UNIVERSITY OF MACEDONIA
DEPARTMENT OF APPLIED INFORMATICS
GRADUATE PROGRAM

Assessing the Complexity of BPMN Models for Redesign using Cluster Analysis

M.Sc. THESIS

of

Chrysoula Fotoglou

Thessaloniki, October 2019

Assessing the Complexity of BPMN Models for Redesign using Cluster Analysis

Chrysoula Fotoglou

B.Sc. in Applied Informatics, Department of Applied Informatics
University of Macedonia, 2017

M.Sc. Thesis

submitted as a partial fulfillment of the requirements for

THE DEGREE OF MASTER OF SCIENCE IN APPLIED INFORMATICS

Supervisor: Dr Kostas Vergidis

Approved by examining board on 29 October 2019

Prof Maro Vlachopoulou Prof Alexander Chatzigeorgiou Dr Kostas Vergidis

Abstract

Business Processes exist at the core of each organisation and their efficient management is a main objective for those aiming to benefit from a process-centric approach. Today's rapidly changing economic environment introduces the challenge of analysing, maintaining and optimizing increasingly complex processes. Increased complexity is considered to have a negative impact on the success of Business Process Redesign initiatives. Several aspects of Business Process complexity have been studied in literature, mostly focussing on complexity measurement and proposal of appropriate complexity metrics.

The present research leverages metrics introduced in research for the development of complexity assessment methods. These methods aim at providing a straightforward way of evaluating a process model's complexity to draw conclusions regarding a model's capability for Redesign. Through the implementation of cluster analysis on BPMN models, based on selected complexity metrics, the identification of highly complex models becomes feasible. The developed methods utilize the K-means algorithm for clustering in order to create a model for distinguishing between complexity levels. The first proposed approach provides reference values for the categorization of future Business Process models, while a second approach combines complexity metrics to a single weighted sum measure. This approach is presented with the purpose of simplifying the assessment of models by the definition of exact thresholds, as well as offering a means of assigning priorities to complexity metrics. The latter is found to aid in the identification of problematic areas on a process model, by examining the impact of each metric on the overall complexity.

Acknowledgements

Writing this thesis for the completion of the MSc in Applied Informatics was a very strenuous yet rewarding process. I would like to express my gratitude to my supervisor, Dr Kostas Vergidis for his continuous support and counseling during this journey. His guidance was indispensable for the complexion of this research, as was the confidence he had in me in every step. I would also like to thank my family for their endless support throughout the course of my studies.

Table of contents

List of figures

List of tables

CHAPTER 1

Introduction

In today's rapidly changing environment, organizations are faced with the challenge of handling increasingly larger and more complex processes. As a result, Business Process Management (BPM) is widely used among organizations and is being developed extensively in the last decade. A significant part of BPM is the efficient representation, analysis and optimization of Business Processes, which is the main subject of study for the Business Process Modeling field. Various modeling languages for Business Processes (BPs) have been proposed in literature (e.g., EPCs, IDEF0, BPMN 2.0). The Business Process Modeling Notation 2.0 (BPMN 2.0) standard is one of the most popular one in the field. Being developed specifically for business process modeling, it incorporates many elements and artifacts that aid in the simple representation of complex concepts in the BPM field, hence its popularity.

An important aspect of business process models, that affects their capability of being analyzed, transformed and optimized, is their complexity. High complexity results in low understandability, limited maintainability and higher error probability. Most studies address process complexity as a measurable property, in an attempt to measure the quality of a process and the ease of understanding or modifying a process. Several measures have been proposed to quantify the complexity of a business process from various aspects.

The present thesis aims at developing methods to evaluate complexity of business processes in relation to their capability for Redesign. The concept of business processes is introduced in the next section, to clarify the use of the term for the rest of the study. In addition, this chapter specifies the main purpose of this thesis and concisely presents the thesis structure.

1.1 Motivation

This section briefly introduces the main concept of this research, which is business processes, and the motivation behind it. Essentially, a business process documents organisational activity. Business processes reveal the behaviour of an organization and are rarely visible to the external environment, e.g., customers. They are crucial for understanding how an organisation operates and are commonly utilized for the design of information systems. Organisations that adopt a process-centric approach enjoy benefits that include increased customer satisfaction and the ability to quickly adapt and evolve in a competitive environment [1].

A variety of formal definitions for what constitutes a business process exist in literature. For the premise of this thesis the definition proposed by Vergidis [2] is adopted. According to this author a business process is "a collective set of tasks that when properly connected and sequenced perform a business operation. The aim of a business process is to perform a business operation, i.e., any service-related operation that produces value to the organization".

Business Process Management (BPM) refers to a set of methods aimed at supporting the design, analysis and optimization of processes [1]. A core aspect of BPM is the utilization of modeling techniques to represent business processes in an efficient and understandable way, that facilitates the application of redesign initiatives. However, it is a common occurrence that low model quality and high complexity hinder the success of such initiatives. A complexity assessment of BP models is regarded essential in order to evaluate a model's capability for redesign. With respect to the latter acknowledgement, the aim of this thesis is presented next.

1.2 Aim and Objectives

The main purpose of this thesis is to measure the complexity of BP models in relation to transformation and redesign. Methods to quantify the complexity of BPMN Models, in order to classify them regarding their capability for redesign, are developed. The evaluation of the complexity is performed through the scope of redesign, using metrics selected for this goal from literature. By means of a cluster analysis of models included in the SOA-based Business Process Database, the identification of complex models becomes feasible. The main objective is to establish methods that enable the highly complex models' identification. The latter, ultimately, can be used as a means to reveal a model's capability for Redesign and potential need for normalization.

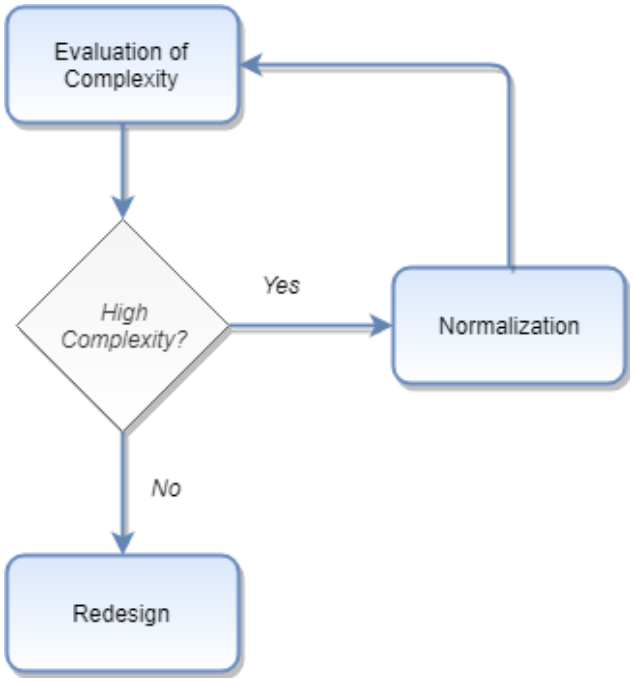

Figure 1: Thesis Overview

In particular, as depicted in Figure 1, the aim of this Thesis is to assess BP models represented in BPMN based on their complexity. This is achieved by implementing algorithms that, leveraging complexity metric values, safely yield a verdict about a BP's capability for Redesign or its necessity for normalization. In case a BP is revealed to exceeds thresholds set for the selected metrics, it is a clear indication of high complexity. This, in turn, reveals the difficulty of applying Redesign methods and practices in that process, since high complexity is presumed to hinder Redesign initiatives. To reduce complexity, the utilization of normalization techniques found in literature is required.

To summarize, the main objectives of this research include:

1. Selecting suitable measures from literature to measure complexity in relation to Redesign
2. Adapting these measures to be used on BPMN models
3. Developing methods that leverage on those measures and provide efficient complexity assessment for BP models
4. Using those methods to define reference values for the selected metrics and facilitate the identification of highly complex models

1.3 Thesis structure

The rest of the thesis is organized in a way that reveals the path followed by the author for the accomplishment of the defined objectives. Following, the structure of the thesis is provided, an outline of which is also displayed in Figure 2.

Chapter 2: This chapter offers the necessary theoretical background for the development of the research. In essence, it introduces the main concepts of Business Process Modeling, to be discussed throughout this thesis. Among others, this chapter concerns Business Process Complexity, its impact on Business Process Redesign and how it can be managed efficiently.

Chapter 3: Related research on complexity measurement and evaluation is presented in this chapter. Studies involving definitions of complexity metrics and establishment of reference values for these metrics are discussed in an attempt to determine the research gap addressed by the present thesis.

Chapter 4: This chapter offers an overview of the research methodology and clarifies the steps of the present study. Initially, it justifies the selection of specific measures from literature and presents them in detail. Next, the procedure to obtain the data for the development of the methods to evaluate complexity is presented. Last, technical details regarding the parameters used for the implementation of the applied algorithm are specified.

Chapter 5: In this chapter, the methods developed to accomplish complexity assessment are presented. Both methods employ cluster analysis to categorize models regarding complexity. The first approach provides reference values, while the second approach combines metrics to one single measure so as to offer a more straightforward definition of threshold values. The chapter concludes with the assessment of BPMN examples from literature, using the methods proposed.

Chapter 6: The last chapter includes an overview of the thesis along with the main conclusions derived by the development of the complexity assessment methods. In addition, the limitations of the research are presented, including directions for future work.

