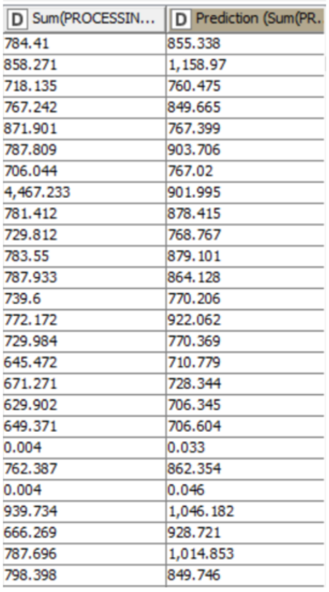
Multiple cases based on SARIMA predictive model

Predictors:
- total time
- season
- workdays
- start time

Time taken for a single case

Figure 4.12: Two-step predictive model

As the predictors for the first step, I use process variants number, descriptive statistics for times (flow time, waiting time), the cargo type and the berth number. Second step: coefficients for statistical predictive model are taken from the section above (SARIMA (313) $(110)_{12}$). The predictors for second step are total time, start time, season and workdays. Since the goal for the use case is to predict total time, I do not use here information about bottlenecks, activities order and so on. The results of the two-step predictive model are presented in the Figure 4.13. We can see that the graph of predicted times is close to the actual values. MSE for this predictive model equals 152.63, and RMSE =12.3.



D Sum(PROCESSIN...	D Prediction (Sum(PR..
784.41	855.338
858.271	1,158.97
718.135	760.475
767.242	849.665
871.901	767.399
787.809	903.706
706.044	767.02
4,467.233	901.995
781.412	878.415
729.812	768.767
783.55	879.101
787.933	864.128
739.6	770.206
772.172	922.062
729.984	770.369
645.472	710.779
671.271	728.344
629.902	706.345
649.371	706.604
0.004	0.033
762.387	862.354
0.004	0.046
939.734	1,046.182
666.269	928.721
787.696	1,014.853
798.398	849.746

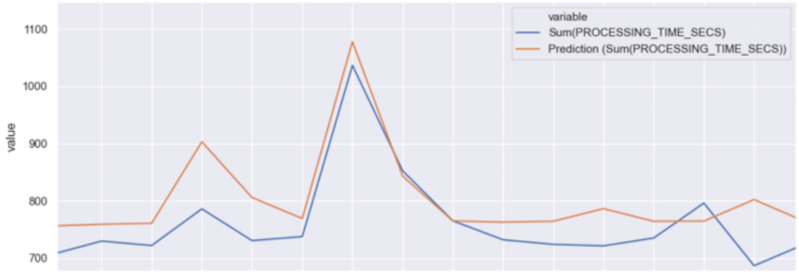


Figure 4.13: Evaluation of two step predictive model – the sample of test dataset and predicted values (on the left, made in KNIME tool), the visualization for predictive model, where blue line is the actual values and red line is predicted values (on the right, made by python visualization) [66]

Thus, based on all calculations we can admit that transition system showed better result for each activity, but for whole process with different variants the two-step model gives better results. The two-step predictive model was also applied to manufacturing processes and the detail description of the implementation can be found in the work [66].

Summary of Chapter 4

Since I have graduated from the maritime university and was working with international projects related to seaport processes optimization, the case for the work was taken in seaport logistics area, specifically a *ship-handling process in the harbor*. This choice helped to combine expert knowledge with PM analytical tools withing data processing and process modelling steps.

In a turn, the reasons for applying process mining in terms of logistics processes include the following: reducing the idle time of ship handling, improving of key performance indicators, maximizing the employment of port resources, etc. There are usually basic patterns of work and people from industry are not willing to waste their time without previously known results. Analyzed process models strongly depend on different scenarios in logistics. In case of transport logistics, we can define different cases such as ships, warehouses, containers, documents, forwarders, terminals. For my study, the ship handling process was chosen to present possibilities of applying process mining techniques. The ship handling process at an oil terminal is supposed to be clear and straight. However, it remains a black box for analysts. There are many simulations and normative models to understand what the ship-handling process is, but most of them do not focus on reflecting the real/descriptive model. Using algorithms of process mining and data from timesheets, the real process model was constructed. The real process model can be used:

- for audit (comparing normative and real models, finding deviations),
- training new workers,
- creating ship-scheduling,
- prediction of threats,
- detection of bottlenecks and their causes,
- and also, for recommendation of countermeasures.

The real descriptive model is the basis for valuable types of analysis like conformance checking (comparing two models – descriptive and normative, or the real model with event log for a new month to find deviations), streaming (monitor the production process in real time), performance analysis (bottlenecks, resource analysis), recommendation and decision support system (decision rules for KPI – key performance indicators). Each node in the process can be enriched with required data like resources, risks, probabilities, productivity characteristics.

In this chapter, the framework which combine process mining, statistics, machine learning techniques with context awareness is applied to logistics processes. I explored the part of the seaport process for loading/unloading operations – ship handling process.

After main data pre-processing steps, I could create event logs with different features and context awareness to obtain the process model and activity related statistic information. Using the event log, I built process models and optimized the visualization by adding context awareness. The Figure 4.2 shows a visible change in the process model visualization, what is valuable to work with *spaghetti* models.

The bottleneck detection is a crucial part of the continuous improvement of processes. Using the TimeLag method and the confidence interval method, the workflow will identify the bottlenecks from the dataset and add the root cause for the bottlenecks based on the operation and delay incurred by downstream or upstream process. The biggest bottleneck was found in EQT (time in queue) before the operation A012800. It can be taken as an additional feature of the process variant I as well as the recommendations. It is necessary to add, that finding bottleneck for one case represents the same bottlenecks for batch of cases depending on buffer size of the process line. Underlying process model makes the detection more transparent, accurate and valuable.

Last part of the framework – prediction model – was applied to the case. Based on all calculations we can admit that transition system showed better result for each activity, but for whole process with different variants the invented two-step model gives better results. The two-step predictive model was also applied to manufacturing processes and the detail description of the implementation can be found in the work.

Chapter 5

Conclusion and future work

5.1 Conclusions

There is an increasing interest in logistics process modelling research and new directions for modeling are developing quickly. A key challenge is to populate the modelling frameworks with the descriptive models of logistics decision making behavior. The processes are human-centric, diverse, flexible, and complex. That leads to obtaining unstructured unreadable process models called *spaghetti*. Thus, process remains to be unknown black-box and has no influence on further analysis.

The goal of my research was to extract knowledge and models from real-world raw event logs from interconnected logistics processes and to identify problems to further improve the processes. I propose an approach for modelling hidden and unknown processes and subprocesses in the example of a seaport logistics area. Having the underlying process model makes it possible to exploit more advanced algorithms since deviations and main paths are becoming visible and better controlled. The obtained model is the foundation for the core research of this work and will be enriched with key performing indicators and their forecast by applying advanced process mining, statistics, and machine learning techniques. Combining these powerful tools expands the possibilities of the research and empowers results of each other. The goal was achieved and applied to the real data for ship-handling processes, which is the main part of logistics. The model was built by applying fuzzy mining and DFG notation.

The main difference of the approach is that we take as a target variable not any specific value, but the object – a process variant or a process type with a set of parameters. $F_{process}$ and $F_{activities}$ were taken to deal with *spaghetti* models by building subprocesses with different perspectives and abstract levels. (e.g. getting just cargo operations or building the subprocess for master data only). In turn, $F_{variant}$ impacts mostly enhancement type of process mining. Bottleneck analysis, from one side, and predictive analysis, on the other hand, are enforced with context-aware information, especially with these additional objective process attributes.

A new-found bottleneck detection method helps to highlight the risk/weak or threat places of the model. The method also is connected with a process model by taking a process variant as a trace for capacities (time or cases) detection. Knowing process trace, bottleneck spot and root direction (upstream, downstream) make it possible to create recommendations for bottleneck reasons. These bottleneck places as well as recommendations for them can be used for the predictive model as an extra attribute of process variant and impact the prediction of key parameters. The advantage of method is that it is automated and oriented on the descriptive real process model with hidden spots.