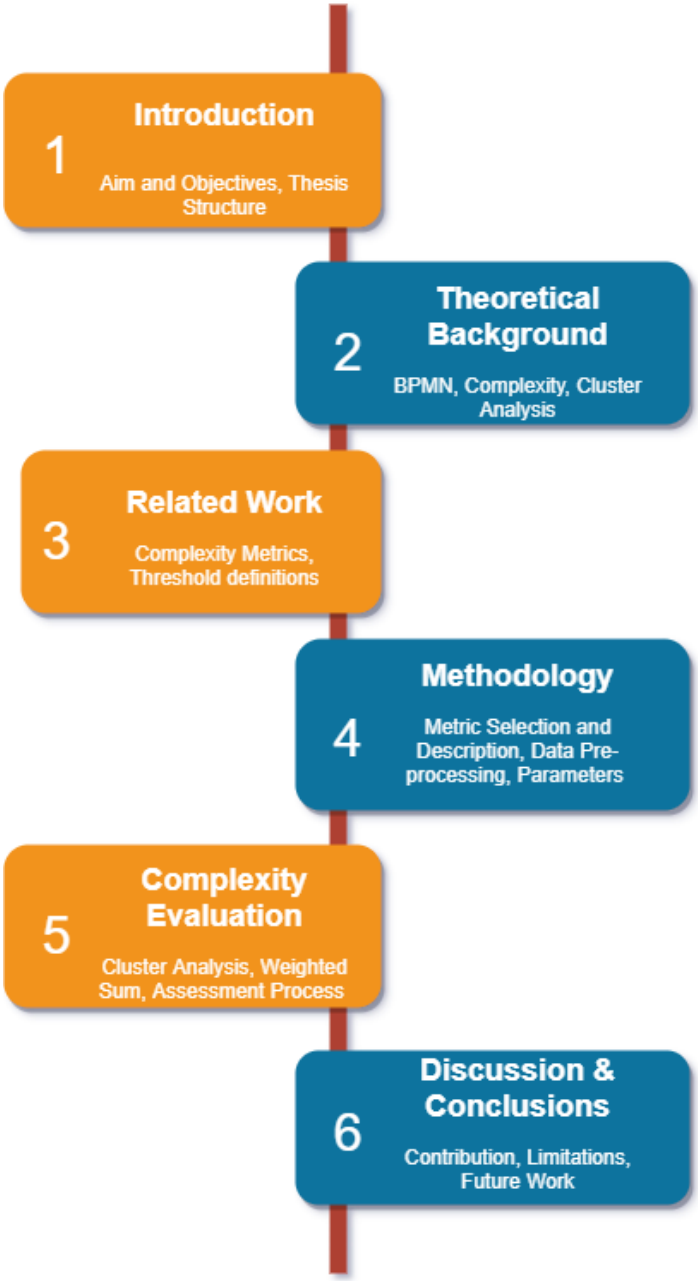

Figure 2: Thesis Structure

1.4 Summary

This chapter provided a brief description of the present thesis, which concerns complexity evaluation of BPMN models. An essential introduction of Business Processes is presented at first, leading to a clarification of the main purpose of the thesis and the defined objectives. The outline, including an overview of the main content of each chapter, is also described. The next chapter explains basic concepts concerning the subject of the present thesis.

CHAPTER 2

Theoretical Background

This chapter includes the main concepts to be discussed in the premise of this thesis. These concepts derive mostly from the business processes field and include definitions, modeling techniques, process redesign and process complexity. Special attention is given to the Business Process Modeling Notation (BPMN), a modeling technique used in the current research. In addition, the main idea of cluster analysis is explained and the algorithm K-means, which is the method utilized for analysing the data, is presented in detail.

2.1 Business Process Definitions

Several definition of what constitutes a business process (BP) are available in literature, with most researchers citing the definition proposed by Hammer and Champy [3] and Davenport [4]. Davenport [4] defines a business process as "the chain of activities whose final aim is the production of a specific output for a particular customer or market". In their work Hammer and Champy [3] view a business process as "a collection of activities that takes one or more kinds of inputs and creates an output that is of value to the customer". Many more proposals of a formal definition for business processes ensued. The multitude of definitions acts as an indicator of the variety and diversity of the term in literature, however it is evident that most definitions are quite similar and use similar concepts to describe a business process [2].

The main concepts present in business process definitions regard activities, input and output. Activities constitute the core element of a business process, that utilizes resources and are executed to transform these resources to required output. Generally, task is another term used to describe an activity. Input represents the resources utilized by the activities, while output refers to the transformed resources corresponding to a specific goal, mostly aimed at fulfilling a customer's need [2].

Opposite to this customer-centric approach, Vergidis [2], maintaining the same main concepts, but focusing mainly on the operational aspect offers a more inclusive definition of business processes. He defines a business process as "a collective set of tasks that when properly connected and sequenced perform a business operation. The aim of a business process is to perform a business operation, i.e., any service-related operation that produces value to the organization".

Efficient managing of business processes is a key objective for organization that follow a process-centric approach. Business Process Management (BPM) is a discipline involving a combination of business process control, automation, measurement, modeling and optimization, with the purpose of supporting enterprise goals, engaging employees, customers internal and external partners [5]. BPM attempts to better understand the key mechanisms of an organization in order to improve business performance by identifying opportunities, e.g., potential outsourcing or adaptation of new technologies [6]. For this reason BPM is recognized as a top priority for organization, hence the increased focus on discovering appropriate solutions for efficient BPM, such as Business Process Modeling initiatives [7]. In the next section, Business Process Modeling and how it facilitates organizational goals are further explained.

2.2 Business Process Modeling

The acknowledgment of BPM's significance for organizational success led to increased academic and enterprise interest on Business Process Modeling Techniques. BP Modeling refers to the representation of processes in a way that enables their analysis, improvement and automation. Given the complex nature of business processes nowadays, BP Modeling is perceived of outmost importance for providing specification and documentation to organizational procedures [8].

Process models are used to illustrate the main internal elements of business processes, including activities, sequence flow, dataflow and actors involved, as well as their relationships [6]. Their goal is to accurately describe a process, in a simple and understandable way, so that it facilitates communication among the stakeholders of the process. According to Vergidis [2], a business process design is "the representation of a business process depicting the participating tasks and their connectivity patterns that determine the flow of the process. The aim of the design is to capture, visualise and communicate a business process". Essentially, BP models offer an abstract view of a process, that enables decision-makers to disregard irrelevant complexities and focus on the important parts of the process, when examining a system [9].

Vergidis et al. [10] argue that a business process is as expressive and as communicative as is the technique that has been used to model it. Thus, the selection of the modeling technique is of critical importance. As a result, a variety of techniques and standards have been proposed in literature over the years, capturing different aspects of a business process. An overview of the most influential ones follows, with more attention directed to the BPMN standard, which is used for the current study.

2.2.1 Brief Overview of Business Process Modeling Techniques

As mentioned earlier, there is a multitude of BP modeling techniques specified in literature. Each techniques offers diverse capabilities for expressing business domains and give emphasis to different aspects of processes [7]. In this section, popular modeling techniques in academia, as well as in practice, are classified and briefly introduced.

In his work Vergidis [2] offers a classification scheme for business process modeling techniques, identifying three main categories: diagrammatic models, mathematical models and business process languages. The classification, as presented in [10] is visible in Figure 3.

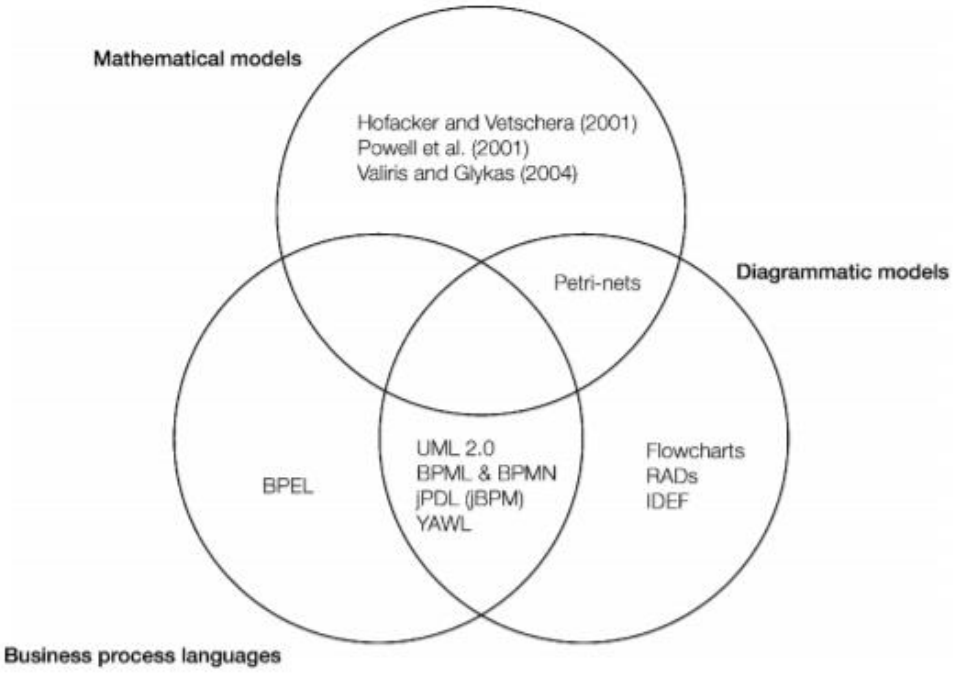

Figure 3: Classification of Modeling techniques [10]

The most common method of representing a BP is through a diagrammatic model, that has the ability to visually illustrate a process and present it in a simple and understandable way. Sometimes a diagrammatic model is enriched with executable features, that allows their classification as a business process language, e.g., BPMN. The diagrammatic category of modeling standards includes the most established methods of representing BPs and constitutes the primary focus regarding the context of this thesis.

Simple flowcharts were the first attempts to symbolize the flow of a Business Process. A flowchart consists of basic elements representing different types of actions or steps in a process, combined with arrows indicating the sequence of those steps. The basic elements comprising a flowchart are included in Figure 4 and an example of a simple process be found in Figure 5. Flowcharts offer a high-level approach of a process and are quite straightforward and easy to construct. However, they are not suitable for complex processes that include much information, since they incorporate basic facilities and provide limited room for detail [9].

Symbol	Name	Function
	Start/end	An oval represents a start or end point
	Arrows	A line is a connector that shows relationships between the representative shapes
	Input/Output	A parallelogram represents input or output
	Process	A rectangle represents a process
	Decision	A diamond indicates a decision

Figure 4: Basic flowchart symbols from smartdraw.com

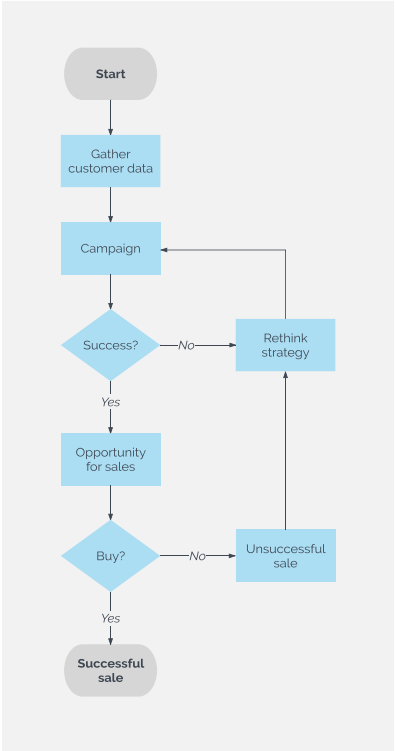

Figure 5: Simple Example of a Flowchart from lucidchart.com

Event-driven Process Chains (EPCs) are a modeling technique, that describe processes on a business level and is mostly aimed at domain experts. This type of modeling technique represents the control flow structure of the process as a chain of events and functions [11]. An EPC models consist of three basic element types: Events, Functions and Connectors. Functions refer to activities and are the main elements of a model. Events describe the circumstances of a function's execution and links function with each other. Connectors represent the logical relationships between the elements [11]. There are three connector types, i.e., AND (symbol ∧), OR (symbol ∨) and XOR (symbol ×). They can also be distinguished to join connectors, and split connectors [12]. The basic elements of EPCs are displayed in Figure 6. An example process modelled using the EPCs technique is displayed in Figure 7. In general, EPCs are criticized for offering an abstract representation of processes and allowing limited visual expressiveness [13].

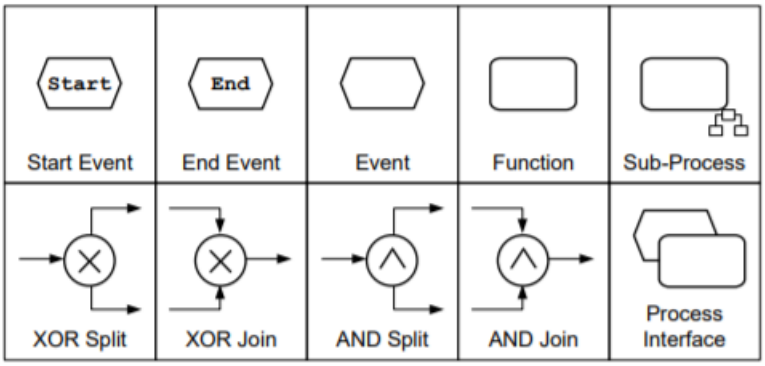

Figure 6: Basic elements of EPCs [13]

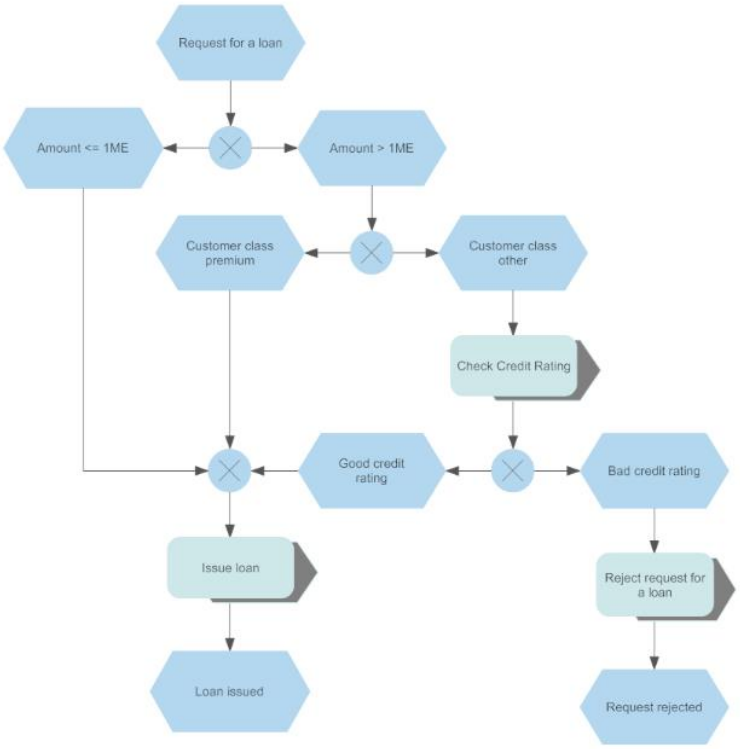

Figure 7: Example of a loan application process in EPC from smartdraw.com

Another modeling standard is the Unified Modeling Language (UML), a multi-purpose modeling standard, formally specified by Object Management Group (OMG) [14]. Initially, it was created to aid software development, but it evolved in a standard that covers different aspects of software structure and behaviour [15]. UML offers the capability of representing, apart from static models, dynamic (or behavioural) ones. A corresponding representation for a business process is a UML Activity Diagram, which focuses on the dynamic behaviour of the process, by expressing collaborations among objects and changes to the internal states of objects. UML activity diagrams are able to represent both computational and organizational processes (i.e., workflows), as well as the data flows interconnecting with the related activities [16]. Activity diagrams have been the recipients of criticism regarding the lack of syntax and semantics on some of the included constructs [16]. In Figures 8 and 9 the core elements of UML activity diagrams and an example process are displayed, respectively.

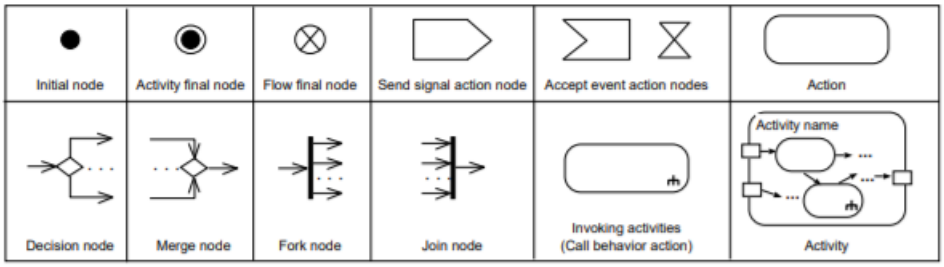

Figure 8: Basic Elements of a UML Activity Diagram [13]

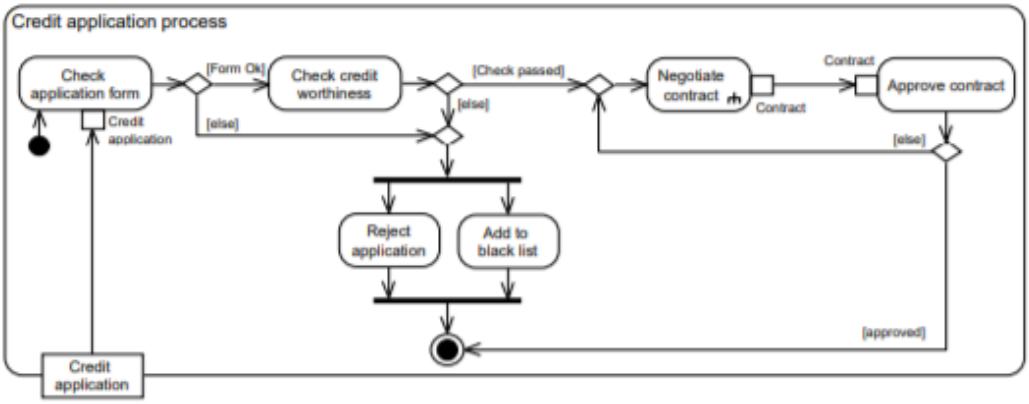

Figure 9: Example of a Credit Application Process UML Activity Diagram [13]

Another family of modeling techniques is the Integrated Definition (IDEF) suite of modeling languages, which first emerged in the 1970s from the U.S. Air Force Integrated Computer Aided Manufacturing (ICAM) program, with the purpose of increasing manufacturing productivity [17]. The suite includes methods that relate to the business process modeling field, such as a functional modeling method (IDEF0) and a process description capture method (IDEF3) [9], [17].