Furthermore, by closely working with domain experts, the hybrid predictive model based on machine learning and statistics was developed for the prediction of time flow. Among input predictors we also use context awareness – process variant with its own attributes. Depending on requirements, predictive model can be applied to forecast resource needs, activities, operations, and potential delays, and deviations. This predictive model shows better results in comparison to other models, since it has strong connection with the process and the same time do not overload the process model with unnecessary time transitional parameters.

The work primarily focuses on the design of algorithms and methods for supporting logistics data analysis. However, it can be adjusted and be applicable to other areas accordingly, which makes the approach flexible and versatile. This analysis of processes with their attributes might be used for decision-making systems and process maps in future. Furthermore, the support of the descriptive (*As is*) current process model with certain notation and the integration with relevant bottleneck and predictive methods compromise the advantages of the approach.

Some challenges still lie in data quality and limitations of event logs and the domain area. I offered some robust methods to preprocess and clean data which were applied and helped to get higher results in the further analysis. However, the last step of the proposed framework – building of process map was described just theoretically in the work. The scope of this thesis does not cover the development of the process map and suggests it for future research.

5.2 Theoretical contribution

The results of the work are presented as a consistent set of methods and techniques, which are strongly interconnected and depend on each other results:

- the framework to deal with unstructured and unknown complex process modelling using additional parameters of the processes, activities and variants. This framework enables to rapidly construct a robust, reliable and transparent model to which others, in turn, can connect their own services. In terms of process mining, the framework shows the way of revealing hidden (also unused and unknown) subprocesses and handling *spaghetti* (non-structured, non-readable, poorly regulated and entangled processes) process models as well as presenting new perspectives/types of the process. Once it is developed, the process model can be used as a foundation and be enriched with additional parameters for recommendation or decision-support systems as well as for process maps.
- Proposal of the combined bottleneck detection method for the defined process model using context-aware information to detect weak points.
- Develop hybrid predictive models based on machine learning, statistics and process mining technologies using process variants and results of bottleneck algorithm as predictor variables. The operator can get the opportunity not just to better plan schedules, but also know risks, times, costs, next possible activity in the process.

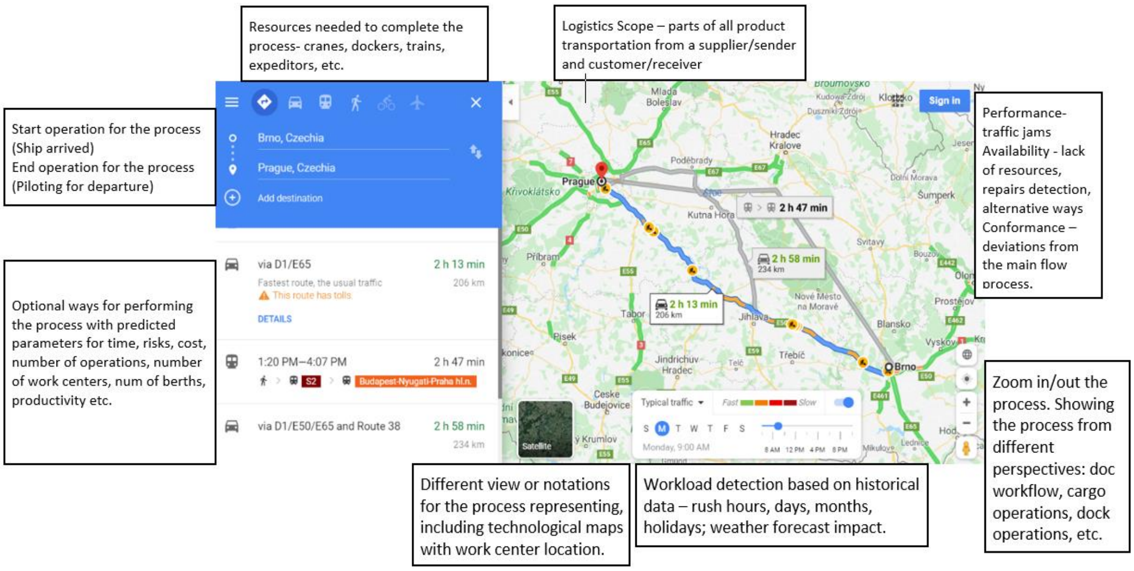
Besides these results, the thesis summarizes recent research, technologies, methods, and challenges of the current state of process mining, forecast models and bottlenecks detection methods as well as logistics domain.

5.3 Practical contribution

In the scope of this thesis, there was proposed a set of methods to develop a complex process model, reinforcing it with context awareness, bottleneck detection and predicted values. The model with all additional parameters can be merged and added to the process map, which is a powerful tool for process monitoring, audition, planning and forecasting. The advantage of the method is flexibility, so even if all steps are interconnected between each other and to obtain a substantial satisfying result – process map – all steps need to be completed, each step can be used independently. For instance, a two-step predictive model can be applied with another set of parameters on the input, excluding context awareness. It degrades the results but helps to make a fast analysis as well as to see dependencies between parameters in the first step – regression trees. The broad method described in the third part was applied partly to logistics processes in the seaport (ship handling process) and logistics in manufacturing processes (production line). All steps of the framework described in the Methodology part, except for the Enriched process map, ***were executed, implemented, and evaluated***. Some of them became part of the usual workflow in the ship handling process or production line optimization. Moreover, the web application for the prediction model with context-aware parameters was developed based on algorithms described in the work. More details of the application can be found in work [66].

5.4 Future work – Interactive process map

Despite the results achieved in this work, this topic still has many open areas for research. The part of the research described here has already covered the main pieces of the development of a process map for different types of processes. To offer new directions for future research (additionally to challenges from the Chapter 2), the process map presentation was created accordingly to the existing navigation system – google maps. If we consider that all roads represent processes traces, traffic jams are bottlenecks, prediction parameters are estimated performance characteristics for the situation, then the process map should look like in the Figure 5.1.



Resources needed to complete the process- cranes, dockers, trains, expeditors, etc.

Logistics Scope – parts of all product transportation from a supplier/sender and customer/receiver

Start operation for the process (Ship arrived)
End operation for the process (Piloting for departure)

Optional ways for performing the process with predicted parameters for time, risks, cost, number of operations, number of work centers, num of berths, productivity etc.

Performance- traffic jams
Availability - lack of resources, repairs detection, alternative ways
Conformance – deviations from the main flow process.

Zoom in/out the process. Showing the process from different perspectives: doc workflow, cargo operations, dock operations, etc.

Different view or notations for the process representing, including technological maps with work center location.

Workload detection based on historical data – rush hours, days, months, holidays; weather forecast impact.

Figure 5.1: Process map - navigation for logistics processes

The core of this map was already proposed in the work – the process model with zoom-in/out and different perspectives (framework to process *spaghetti* process models with context awareness), traffic jam and their causes (bottleneck detection method with process variants) and estimation of travel time (predictive models with additional calculated parameters from process mining explorative analysis).

The following recommendations can be carried out in future to maximize the impact of this work:

- Integrating results of the bottleneck and predictive model to the process mining workflow simulation, emulating the process animation.
- Adding more variants and specific process flows as well as resource analysis to the simulation model.
- Following the production process including manual resources (so-called hidden processes, information about which is not digitalized so far like manual operations).
- Creating a heatmap (performance) during day/week/year to define *prime time* in the process.
- Discussing additional metrics, methods and results with process experts to approach the real situation and to consider cases, in which data are not reflected in the data sets.

The creation of the process map requires more collaboration with current projects and the involvement of process owners as well as software engineers. Critical elements and methods of the process map are connected to the ones, explained in the research. I have provided a guideline to the logistics analysts in extracting event logs, processing and cleaning data, building the descriptive model and performance analysis. All obtained knowledge may weaken previous assumptions about the logistics process and the same time give the analysts a powerful tool to communicate with process owners/ stakeholders and see new perspectives of research.