IDEF0 supports process modeling by offering a hierarchical decomposition of activities, which aids in avoiding unnecessary complexity. The two primary modeling elements of an IDEF0 model are functions, represented by boxes, and data and objects, represented by arrows. Activities can be described by their inputs, outputs, controls, and mechanisms (ICOMs), as depicted in Figure 10. An important advantage of this method lies in its ability to provide hierarchical decomposition of activities, essentially allowing the analyst to decide the level of detail required at any time. An example of the IDEF0 modeling technique is presented in Figure 11. IDEF0 main objectives include organizing the analysis of a system and assisting effective communication between the analyst and the customer through simplified graphical models [18].

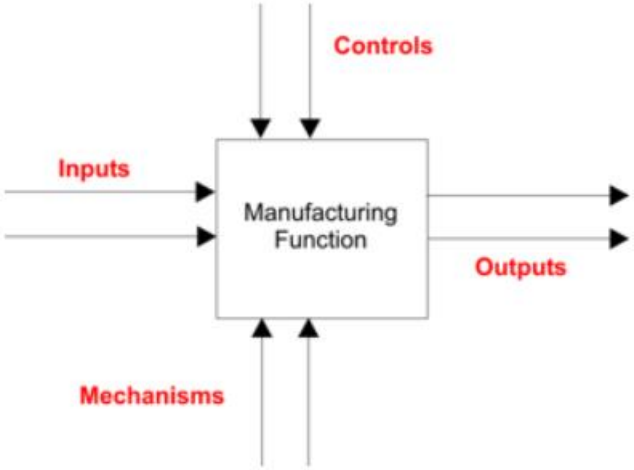

Figure 10: Basic syntax of an IDEF0 activity

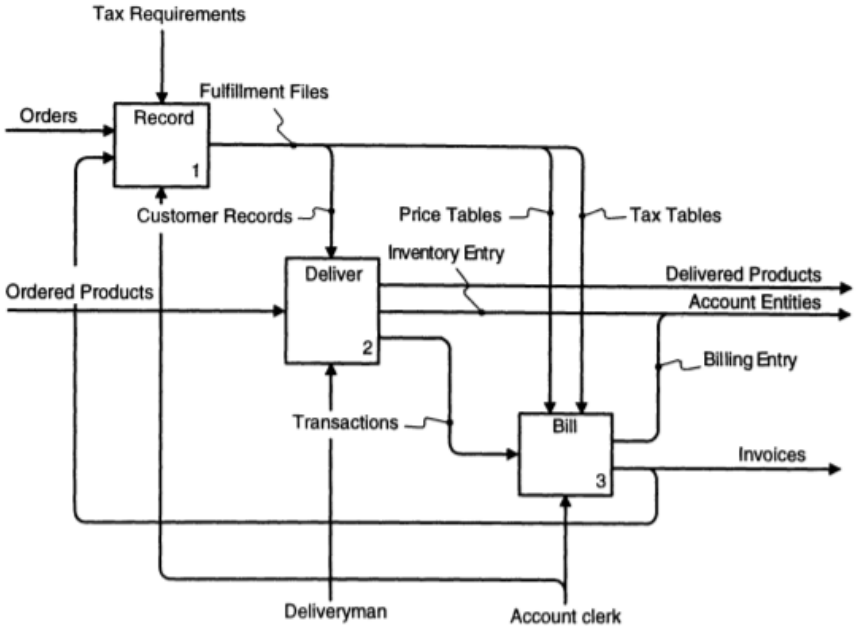

Figure 11: An Example of an IDEF0 diagram [17]

IDEF3 is a process description capture method that belongs to the next generation IDEF languages [19]. IDEF3 represents processes as a sequence of events or activities and is a scenario-driven process flow modeling technique based on the direct capture of relations between situations and events. Essentially, IDEF0 shows what is done within an organization or system, while IDEF3 shows how things work, acting complementary to IDEF0 [20].

The main construct of the IDEF3 is the Unit of Behaviour (UOB). UOBs may represent functions, activities and other processes based on the surroundings and context [19]. The basic syntax of a UOB is presented in Figure 12, along with the basic connections. The process in an IDEF3 diagram is organised within a scenario [19]. An Example of an IDEF3 diagram is depicted in Figure 13.

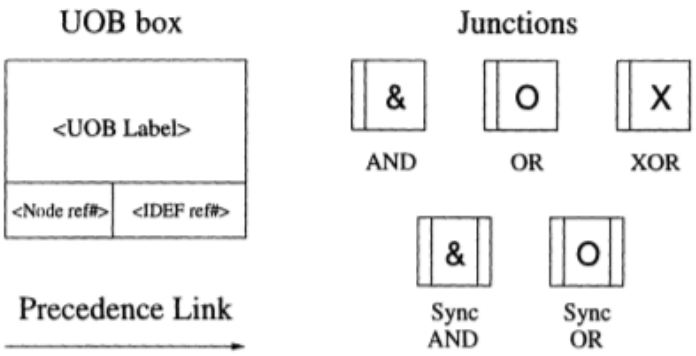

Figure 12: Basic Elements of the IDEF3 method [17]

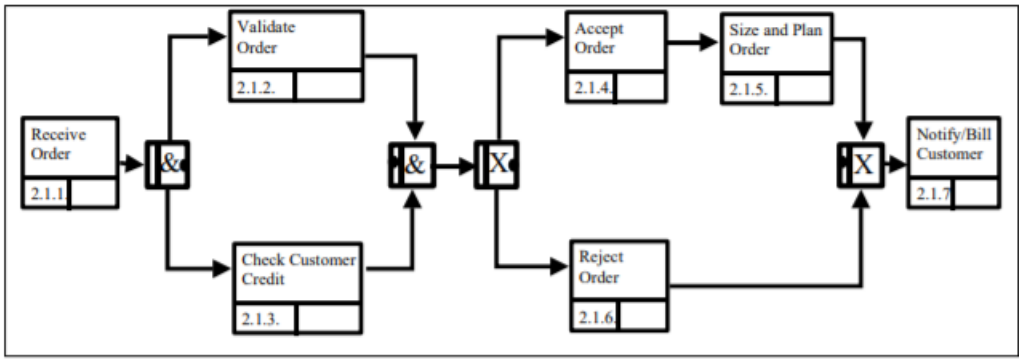

Figure 13: Example of an Order Process in IDEF3 [9]

With the aim to provide a complete notation that addresses all the mentioned weaknesses of previous techniques and is suitable for BPs, a new notation, namely BPMN, was introduced. Designed specifically for the sole representation of business processes, it quickly become very popular both in academia and enterprise world. A more detailed presentation of BPMN follows.

2.2.2 BPMN

The Business Process Model and Notation (BPMN) is a graphical standard created for the representation of business processes. The primary purpose of the technique is to address the communication gap between the users involved in the process design, implementation and monitoring, e.g., analysts, developers, business executives [21]. The initial version of the standard was introduced by the Business Process Modeling Initiative (BPMI) in 2004, while in 2011 the specification for BPMN 2.0 was released by Object Management Group (OMG) [22].

BPMN specifies a business process to a Business Process Diagram (BPD), comprising of many graphical elements. These elements allow the accurate representation of business concepts in a simple and comprehensible way. Essentially, BPMN is similar to EPCs and UML activity diagrams, but with a plurality of elements for the specific purpose of modeling business concepts. This is one main advantage of BPMN that has resulted in the wide adoption of the standard by both business professionals and IT experts [23].

A BPMN model consists of four main categories of graphical elements, that include different types of core elements [21]:

- Flow Objects
  - Events
  - Activities
  - Gateways
- Connecting Objects
  - Sequence Flow
  - Message Flow
  - Association
- Swimlanes
  - Pool
  - Lane
- Artifacts
  - Data Object
  - Group
  - Annotation

Flow objects are the core elements of a BPD. There are three main types of flow objects, namely Events, Activities and Gateways. An Event represents something that happens during a process and affects the flow of the whole process. Three main types of events exist on a model, Start, End and Intermediate events. These elements are symbolized by a circle, while a variety of different symbols is used to illustrate the nature of the events in more detail. Activities are the most important part of a process model, that embody the performance of actions to achieve an outcome, as is in most modeling techniques. They are symbolized by a rectangle and may represent a single Task or a collection of Tasks within a Sub-Process. Last, Gateways symbolize decisions that control the flow of a process. They represent forking, merging and joining of paths and are symbolized by a diamond shape. The most common Gateway types are Exclusive (XOR), Parallel (AND) and Inclusive (OR) Gateways, in accordance to most flowcharting techniques. More variations of the simple Gateways are specified to represent more complex connections.

The second category of elements comprising a BPD are Connecting Objects. Essentially, they are used to connect the Flow objects to create the basic structure of the process. There are three types of Connecting objects. The Sequence Flow, which is used to connect the Activities, the Message Flow, which show the flow of messages between the participants of the process, and the Association, which associates Artifacts to Flow Objects.

Swimlanes are used to organize the Activities to distinct categories. Two main constructs are used for this reason, Pools and Lanes. A Pool represents a participant of the process and is symbolized by a rectangle container that collects all the Activities performed by this participant. A Lane is an additional partition of a Pool, that may further separate the Flow Objects. Swimlanes are vital for visualizing a process with higher level of precision.

Finally, Artifacts offer added flexibility to the illustration of processes and enrich the modeling capabilities. Three main types of Artifacts are defined, which include Data Objects, Groups and Annotations. A Data Objects demonstrates the flow of data required for the execution of the Activities. A Group is used to visually show that a number of activities belong in the same group. Last, an Annotation serves the need to provide additional information in the form of text for better understanding of a model.

The core elements of BPMN are presented in Figure 14. For the representation of complex business notions and functions a multitude of symbols are utilized to offer more detail. There is a variety of different Tasks, Events and Gateways that explicitly define the behavior of a process. Several of those symbols are displayed in Figure 15 as an example of the expressiveness of the standard.

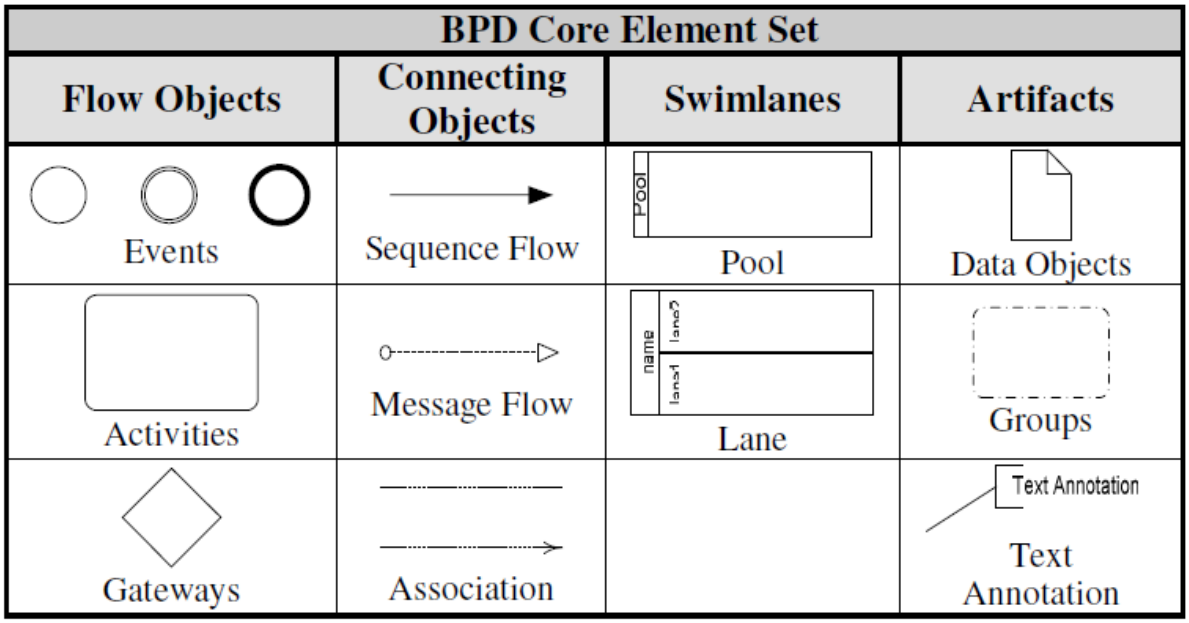

Figure 14: Basic BPMN Elements [24]

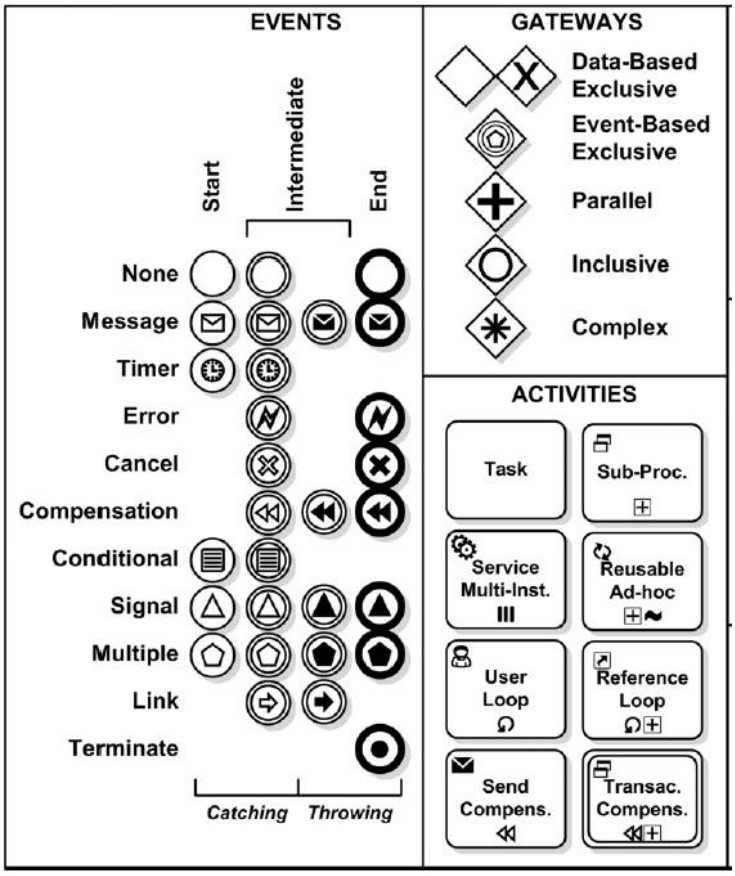

Figure 15: Different Types of Flow Objects in BPMN [25]

An important advantage of BPMN lies on its connection to Business Process Execution Language (BPEL). BPEL is, essentially, an XML based executable language for specifying business processes with web services. BPMN includes a partial mapping to BPEL, that allows the translation of BPMN constructs to executable code [25].

The plurality of elements is a key characteristic that supports the expressiveness of the BPMN standard compared to other modeling techniques. Given the popularity the standard has gained over the years, it is safe to assume it has accomplished its goal. However, researchers have pointed out that the enriched set of shapes of BPMN intensifies the need for formal specification and, at the same time, complicates the learning process of the standard [13].

Nevertheless, BPMN is a widely adopted standard by business process practitioners, that offers a multitude of capabilities for users. It is considered the state-of-the-art approach to business process modeling. Therefore, it is selected as the modeling technique to be used for the present research.

The processes that concern us for the premise of this thesis are modeled using the BPMN standard. An example of a BPMN model is displayed in Figure 21.

2.3 Business Process Redesign (BPR)

Business Process Redesign (BPR) is an important aspect of the Business Process Management (BPM) lifecycle [26]. Pertaining to the fundamental elements of BPR, Process Analysis (PA) utilizes measures to evaluate the performance of a process. The definition of such measures is strongly driven by an organization's desired objectives; the latter, essentially, implies the presence of subjectivity in the course of defining a measure, thus, a wide variety of them can be found in the literature [27]–[30]. PA offers insight into the "as-is" process, allowing the identification of process steps that necessitate improvement. During the redesign of the process, such issues are being resolved while, at the same time, opportunities for improvement are being discovered and implemented, generating the "to-be" process [31].

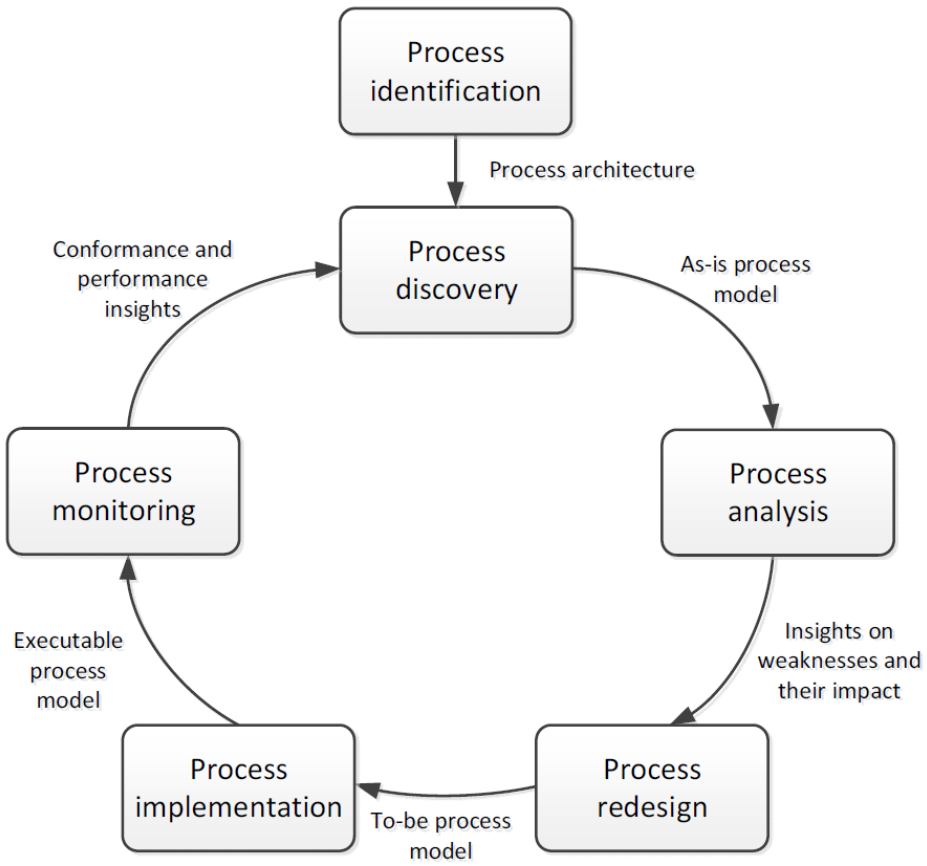

Figure 16: Business Process Management Lifecycle [26]

BPR goes hand in hand with innovation. Continuous changes in the business ecosystem, along with the rising competition, lead to more complex processes for organizations and establishes BPR as a key factor for improvement and success. Innovation proves an important motivator for enhancing the performance of processes, therefore, organizations try to incorporate innovative changes to satisfy the emerging needs and raise performance indicators [32].