Since the processes can be found everywhere, one possible future development of this work is to implement all steps of the proposed framework and to create advanced knowledgebase about different kinds of processes (financial, administrative, educational, web, security, etc.) as well as to build flexible and easy adjustable process maps templates for each of them.

Bibliography

[1] Chebbi, I., Wadii, B., and Imed, R. F. (2015). Big data: Concepts, challenges and applications. In *Computational collective intelligence*, pp. 638-647.

[2] Palmer, M. (2006). Data is the new oil. *ANA marketing maestros 3*.

[3] Van Der Aalst, W. (2016). Process mining: data science in action. In *Heidelberg: Springer*, vol.2.

[4] Provost, F., and Tom, F. (2013). Data Science for Business: What you need to know about data mining and data-analytic thinking. In *O'Reilly Media, Inc.*

[5] Aguilar, S., Ruth, S. (2004). Business process modelling: Review and framework. In *International Journal of production economics*, vol. 90.2, pp. 129-149.

[6] Munoz-Gama, J. (2016). Conformance checking and diagnosis in process mining. In *Springer International Publishing AG*.

[7] Curtis, B., Marc, I. K., and Over, J (1992). Process modeling. In *Communications of the ACM*, vol. 35.9, pp. 75-90.

[8] Mans, Ronny S., et al. (2008). Application of process mining in healthcare–a case study in a dutch hospital. In *International joint conference on biomedical engineering systems and technologies*. Springer, Berlin, Heidelberg.

[9] Mahendrawathi, E. R., Hanim, M. A., and Ayu N. (2015). Analysis of customer fulfilment with process mining: A case study in a telecommunication company. In *Procedia Computer Science*, vol.72, pp. 588-596.

[10] Poggi, N., et al. (2013). Business process mining from e-commerce web logs. In *Business process management*. Springer, Berlin, Heidelberg, pp. 65-80.

[11] Yampaka, T., and Prabhas, C. (2016). An application of process mining for queueing system in health service. In *13th International Joint Conference on Computer Science and Software Engineering (JCSSE)*. IEEE.

[12] Rubin, V., Günther, C. W., Van Der Aalst, W., Kindler, E., Van Dongen, B. F., and Schäfer, W. (2007). Process mining framework for software processes. In *International Conference on Software Process*, pp. 169–181.

[13] Trcka, N., Pechenizkiy, M., and van der Aalst, W. (2010). Process mining from educational data. In *Hands. Educ. Data Min*., pp. 123–142.

[14] Wang, Y., et al. (2014). Acquiring logistics process intelligence: Methodology and an application for a Chinese bulk port. In *Expert Systems with Applications*, vol. 41.1, pp.195-209.

[15] Veenstra, A. W., and LA Harmelink, R. (2022). Process mining ship arrivals in port: the case of the Port of Antwerp. In *Maritime Economics & Logistics*, vol. 24.3, pp. 584-601.

[16] Rozinat, A. and Van der Aalst, W. (2008). Conformance checking of processes based on monitoring real behavior. In *Information Systems*, vol. 33, no. 1, pp. 64–95.

[17] Rozinat, A., and van der Aalst, W. (2006). Decision mining in ProM. in *International Conference on Business Process Management*, pp. 420–425.

[18] Leemans, S. J., Fahland, D., and van der Aalst, W. (2014). Process and Deviation Exploration with Inductive Visual Miner. In *BPM Demos*, vol. 1295, p. 46.

[19] Huser, V. (2012). Process mining: Discovery, conformance and enhancement of business processes. pp. 1018-1019.

[20] de Medeiros, A., Weijters, A. and van der Aalst, W. (2007). Genetic process mining: an experimental evaluation. In *Data mining and knowledge discovery*, vol. 14.2, pp. 245-304.

[21] Angluin, D. C. (1976). An application of the theory of computational complexity to the study of inductive inference. *University of California, Berkeley*.

[22] Rudnitckaia, J., et al. (2019). Applying process mining to the ship handling process at oil terminal. In *2019 IEEE International Conference on Industrial Cyber Physical Systems (ICPS*). IEEE.

[23] Günther, C. W., and van der Aalst, W. (2007). Fuzzy mining– adaptive process simplification based on multi-perspective metrics. in *International conference on business process management*, pp. 328–343.

[24] van der Aalst, W. (2019). A practitioner's guide to process mining: limitations of the directly-follows graph. In *Procedia Computer Science*, vol.164, pp.321-328.

[25] van der Aalst, W. Process mining: discovery, conformance and enhancement of business processes. (2011). Springer-Verlag.

[26] van der Aalst, W., et al. (2011). Process mining manifesto. International conference on business process management. Springer, Berlin, Heidelberg.

[27] van der Aalst, W. (2020). Academic view: development of the process mining discipline. *Process Mining in Action*. Springer, Cham, pp.181-196.

[28] van der Aalst, W., et al. (2017). Learning hybrid process models from events: process discovery without faking confidence. In *International Conference on Business Process Management (BPM 2017*). Springer, Berlin.

[29] van der Aalst, W. (2019). Object-centric process mining: dealing with divergence and convergence in event data. In *Software Engineering and Formal Methods (SEFM 2019).* Springer, Berlin.

[30] Berti, A., and van der Aalst, W. (2020). Discovering Multiple Viewpoint Models from Relational Databases. In *International Symposium on Data-driven Process Discovery and Analysis*, vol. 379 of Lecture Notes in Business Information Processing, pp 24–51. Springer-Verlag, Berlin.

[31] Becker, T., and Wacharawan, I. (2017). Context aware process mining in logistics. In *Procedia Cirp*, vol. 63, pp. 557-562.

[32] Rafiei, M., and van der Aalst, W. (2021). Privacy-preserving continuous event data publishing. In *International Conference on Business Process Management*. Springer, Cham.

[33] Baldauf, M., Dustdar, S., and Rosenberg, F. (2007). A survey on context aware systems. In *Journal Ad Hoc Ubiquitous Comput*., vol. 2, no. 4, pp. 263–277.

[34] Bottazzi, D., Corradi, A., and Montanari, R. (2006). Context-aware middleware solutions for anytime and anywhere emergency assistance to elderly people. In *IEEE Commun*. Mag., vol. 44, no. 4, pp. 82–90.

[35] Abowd, G. D., Dey, A. K., Brown, P. J., Davies, N., Smith, M., and Steggles, P. (1999). Towards a better understanding of context and context awareness. In *International Symposium on Handheld and Ubiquitous Computing*, pp. 304–307.

[36] Chow, H., King, L. C., and Wing, B. L. (2007). A dynamic logistics process knowledge-based system–An RFID multi-agent approach. In *Knowledge-Based Systems*, vol. 20.4, pp. 357-372.

[37] Vernimmen, B., Dullaert, W., and Engelen, S. (2007). Schedule unreliability in liner shipping: Origins and consequences for the hinterland supply chain. In *Maritime Economics & Logistics*, vol. 9.3, pp.193-213.

[38] Douma, A. (2008). Aligning the operations of barges and terminals through distributed planning. *University of Twente*, Enschede, Netherlands.

[39] Loklindt, C., Moeller, M.P., Kinra, A. (2018). How Blockchain Could Be Implemented for Exchanging Documentation in the Shipping Industry. DOI:10.1007/978-3-319-74225-0_27.

[40] Jensen, T., Bjørn-Andersen, N., Vatrapu, R. (2014). Avocados Crossing Borders: The Missing Common Information Infrastructure for International Trade. In *Computational Social Science Laboratory (CSSL).*

[41] Talley, W. K., and ManWo Ng. (2016). Port multi-service congestion. In *Transportation Research Part E: Logistics and Transportation Review*, vol. 94, pp. 66-70.