In [33], Reijers and Mansar describe an evaluation mechanism for redesigning a business process, based in four performance dimensions: cost, time, quality and flexibility. The proposed mechanism, known as the devil's quadrangle, reveals a clear trade-off between the four dimensions, meaning that a positive effect on one of them may negatively influence another. The authors point out the significance of understanding this trade-off in order to achieve an actually improved process. This framework establishes an abstract approach in evaluating redesign attempts and deals with four dimensions that can be subjected to various interpretations, depending on the objective of each organization.

Redesign Heuristics, meaning rules and principles used to generate alternate redesign scenarios are classified in [26]. The Redesign Heuristics are divided into categories based on the main elements of business processes, i.e., customers, business process operation, business process behavior, organization, information, technology, and the external environment. Table 1 describes a full list of popular heuristic approaches for Redesign, as classified in [26]. Several approaches for each aspect of a business process are presented. Mostly deriving from successful industry practices and expert knowledge, these heuristics target the four performance dimensions of the Devil's Quadrangle. An organization's performance objective reflects on the heuristic that needs to be applied. The evaluation of the generated designs is a significant step in the Heuristic Redesign process, that may also lead to the reassessment of the goals set by an organization.

Category	Heuristic	
Customer Heuristics	*Control relocation*	Move controls towards the customer
	Contact reduction	Reduce the number of contacts with customers and third parties
	Integration	Consider the integration with a BP of the customer or a supplier
Business Process Operation Heuristics	*Case Types*	Determine whether activities are related to the same type of case and, if necessary, distinguish new BPs
	Activity Elimination	Eliminate unnecessary activities from a BP

	Case-based work	Remove batch-processing and periodic activities
	Triage	Split an activity into alternative versions
	Activity Composition	Combine small activities into composite activities
Business Process Behavior Heuristics	*Resequencing*	Move activities to their appropriate place
	Parallelism	Place activities in parallel
	Knock-out	Order knock-outs in an increasing order of effort and in a decreasing order of termination probability
	Exception	Design BPs for typical cases and isolate exceptional cases from the normal flow
Organization Heuristics	*Case Assignment*	Let participants perform as many steps as possible
	Flexible assignment	Keep generic participants free for as long as possible
	Centralization	Let geographically dispersed participants act as if they are centralized
	Split responsibilities	Avoid shared responsibilities for tasks by people from different functional units
	Customer Teams	Consider composing work teams of people from different departments that will take care of the complete handling of specific sorts of cases
	Numerical Involvement	Minimize the number of departments, groups and persons involved in a BP
	Case Manager	Appoint one person to be responsible for the handling of each type of case
	Extra Resources	If capacity is insufficient, increase the available number of resources
	Specialize	Consider deepening the skills of participants
	Empower	Give workers decision-making authority instead of relying on middle management
Information Heuristics	*Control Addition*	Check the completeness and correctness of incoming materials and check the output before it is sent to customers
	Buffering	Instead of requesting information from an external source, buffer it and subscribe to updates
Technology Heuristics	*Activity automation*	Consider automating activities
	Integral technology	Elevate physical constraints in a BP by applying new technology
External Environment Heuristics	*Trusted party*	Use the insights of a trusted party
	Outsourcing	Consider outsourcing a BP completely or parts of it
	Interfacing	Consider a standardized interface with customers and partners

Table 1: Redesign Heuristics [26]

Several attempts of presenting and evaluating BPR techniques have been made in research, nevertheless, redesign is still regarded as a severe challenge for organizations. Limited technical information is available for the Redesign process itself and limitations are recognized in literature for the existing Redesign methodologies [33]. In addition, it has been observed that researchers use a variety of labels to refer to BPR (e.g., Process Improvement, Process Reengineering), creating confusion among the scientific community. Studies also indicates the lack of information about redesign methodologies, regarding evaluation metrics, data collection and analysis [34].

The value of BPR is undeniable for organizations. Not only improving business processes is beneficial for business performance, but also constitutes an inevitable reality for organizations that seek to maintain their profitable status and continue to advance [26]. However, the altering of core business processes, may also pose a great risk for performance and customer satisfaction [35]. Business Process Modeling enables redesign initiatives by comprehensibly representing the process, allowing analysis and simulation in a conceptual level.

The representation of a process through an understandable and well-structured model is key when the intention is the application of Redesign Heuristics. Highly complex models tend to exhibit limited flexibility for change, meaning the implementation of Redesign practices becomes a difficult endeavor. The possible use of the Resequencing heuristic, for example, could prove difficult to recognize at first, and, ultimately, unfeasible to apply in a complex, poorly structured model. Moreover, the success of Redesign initiatives has been found to strongly relate to model quality [36]. For this reason, the evaluation of process models' complexity and general quality, through the scope of redesign, is considered of great importance. In many cases, normalization, which is further investigated latter, acts as a means to address the increased complexity of a process model [37], allowing room for Redesign.

2.4 Business Process Complexity

Business Process Modeling offers, as previously established, many advantages to organization that follow a process-oriented approach. The main advantage of BP modeling is that it simplifies the representation of processes, so as that they are easier to understand and maintain. In this way the communication between stakeholders is enabled, which is a key objective of BP modeling [38], [39].

Quality is a desirable property for BP models and is widely studied in literature [40]. BP models need to be simple, comprehensive and easy to understand. Low quality of BP models is associated with increased complexity, low understandability, low modifiability, limited redesign capabilities and error-proneness. In particular, complexity is considered an important quality characteristic and is often used interchangeably with the terms understandability, modifiability and maintainability [40]. Increased complexity levels have been proven to negatively impact BPR initiatives as well [36], therefore methods to measure complexity are deemed necessary.

Complexity, as a measurable property, appears in scientific literature, where several definitions exist [41]. According to Cardoso [42], the definition that is better suited to describe process complexity can be derived from IEEE Standard Computer Dictionary, and is defined as *the degree to which a system or component has a design or implementation that is difficult to analyze, understand or explain* [43].

Complexity of business processes is a high-level notion, that can be studied from many different aspects [27]. The complexity measurement of a process model, often associated with the model's understandability and modifiability, is a common objective in relevant research [40], [44]–[49]. Error-probability is another quality aspect related to complexity, widely investigated in literature [28]. Structuredness and its impact on complexity is an additional characteristic explored by researchers [45], [50], [51]. The common objective identified in research concerns the definition of appropriate complexity metrics that quantify the discussed aspects.

Generally, complexity is investigated without considering the modeling language used in each case. Generic metrics are defined, that take into account the main elements comprising a BP model. The size of a model is considered to impact the complexity level of a model and the presence of too many activities is a clear indication of increased complexity [27]. Loops and gateways have been found to add to the cognitive effort required to understand a model [52] and for this reason they are commonly used for the development of metrics [42], [53]. In addition, the interrelationships between the elements in a model are thought to reflect on complexity, thus connectivity between elements is examined by defined measures [38]. Lastly, the structure of a process model is a key characteristic that inspires the proposal of complexity metrics [28], [45]. The existing measures utilized to evaluate complexity for BP models are further analyzed in chapter 3, where related work is presented.

To summarize, complexity as a quality indicator has been investigated from many different angles in literature. There is a multitude of metrics proposed that evaluate complexity, associated with each researcher's perspective. For this reason, there are several terms used to describe complexity that interrelate to each other, e.g., understandability. For the premise of this thesis the term complexity will be used to describe the difficulty of understanding and modifying a process. The purpose of this thesis is related to a process' capability of Redesign, which is hindered by increased complexity. The next section delves into techniques to address increased complexity.

2.5 Addressing Business Process Complexity

Complexity of business process models, as defined in the previous section, is regarded as the root of many problems when it comes to analysing, maintaining, optimizing or redesigning processes. Despite the fact that there are multiple suggestions on how to use the BPMN notation correctly, modelers have been found to ignore them [54], leading to more complex representations of processes. Therefore, the task of managing excessive complexity of business process models is a challenge for many organizations.

Redesign, being a significant objective for many organisations, is greatly affected by model complexity and quality [36]. Reducing complexity is often a requirement, in order to apply redesign methods to process models [37]. Accomplishing that, increases the understandability of a process and its potential for transformation [55].

The process of employing techniques to achieve a behaviourally equivalent, but better structured and less complex model, is called normalization. Via normalization, it becomes feasible to reduce the complexity of a process and, at the same time, maintain its behaviour. In Business Process Modeling the same meaning can be conveyed using various syntactic structures, especially to enriched standards such as BPMN. Complexity reducing mechanisms, e.g., normalization, leverage on the formal structure of process model elements and their interrelationships [56].

Improving the structure and quality of a model is achieved by adhering to modeling guidelines and applying transformation rules, based on the previous principle. Essentially, when a model is normalized, practices and rules are applied to it, in the form of behaviourally equivalent structures, that replace problematic areas. Several studies discuss syntactical equivalent patterns in BPMN or introduce modeling guidelines to improve model quality [54]–[57].

In their study, which has been proven very influential in the field, Mendling et al. [57] introduce seven modeling guidelines for Business Process Modeling. These Guidelines are largely based on previous empirical research, performed by the same authors regarding understandability and error-probability of business process models [44], [58]. In essence, these guidelines are a set of recommendation that lead to the creation of better models or improve the quality of existing ones. These are presented below.