[42] Ascencio, L. M., González-Ramírez, R. G., Bearzotti, L. A., Smith, N. R., Camacho-Vallejo, J. F. (2014). A collaborative supply chain management system for a maritime port logistics chain. In *Journal of applied research and technology*, vol. 12(3), pp. 444-458.

[43] Loh, H. S., Thai, V. V. (2016). Managing port-related supply chain disruptions (PSCDs): a management model and empirical evidence. In *Maritime Policy & Management*, vol. 43(4), pp.436-455.

[44] Kaljouw, S., Bouman, P. C., and Sharif Azadeh, S. (2019). Tugboat resting location optimization using AIS data analysis. *Erasmus University Rotterdam*.

[45] Keogh, K., Sonenberg, L. et al. (2012). Adaptive coordination in distributed and dynamic agent organizations. In *S. Cranefield et al. (Eds.), COIN 2011, LNCS*, vol. 7254, pp. 38-57.

[46] Fransen, R. W., and Davydenko, I. Y. (2021). Empirical agent-based model simulation for the port nautical services: A case study for the Port of Rotterdam. In *Maritime Transport Research,* 2:100040.

[47] van Putten, M. P. (2005). Improving yard deployment efficiency at APM Terminals Rotterdam. *Stan Ackermans Instituut*.

[48] Loklindt, C. M., Marc-Philip, Kinra, A. (2018). How Blockchain Could Be Implemented for Exchanging Documentation in the Shipping Industry. DOI:10.1007/978-3-319-74225-0_27.

[49] Adeyinka, A., Kareem, B. (2018). The application of Queuing Theory in Solving Automobile Assembly Line Problem. In *International Journal of Engineering and Technical Research*, vol. 7.

[50] Roser, C., Lorentzen, K., Deuse. J. (2015). Reliable shop floor bottleneck detection for flow lines through process and inventory observations: the bottleneck walk. In *Logistics Research*, vol. 8.1, pp.1-9.

[51] Cox III, James, F., and Schleier, Jr. (2010). Theory of constraints handbook. *McGraw-Hill Education*.

[52] Bertagnolli, F. (2022). Lean Management: Introduction and In-Depth Study of Japanese Management Philosophy. *Springer Nature*

[53] Moore, R., Scheinkopf, L. (1998). Theory of constraints and lean manufacturing: friends or foes. *Chesapeake Consulting Inc*.

[54] Rudnitckaia, J. (2017). Methods of detection of noises and outliers in the data structure on the example of ship handling duration at an oil terminal. In *Vestnik Gosudarstvennogo Universiteta Morskogo I Rechnogo Flota Imeni Admirala Makarova*, vol.4.44, pp.866-873.

[55] Roser, C., Lorentzen, K., Deuse, J. (2015). Reliable shop floor bottleneck detection for flow lines through process and inventory observations: the bottleneck walk. *Logistics Research*, 8(1).

[56] Rausch, P., Alaa, F. Sheta, A. (2013). Business intelligence and performance management: theory, systems and industrial applications. *Springer Science & Business Media*.

[57] Box, G. (2015). Time series analysis: forecasting and control. *John Wiley & Sons.*

[58] Ruppert, D., Matteson, D. (2015). Statistics and data analysis for financial engineering: with R examples. *Springer*.

[59] Brockwell, P. J., Davis, R.A. (2016). Introduction to time series and forecasting. *Springer*.

[60] NIST/SEMATECH e-Handbook of Statistical Methods. Forecasting with Single Exponential Smoothing. Retrieved on 2017/06/20 from http://www.itl.nist.gov/div898/handbook/pmc/section4/pmc432.htm

[61] van der Aalst, W., Schonenberg, M., Song, M. (2011). Time prediction based on process mining. *Information systems*, vol. 36.2, pp.450-475.

[62] van der Aalst, W, Rubin, V., van Dongen, B.F., Kinfler, E., Gunther, C. W. (2016). Process Mining: A two-step loose approach using transition systems and regions. *BPM Center Report BPM-06-3*0, BPM Center

[63] Rozinat, A. (2015). Disco User's Guide. Discover your Process.

[64] Rudnitckaia, J., Hruska, T. (2018). Prediction of times and delays for ship handling process based on a transition system. In *Proceedings of the Int. Conf. on Harbor Maritime and Multimodal Logistics*, pp. 39–43.

[65] Bichou, K., Gray, R. (2004). A logistics and supply chain management approach to port performance measurement. In *Maritime Policy & Management*, vol. 31.1, pp.47-67.

[66] Venkatachalam, H.S. (2020). Value Stream Analysis of Gas Meter Production by Simulation and Forecasting. *Master thesis.*

[67] Rudnitckaia, J., Intayoad, W., Becker, T., Hruška, T. (2019). Applying process mining to the ship handling process at oil terminal. In *2019 IEEE International Conference on Industrial Cyber Physical Systems (ICPS),* pp. 552-557.

[68] Rudnitckaia, J., Venkatachalam, H. S., Essmann, R., Hruška, T., Colombo, A. W. (2022). Screening Process Mining and Value Stream Techniques on Industrial Manufacturing Processes: Process Modelling and Bottleneck Analysis. *IEEE Access*, vol. 10, pp. 24203-24214.

[69] Kourounioti, I., Kurapati, S., Lukosch, H., Tavasszy, L., Verbraeck, A. (2018). Simulation Games to Study Transportation Issues and Solutions: Studies on Synchromodality. In *Transportation Research Record*, vol. 2672(44), pp. 72–81. https://doi.org/10.1177/0361198118792334.

[70] van Geffen, F., Niks, R. (2013). Accelerate DMAIC using Process Mining. *In BPIC@ BPM*.

[71] Rudnitckaia, J., Hruška, T. (2017). Time Series Analysis and Prediction Statistical Models for the Duration of the Ship Handling at an Oil Terminal. In *International Conference on Reliability and Statistics in Transportation and Communication*, pp. 127-136. Springer, Cham.

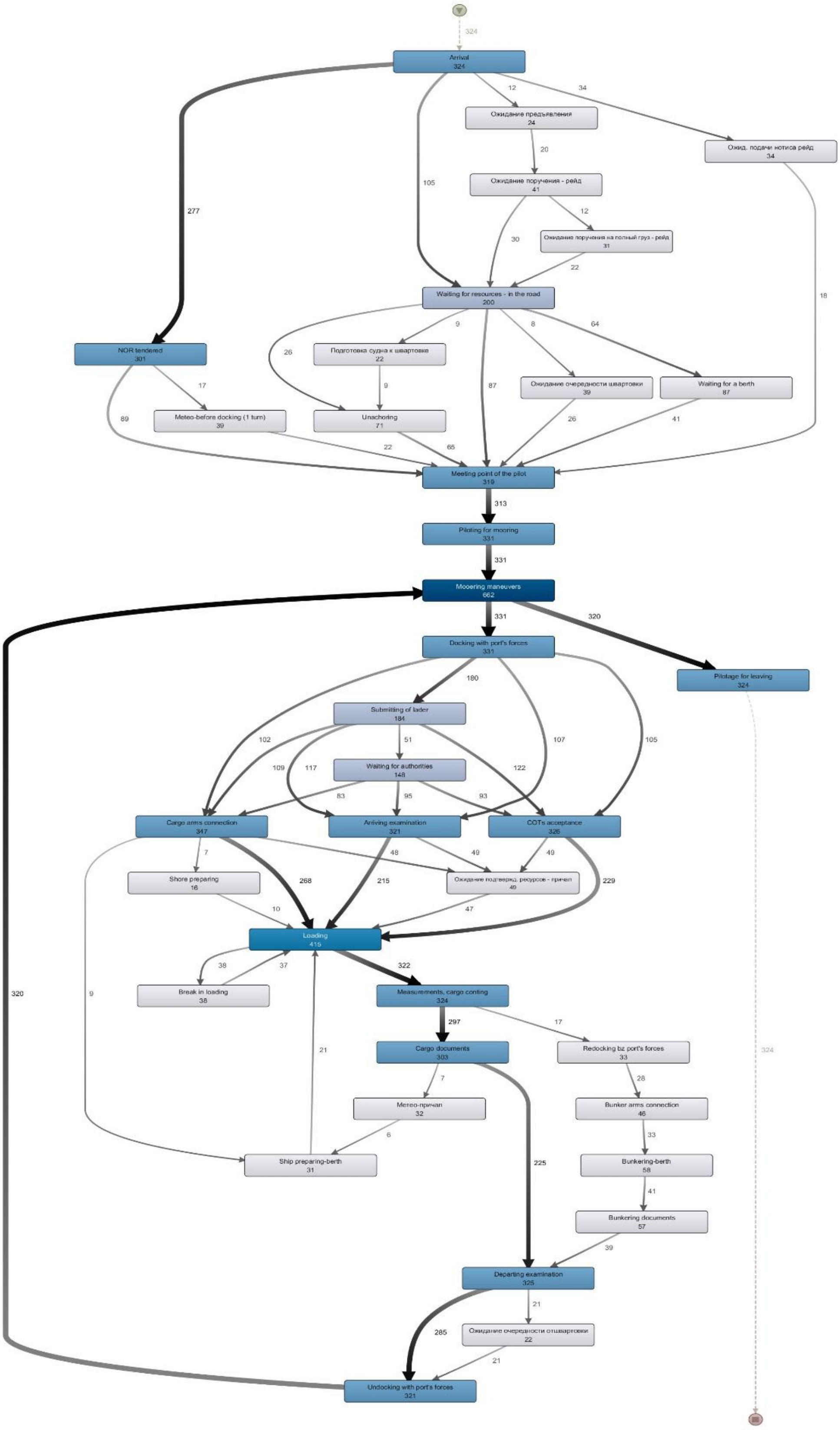
[72] Doherty, D., O' Riordan, C. (2007). A phenotypic analysis of GP-evolved team behaviours. pp.1951-1958. DOI:10.1145/1276958.1277347.

Abbreviations

AI – Artificial Intelligence
API – Application Programming Interface
ARIMA – Autoregressive Integrated Moving Average
BPM – Business Process Management
BPMN - Business Process Model and Notation
CRM – Customer Relationship Management
DFG – The Directly-Follows Graph
DM – Data Mining
EL – Event Log
EPCs – Event-Driven Process Chain
ERP – Enterprise Resource Planning
FACT challenge – Fairness, Accuracy, Confidentiality, Transparency
IoT – Internet of things
KPI – Key Performance Indicators
MAE – Mean absolute error
MAPE – Mean absolute percentage error
MSE – Mean squared error
MXML – Mining eXtensible Markup Language
ODS – Operational Data Store
PM – Process Mining
RMSE – Root Mean Square Error
ROC curve – Receiver Operating Characteristics curve
RPA – Robotic process automation
SAP – Systemanalyse und Programmentwicklung (eng. System Analysis and Software Development)
SARIMA – Seasonal Autoregressive Integrated Moving Average
SLA – Service Level Agreement
TOC – Theory of constraints
TS – Time series
UML – Unified Modeling Language
WF – Workflow net
WIP – Work-in-Progress
XES – eXtensible Event Stream
YAWL – Yet Another Workflow Language Xtensible Markup Language

Appendix A

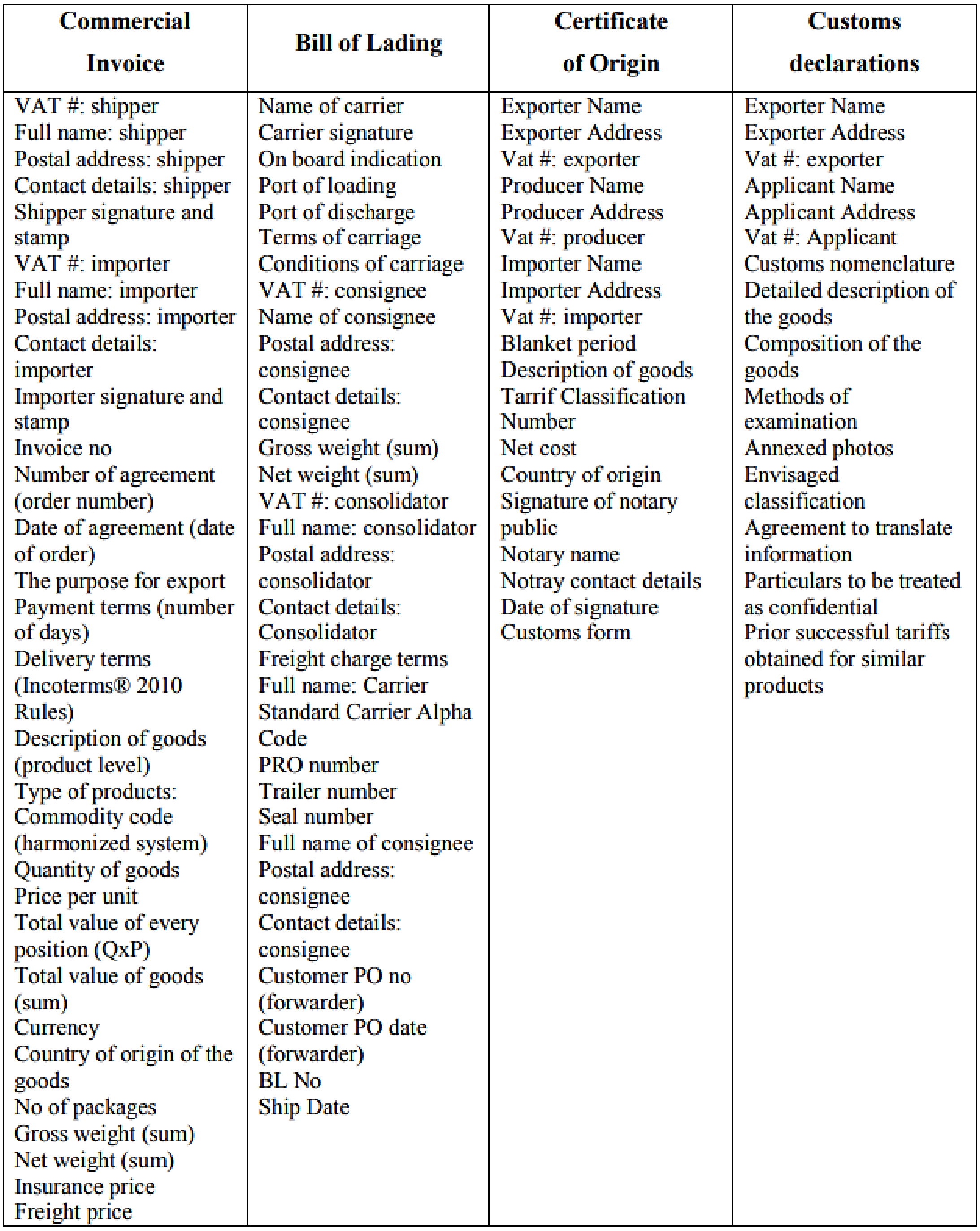
Complete descriptive process model for a ship-handling process made by Fuzzy miner and presented in DFG.