G1: Use as few elements in the model as possible

G2: Minimize the routing paths per element

G3: Use one start and one end event

G4: Model as structured as possible

G5: Avoid OR routing elements

G6: Use verb-object activity labels

G7: Decompose a model with more than 50 elements

Leopold et al. [54], following a similar reasoning, conducted a study on 585 BPMN 2.0 models collected from industry, in order to examine their adherence to fundamental quality principles. Even though the companies consisted of trained personnel, results revealed several violations of formal guidelines in practice. The researchers identified multiple errors, inconsistences and violations that urged them to propose five additional recommendations for improving BPMN modeling. The recommendations include:

Recommendation 1: Avoid implicit splits and joins

Recommendation 2: Provide tool support for proper model decomposition

Recommendation 3: Omit the throwing message event

Recommendation 4: Establish a centrally maintained glossary

Recommendation 5: Provide tool support for linguistic checks during modeling

More recently, in their work Corradini et al. [39] propose a set of 50 modeling guidelines, specific for the BPMN 2.0 standard. These guidelines derive from literature and are collected and adapted

by the authors in an attempt to offer a homogenous framework for their classification. This framework aims at facilitating the use of modeling guidelines from modelers, in order to improve model quality, especially with regard to understandability. A tool, called BEBoP (understandability verifier for Business Process models), is developed to automate the application of the guidelines in BP modeling and simplify their adoption by modelers.

The importance of following modeling principles has been recognized by many researchers. In order to facilitate their effectiveness in business process modeling, further steps towards standardization are required. Modeling languages, especially contemporary ones like the BPMN standard, allow multiple representations for the exact same process behavior. An important amount of research has been conducted on establishing ways of representing semantically identical meaning, while using syntactically different structure. Transformation rules and equivalent pattern repositories are employed by researchers to support complexity management [55], [56], [59], [60].

La Rosa et al. [55] discuss syntax modifications for the purpose of reducing complexity of business process models, that function regardless of the standard or tool used. The authors propose a set of patterns that improve structuredness and reduce process complexity in an abstract level, while at the same time exploring the effect these modifications have on several metrics found in literature. Table 2 presents collectively all pattern modifications proposed and offers their description and purpose as defined by the authors.

Pattern	Description	Purpose
Block-Structuring	Methods to structure a process model in blocks.	To improve understandability and maintenance through a simpler process model structure.
Duplication	Two model elements are duplicated if they point to the same conceptual definition.	To improve understandability and maintenance through a simpler process model structure
Compacting	Methods to remove redundant elements in a process model without loss of behavior. Elements that can be removed include redundant transitive arcs, superfluous gateways or duplicated tasks.	To reduce model size and thus improve the overall model representation, especially in large process models or models that have undergone a number of updates.

Vertical Modularization	This pattern captures features to decompose a model into vertical modules, i.e., subprocesses, according to a hierarchical structure.	To increase understandability of large process models by "hiding" process details into sub-levels. The maintenance burden of a process model (repository) is also decreased, as a change to a subprocess needs only be performed in one place.
Horizontal Modularization	This pattern captures features to partition a process model into peer modules.	To increase maintainability by breaking down a process model into smaller and more easily manageable parts, the ownership of which may be assigned to different users. Hence, to facilitate collaboration.
Orthogonal Modularization	This pattern captures features to decompose a process model along the crosscutting concerns of the modeling domain, which are scattered across several model elements or modules.	To enable a separation of concerns and distribution of responsibilities. To facilitate maintenance of individual, concern-specific process models.
Composition	This pattern describes features for constructing a consolidated process model from different disjoint modules. Modules may be organized vertically in a hierarchy (in this case Composition will flatten the hierarchy), or horizontally, or orthogonally (where each module represents a domain-specific concern).	To consolidate a family of interrelated modules into a single process model. The effect may be increased maintainability and understandability when there are too many or too small modules.
Merging	This pattern describes features for merging similar process models based on their commonalities, i.e., their identical model elements. The result is a single merged model.	To consolidate a family of similar process models (i.e., process variants) into a single "reference" process model, by removing redundancies.
Omission	Omission (aka Elimination) captures features to remove one or more elements from a process model and reconnect the remaining ones. It implies loss of process behavior.	To focus on specific parts of a process model while abstracting from the rest, due to the requirements of a project or specific audience.
Collapse	Collapse (aka Aggregation) describes features to synthesize multiple model elements into a single one of more abstract nature, where the distinction among the constituent elements is no longer relevant. It implies information synthesis.	To simplify a process model for a specific audience.

Restriction	This pattern captures features to restrict the syntax and semantics of a process modeling language, by removing modeling concepts from the language's meta-model.	To improve understandability and maintenance through a simplified process model.
Extension	This pattern captures features to extend the syntax and semantics of a process modeling language by adding new modeling concepts to the language's meta-model or refining the existing ones.	To obtain either a closer match to the concepts of a particular domain, or a straightforward transformation to executable software.

Table 2: Abstract Syntax Modification Patterns as proposed in [55]

In [56] Kluza and Kaczor insist on the importance of a standardisation for modeling languages and identify the need for a normalization process based on semantically equivalent patterns. More specifically, the authors provide multiple equivalences for BPMN structures, regarding all types of elements comprising a BPMN model. Starting from simple equivalent structures from the BPMN specification, more patterns concerning loops, control flow, gateways, activities, events and serialization methods are presented. The identification of equivalent patterns supports the possible replacement of complex parts of a process model and, combined with the application of modeling principles, is a first step towards a normalization attempt, according to the authors.

To enhance understandability and maintainability of a process Khlif et al. [60] propose a set of transformation rules that consider both structural and semantical information, relating to both the behavioural and organisational aspect of a process. These rules, inspired by the graph optimization domain, directly refactor a BPMN model to reduce control flow complexity. The application of the rules is enabled by a tool called EVARES, that employs a heuristic algorithm to decide which transformation rules should be applied to each model [61].

Combining complexity measures and modeling principles found in literature, Tsakalidis et al. [37], propose an assessment mechanism for BPMN models and their capability for transformation. After evaluating a model's complexity, a normalization step is proposed for highly complex models, employing well-established modeling guidelines. The main purpose of the mechanism is to reduce complexity in order to facilitate a transformation method, through improving structure.

Addressing quality issues in business process modeling is a notion widely investigated in literature. As previously discussed, lowering complexity or improving structure is mostly achieved by following well-established modeling guidelines. In essence, researchers incorporate a normalization phase to address excessive complexity and facilitate evaluation and transformation mechanisms [37], [49], [51], [60]. To accomplish that, standardization efforts and repositories of equivalent patterns are required as means to aid the normalization process. The latter may be performed either by hand or by using a tool. Some tools available for use are Apromore [59] , BPStruck [51], EVARES [61] and BP-Quality [49].

2.6 Cluster Analysis

This section serves as an introduction to the concept of cluster analysis, used throughout this study, since this is the method selected for analyzing the data, resulting from various complexity measures. The formation of the aforementioned data is further explained in chapter 4. By employing a clustering algorithm, it becomes possible to develop a method of assessing complexity for BP models and distinguish between complexity levels.

Cluster analysis is, essentially, the grouping of data points in categories called clusters, based on similarity. Each cluster contains instances considered similar to each other, and different to the instances included in other clusters [62]. Clustering is a very popular data mining technique, deriving from the field of statistics, with many applications in research. It is one of the main methods used for unsupervised learning, meaning there is no requirement for pre-existing labels on the data. Some of the fields employing clustering methods are machine learning, pattern recognition, bioinformatics and computer graphics. A very simple example of clustered data on the 2-dimentional space is displayed in Figure 17.

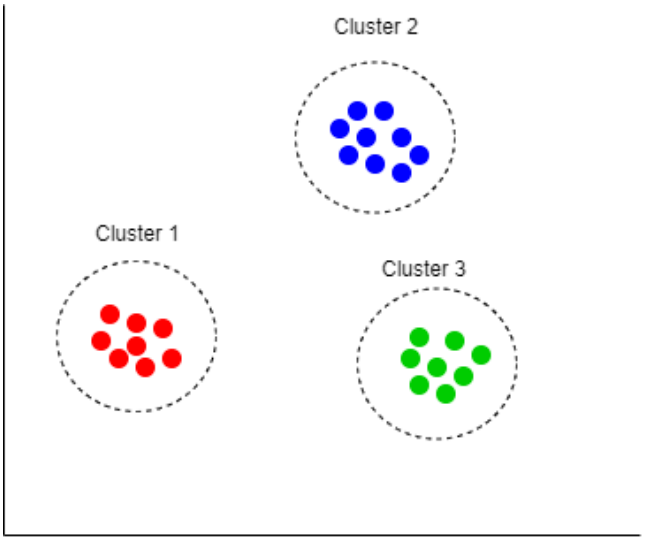

Figure 17: Example of Clustering

Several different approaches on clustering have been proposed over the years, depending on the understanding of the notion of cluster, which is not strictly defined [63]. Popular approaches on the issue involve connectivity models (i.e., models based on distance connectivity), centroid models (i.e., clusters represented by a central vector), distribution models (i.e., models based on statistical distributions) and density models (i.e., models based on dense areas of the data space). As a result, a variety of different algorithms are also developed in order to address this problem. Figure 18 depicts a categorization of clustering approaches and methods presented in [64]. The most popular types of clustering are partitional clustering, which is a way methods to partition data in a way that each instance belongs to one cluster, and hierarchical clustering, which organizes data in nested clusters with a hierarchical structure [65].