Arrival 324
Ожидание предъявления 24
Ожид. подачи нотиса рейд 34
Ожидание поручения - рейд 41
Ожидание поручения на полный груз - рейд 31
Waiting for resources - in the road 200
NOR tendered 301
Подготовка судна к швартовке 22
Ожидание очередности швартовки 39
Waiting for a berth 87
Meteo-before docking (1 turn) 39
Unachoring 71
Meeting point of the pilot 319
Piloting for mooring 331
Mooering maneuvers 662
Docking with port's forces 331
Pilotage for leaving 324
Submitting of lader 184
Waiting for authorities 148
Cargo arms connection 347
Arriving examination 321
COTs acceptance 326
Shore preparing 16
Ожидание подтвержд. ресурсов - причал 49
Loading 415
Break in loading 38
Measurements, cargo conting 324
Cargo documents 303
Redocking bz port's forces 33
Метео-причал 32
Bunker arms connection 46
Ship preparing-berth 31
Bunkering-berth 58
Bunkering documents 57
Departing examination 325
Ожидание очередности отшвартовки 22
Undocking with port's forces 321

Appendix B

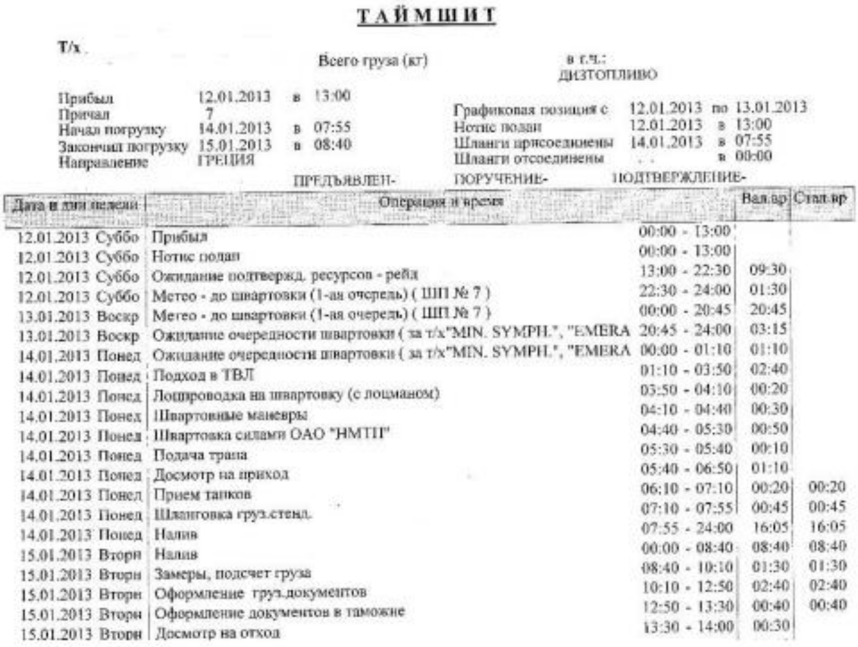
Features description: Loklindt, Christopher & Moeller, Marc-Philip & Kinra, Aseem [39]



Commercial Invoice	Bill of Lading	Certificate of Origin	Customs declarations
VAT #: shipper Full name: shipper Postal address: shipper Contact details: shipper Shipper signature and stamp VAT #: importer Full name: importer Postal address: importer Contact details: importer Importer signature and stamp Invoice no Number of agreement (order number) Date of agreement (date of order) The purpose for export Payment terms (number of days) Delivery terms (Incoterms® 2010 Rules) Description of goods (product level) Type of products: Commodity code (harmonized system) Quantity of goods Price per unit Total value of every position (QxP) Total value of goods (sum) Currency Country of origin of the goods No of packages Gross weight (sum) Net weight (sum) Insurance price Freight price	Name of carrier Carrier signature On board indication Port of loading Port of discharge Terms of carriage Conditions of carriage VAT #: consignee Name of consignee Postal address: consignee Contact details: consignee Gross weight (sum) Net weight (sum) VAT #: consolidator Full name: consolidator Postal address: consolidator Contact details: Consolidator Freight charge terms Full name: Carrier Standard Carrier Alpha Code PRO number Trailer number Seal number Full name of consignee Postal address: consignee Contact details: consignee Customer PO no (forwarder) Customer PO date (forwarder) BL No Ship Date	Exporter Name Exporter Address Vat #: exporter Producer Name Producer Address Vat #: producer Importer Name Importer Address Vat #: importer Blanket period Description of goods Tarrif Classification Number Net cost Country of origin Signature of notary public Notary name Notray contact details Date of signature Customs form	Exporter Name Exporter Address Vat #: exporter Applicant Name Applicant Address Vat #: Applicant Customs nomenclature Detailed description of the goods Composition of the goods Methods of examination Annexed photos Envisaged classification Agreement to translate information Particulars to be treated as confidential Prior successful tariffs obtained for similar products

Appendix C

C1 The sample of timesheet document with all parameters (source: Seaport terminal official documentation)



ТАЙМШИТ

Т/х

Всего груза (кг)

в т.ч.:
ДИЗТОПЛИВО

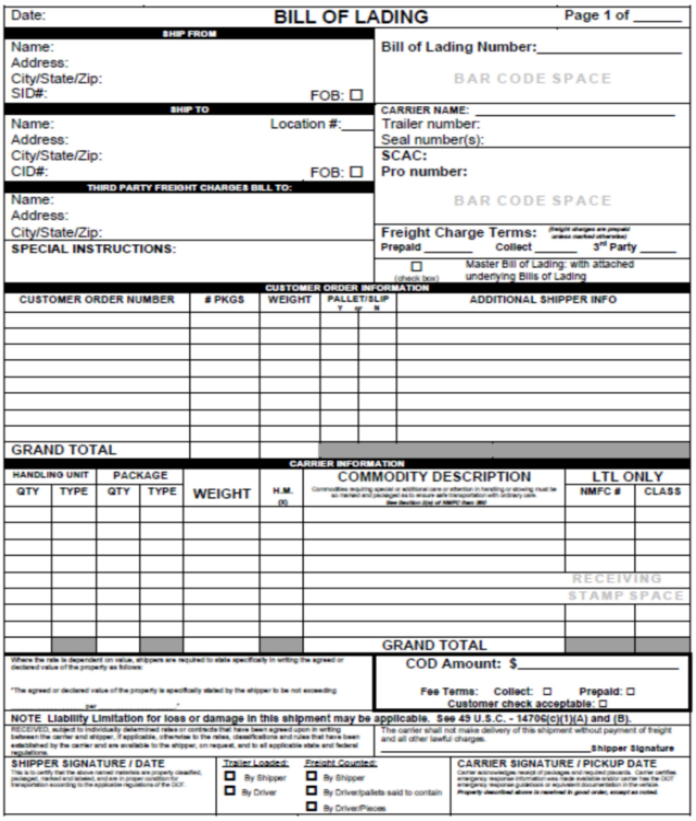
Прибыл	12.01.2013	в	13:00
Причал	7		
Начал погрузку	14.01.2013	в	07:55
Закончил погрузку	15.01.2013	в	08:40
Направление	ГРЕЦИЯ		

Графиковая позиция с	12.01.2013	по	13.01.2013
Нотис подан	12.01.2013	в	13:00
Шланги присоединены	14.01.2013	в	07:55
Шланги отсоединены	. .	в	00:00

ПРЕДЪЯВЛЕН- ПОРУЧЕНИЕ- ПОДТВЕРЖДЕНИЕ-

Дата и дни недели	Операция и время		Вал.вр	Стал.вр
12.01.2013 Суббо	Прибыл	00:00 - 13:00		
12.01.2013 Суббо	Нотис подан	00:00 - 13:00		
12.01.2013 Суббо	Ожидание подтвержд. ресурсов - рейд	13:00 - 22:30	09:30	
12.01.2013 Суббо	Метео - до швартовки (1-ая очередь) (ШП № 7)	22:30 - 24:00	01:30	
13.01.2013 Воскр	Метео - до швартовки (1-ая очередь) (ШП № 7)	00:00 - 20:45	20:45	
13.01.2013 Воскр	Ожидание очередности швартовки (за т/х"MIN. SYMPH.", "EMERA	20:45 - 24:00	03:15	
14.01.2013 Понед	Ожидание очередности швартовки (за т/х"MIN. SYMPH.", "EMERA	00:00 - 01:10	01:10	
14.01.2013 Понед	Подход в ТВЛ	01:10 - 03:50	02:40	
14.01.2013 Понед	Лоцпроводка на швартовку (с лоцманом)	03:50 - 04:10	00:20	
14.01.2013 Понед	Швартовные маневры	04:10 - 04:40	00:30	
14.01.2013 Понед	Швартовка силами ОАО "НМТП"	04:40 - 05:30	00:50	
14.01.2013 Понед	Подача трапа	05:30 - 05:40	00:10	
14.01.2013 Понед	Досмотр на приход	05:40 - 06:50	01:10	
14.01.2013 Понед	Прием танков	06;10 - 07:10	00:20	00:20
14.01.2013 Понед	Шланговка груз.стенд.	07:10 - 07:55	00:45	00:45
14.01.2013 Понед	Налив	07:55 - 24:00	16:05	16:05
15.01.2013 Вторн	Налив	00:00 - 08:40	08:40	08:40
15.01.2013 Вторн	Замеры, подсчет груза	08:40 - 10:10	01:30	01:30
15.01.2013 Вторн	Оформление груз.документов	10:10 - 12:50	02:40	02:40
15.01.2013 Вторн	Оформление документов в таможне	12:50 - 13:30	00:40	00:40
15.01.2013 Вторн	Досмотр на отход	13:30 - 14:00	00:30	

C2 The sample of bill of lading document (source: Law office of Seaton & Husk (2017)).



Date: BILL OF LADING Page 1 of ______

SHIP FROM

Name:
Address:
City/State/Zip:
SID#: FOB: □

Bill of Lading Number:________________

BAR CODE SPACE

SHIP TO

Name: Location #:____
Address:
City/State/Zip:
CID#: FOB: □

CARRIER NAME: ______________________
Trailer number:
Seal number(s):

SCAC:
Pro number:

BAR CODE SPACE

THIRD PARTY FREIGHT CHARGES BILL TO:

Name:
Address:
City/State/Zip:

SPECIAL INSTRUCTIONS:

Freight Charge Terms: (freight charges are prepaid unless marked otherwise)
Prepaid ________ Collect ________ 3rd Party ______

□ (check box) Master Bill of Lading: with attached underlying Bills of Lading

CUSTOMER ORDER INFORMATION

CUSTOMER ORDER NUMBER	# PKGS	WEIGHT	PALLET/SLIP Y or N		ADDITIONAL SHIPPER INFO
GRAND TOTAL					

CARRIER INFORMATION

HANDLING UNIT		PACKAGE		WEIGHT	H.M. (X)	COMMODITY DESCRIPTION	LTL ONLY	
QTY	TYPE	QTY	TYPE			Commodities requiring special or additional care or attention in handling or stowing must be so marked and packaged as to ensure safe transportation with ordinary care. See Section 2(e) of NMFC Item 360	NMFC #	CLASS
							RECEIVING	
							STAMP SPACE	
						GRAND TOTAL		

Where the rate is dependent on value, shippers are required to state specifically in writing the agreed or declared value of the property as follows:

"The agreed or declared value of the property is specifically stated by the shipper to be not exceeding ______________ per ______________."

COD Amount: $________________
Fee Terms: Collect: □ Prepaid: □
Customer check acceptable: □

NOTE Liability Limitation for loss or damage in this shipment may be applicable. See 49 U.S.C. - 14706(c)(1)(A) and (B).

RECEIVED, subject to individually determined rates or contracts that have been agreed upon in writing between the carrier and shipper, if applicable, otherwise to the rates, classifications and rules that have been established by the carrier and are available to the shipper, on request, and to all applicable state and federal regulations.

The carrier shall not make delivery of this shipment without payment of freight and all other lawful charges.

____________________Shipper Signature

SHIPPER SIGNATURE / DATE
This is to certify that the above named materials are properly classified, packaged, marked and labeled, and are in proper condition for transportation according to the applicable regulations of the DOT.

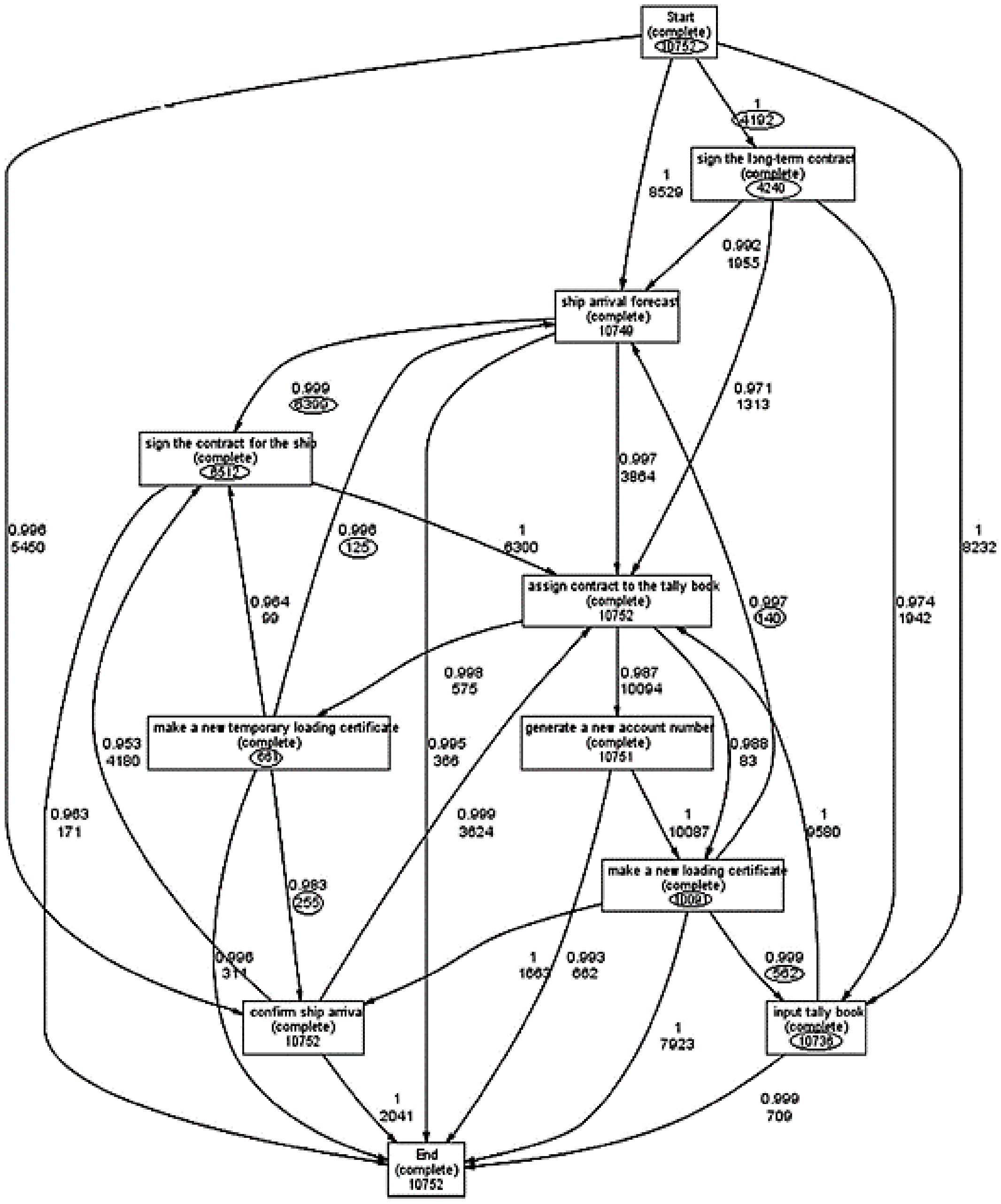
Trailer Loaded:
□ By Shipper
□ By Driver

Freight Counted:
□ By Shipper
□ By Driver/pallets said to contain
□ By Driver/Pieces

CARRIER SIGNATURE / PICKUP DATE
Carrier acknowledges receipt of packages and required placards. Carrier certifies emergency response information was made available and/or carrier has the DOT emergency response guidebook or equivalent documentation in the vehicle.
Property described above is received in good order, except as noted.

Appendix D

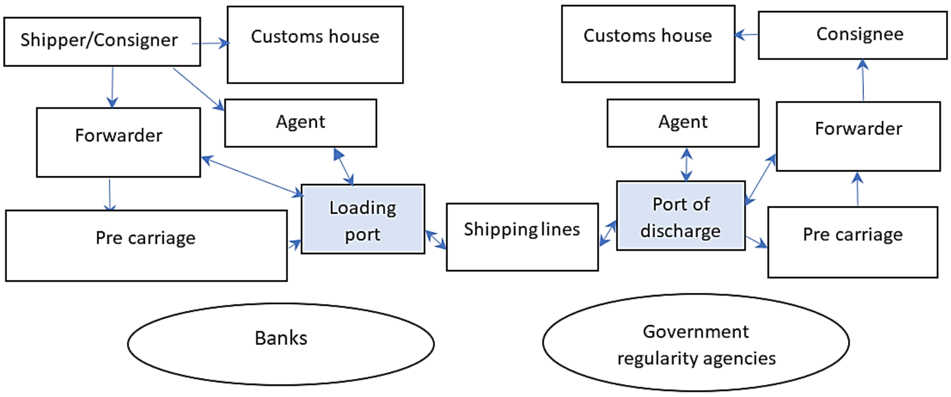
The transition system model, which was created for ship-handling process to explore this type of prediction. The model was created by PROM tool.



Appendix E

Questionnaire for better understanding challenges and needs of the maritime industry, their processes and communications (was used on the interview with logistics companies leads like Hapag Lloyd, MSC, DBH). The results were applied for exploring domain area – logistics processes.

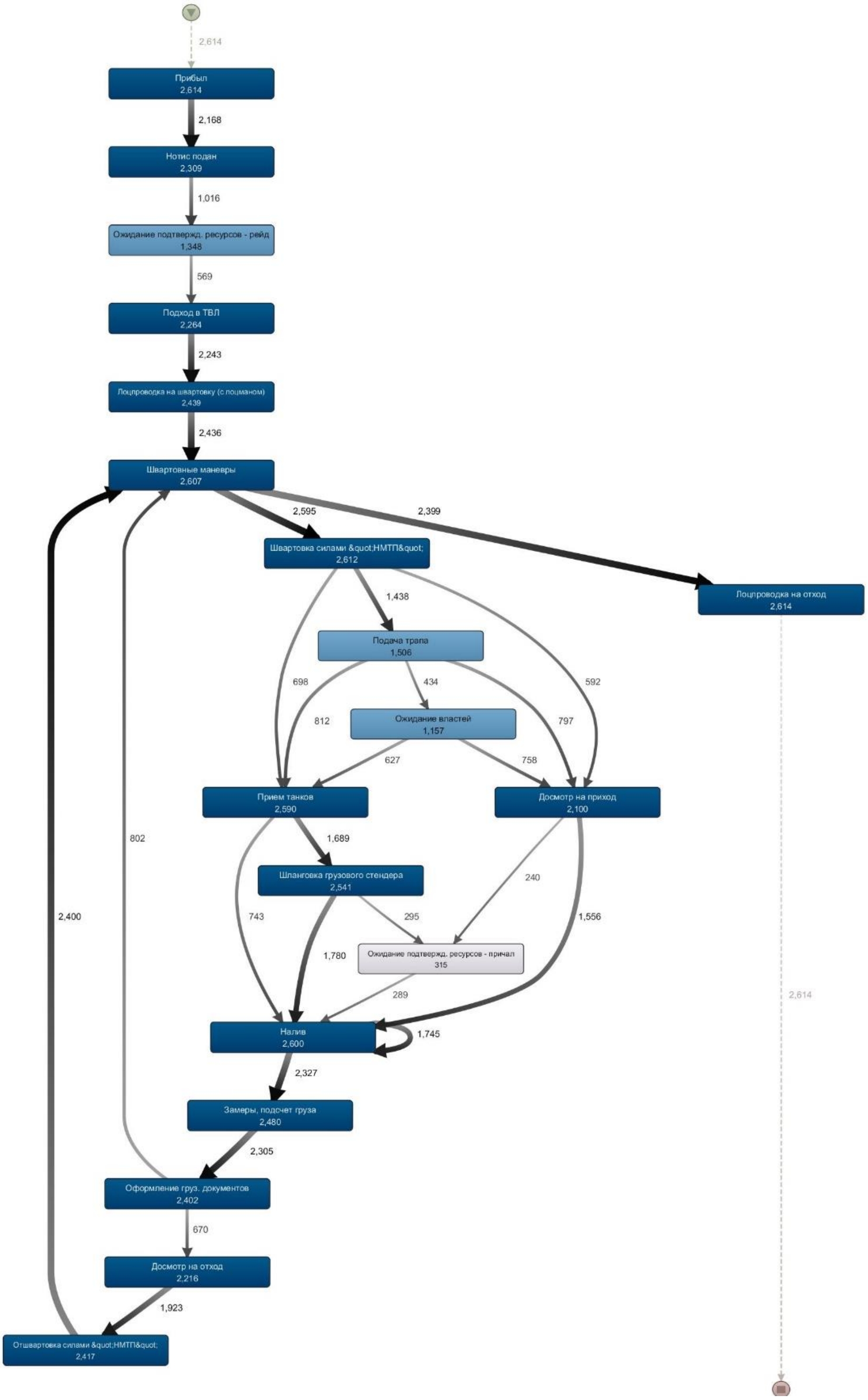
1. What kind of the process do you work with (which relation on the scheme – supply-chain, shipping, ship-handling, etc.)? Which part of the transport logistics process?



2. What data do you usually have from stakeholders?
3. What kind of IS are you working with? (SAP, CRM, ERP…). Do you have a process model you follow? (BPMN, URL diagrams, WF, etc.). Who use the process model?
4. What operations/activities do you conduct within transport process (loading, storage…)?
5. What transport documents are you responsible for? (Bill of lading, Cargo documents, Manifest, Certificate of Origin, Customs declarations, Shipping order, etc.)?
6. Have you ever analyzed data you have? What methods do you use and who make analysis? (statistics methods, AI, data mining, SQL, etc.)
7. What kind of report/knowledge do you usually extract from data you have? (predictions, tracking process of the cargo, recommendations for further improving of the transport process, etc.)
8. Which part of logistics process can be improved/optimized in your opinion? (docflow, dispatcher work, infrastructure, using new methods of information processing, etc.)
9. What challenges can occur in business process with implementing of new methods/technologies?

Appendix F

The complete descriptive process model for a ship-handling process made by Fuzzy miner and presented in DFG (after preprocessing data and adjusting visualization). All operation names are taken in original language and will be encrypted in the thesis to keep confidentiality.


2,614
Прибыл
2,614
2,168
Нотис подан
2,309
1,016
Ожидание подтвержд. ресурсов - рейд
1,348
569
Подход в ТВЛ
2,264
2,243
Лоцпроводка на швартовку (с лоцманом)
2,439
2,436
Швартовные маневры
2,607
2,595
2,399
Швартовка силами &quot;НМТП&quot;
2,612
Лоцпроводка на отход
2,614
1,438
Подача трапа
1,506
698
434
592
812
Ожидание властей
1,157
797
627
758
Прием танков
2,590
Досмотр на приход
2,100
802
1,689
Шланговка грузового стендера
2,541
240
2,400
743
295
1,556
1,780
Ожидание подтвержд. ресурсов - причал
315
289
2,614
Налив
2,600
1,745
2,327
Замеры, подсчет груза
2,480
2,305
Оформление груз. документов
2,402
670
Досмотр на отход
2,216
1,923
Отшвартовка силами &quot;НМТП&quot;
2,417