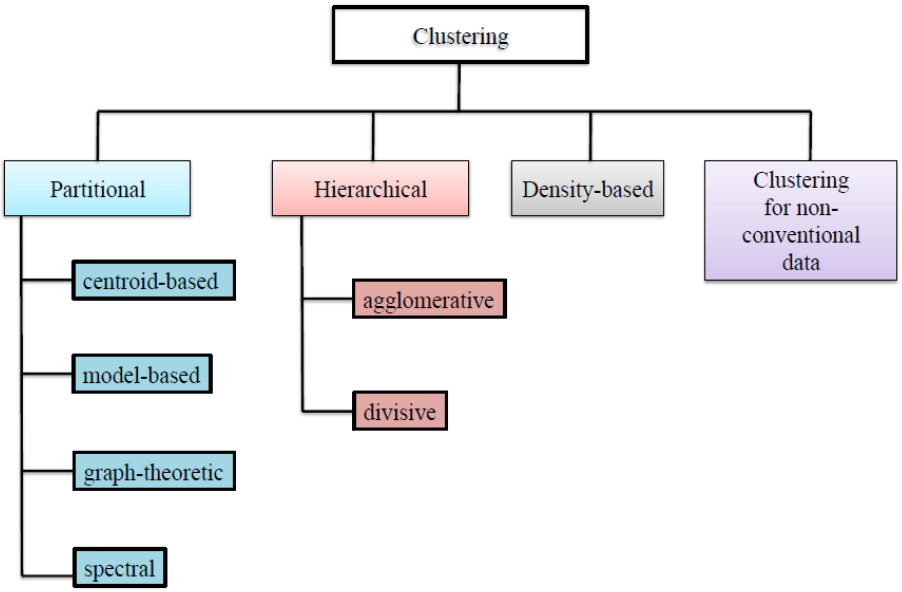

Figure 18: Categorization of clustering approaches [64]

To establish similarity (or dissimilarity) a proximity (or distance) measure needs to be formally defined. Without the use of a properly defined proximity measure a cluster analysis may prove incorrect or useless [66]. There is a variety of measures proposed in literature for interpreting the degree of similarity. According to [64], these can be classified in two main categories: Euclidean and non-Euclidean measures. More specifically, Euclidean measures are based on the notion of point in space and dimensions, while non-Euclidean measures consider other properties of the data to establish similarity. Table 3 summarizes some of the most popular proximity measures in practice as presented in [64].

Type	Measure	Calculation Formula
Euclidean	Euclidean distance	$d(x_i, x_j) = \sqrt{\sum_{k=1}^{n}(x_i^k - x_j^k)^2}$
	Manhattan distance	$d(x_i, x_j) = \sum_{k=1}^{n} \lvert x_i^k - x_j^k \rvert^2$
	Minkowski distance	$d(x_i, x_j) = [\sum_{k=1}^{n}(x_i^k - x_j^k)^p]^{1/p}$
non – Euclidean	Jaccard distance	$d(x, y) = 1 - \frac{\lvert x \cap y \rvert}{\lvert x \cup y \rvert}$
	Cosine similarity	$d(x, y) = \theta = arccos\ \frac{x * y}{\lvert x \rvert \lvert y \rvert}$

Table 3: Popular Proximity Measures [64]

Another very important step for successful cluster analysis is data pre-processing. The pre-processing of data is a key factor in the implementation of clustering methods and may lead to a more accurate solution. Common pre-processing tasks include data reduction, that removes irrelevant attributes or instances from the analysis, removal of noise/outliers and data normalization techniques. Data normalization proves especially important when it comes to distance-based approaches [64].

There is no absolute way to determine which algorithm works the best. This decision depends on various factors including the purpose of the analysis, the dataset used and the desired outcome. In addition, selecting the parameters of the model, such as the distance measure or the number of clusters, is also of great importance for the analysis results. It is important to point out that cluster analysis is not an absolute procedure followed, but mostly involves trial and error, until the outcome of the analysis serves its purpose. For the purpose of the present thesis, the implementation of a centroid-based algorithm, namely K-means, was deemed suitable and is further explained below.

2.6.1 K-Means Algorithm

K-means is one of the most popular algorithms used in unsupervised machine learning. It falls in the category of partitional centroid-based algorithms, which means that clusters are represented by a central vector of means. The idea behind K-means clustering was first proposed more than 50 years ago [67] by several researchers [68], [69]. Many variations of the standard algorithm have been proposed in literature ever since.

A basic requirement for the implementation of K-means is the specification of the number of clusters. The algorithm does not have the ability to identify the number of clusters the data need to be partitioned in. This process is a trial and error procedure and depends mainly on the purpose of the analysis. The standard algorithm, visible in Figure 18, follows the below steps [70]:

1. Specify the number of clusters (K)
2. Randomly select K points as initial cluster centroids
3. Assign the data points to the closest centroid
4. Recompute the centroids of the clusters
5. Repeat the process from step 3
6. Stop when the clusters do not change

1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

Figure 19: Simple K-means algorithm [65]

The specification of the distance measure utilized by the K-means algorithm is of great importance. Several measures can be used by the algorithm to determine proximity, with the Euclidean distance being the most popular method used. The evaluation of the clustering result is performed through

a metric called Sum of Square Errors (SSE). Essentially, for every point the error is defined as the distance of the data point from the nearest cluster. The main objective of K-means is to minimize this measure [67]. By squaring the errors and then summing them, we calculate SSE. Let m_i be the cluster centroid and x the data point. The formula for this metric is:

$$SSE = \sum_{i=1}^{K} \sum_{x \in Ci} dist^2 (m_i, x)$$

Despite being extremely popular in data mining [67], there are several issues recognized by researchers regarding the algorithm. Three main issues are often raised [71]:

- The algorithm requires the number of clusters as an input parameter, which essentially assumes that the number of clusters are known beforehand
- The algorithm is sensitive to the initialization method used, meaning that different initial points may lead to different clustering results
- It is possible for the algorithm to converge to local minima

All things considered, K-means is an extensively used algorithm in data mining and machine learning fields. Its main advantages lie in its simplicity and straightforwardness, combined with ease of implementation and effective performance [67], [72]. Several modifications and adaptations of the algorithm have been proposed in literature and continue to emerge until today, while at the same time its popularity is wide among practitioners.

2.7 Summary

Business Process Modeling is an important ally for organisations seeking to remain competitive, offering them much needed control in a rapidly changing environment. Business Process Redesign is one of the many ways an organisation benefits from employing a process-oriented strategy. Successful BPR initiatives have been proven to reduce costs and raise performance indicators. However, their application is a challenging task, regularly impeded by low model quality and excessive complexity. Therefore, a method to assess business process model complexity through appropriate measures is required. Complexity evaluation is regarded as a significant first step towards the support of Redesign approaches. Process normalization techniques, including

modeling guidelines and transformation rules, facilitate complexity management. The next chapter presents related research on complexity measurement and evaluation.

CHAPTER 3

Related work

This chapter discuses work found in literature involving complexity measurement for business process models. Initially, fundamental research on complexity metrics is presented in detail, providing a complete overview of the field. Moreover, work regarding threshold establishment and validation for metrics in recent years is examined. The chapter concludes with main observations for the field of complexity measurement and threshold extraction and a summary of complexity metrics considered suitable for the current research.

3.1 Complexity Metrics

Measuring the complexity of Business Processes is a task that can be examined from many different standpoints [27]. Nevertheless, the quantification of the notion of complexity is almost always performed with the employment of defined metrics. A multitude of studies in literature concern the definition and proposal of appropriate metrics to measure complexity or aspects of it [24], [27]–[30].

Business Process Complexity metrics proposed in literature, mostly constitute indicators of understandability, maintainability and error-proneness of a process model [41], [73]. At the same time the majority of the metrics for business process models existing in literature are adaptations of software complexity metrics [41], [74].

Cardoso identifies four main complexity perspectives: activity complexity, control-flow complexity, data-flow complexity, and resource complexity [42]. The study presented in [41] concludes, that the majority of complexity metrics proposed in literature, mostly use control-flow complexity as their main objective. In general, control-flow elements are considered to contribute to the complexity of a model and are extensively used in the construction of metrics.

Drawing an analogy between Software and Business Processes, Cardoso et al. [27] propose the Number of Activities metric (NOA), which calculates activity complexity, and was inspired by lines-of-code (LOC) metric [75]. Taking into account the control elements of process models the Number of Activities and Control-flow elements (NOAC) metric is introduced for well-structured models, along with the Number of Activities, Joints and Splits (NOAJS) metric for not well-structured ones.

Albeit these complexity metrics are useful and simple to calculate, it is highly important to complement other forms of complexity. In [53], Cardoso introduces the Control-Flow Complexity (CFC) metric, which also borrows techniques from the software engineering branch of software metrics, namely McCabe's cyclomatic complexity [76]. This metric evaluates the complexity of XOR-split, OR-split, and AND-split constructs and aims to measure the impact of control-flow elements on the perceived complexity of a process.

In their work, Cardoso et al. [27] adapts more software complexity measures to the business process domain. For estimating process length, volume and difficulty the author introduces a set of measures based on the Halstead metrics for software complexity, the Halstead-based Process Complexity (HPC) metrics, namely Process Length, Process Volume and Process Difficulty. They are calculated with the help of primitive measures deriving from source code, adapted to reflect on business processes. Their main purpose regards the quantification of rate of errors and maintenance effort, in a simple and generic way. Several of the metrics, adapted from the software engineering domain and proposed by Cardoso et al. [27] are summarized in Table 4.

Metric	Description
NOA	Counts the number of Activities
NOAJS	Counts the number of Activities, Joins and Splits
NOAC	Counts the number of Activities and Connectors
CFC	Measures Control flow complexity accounting for split constructs
HPC	Estimates process length, volume and difficulty

Table 4: Complexity metrics adapted from the software domain [27]

In his work, Mendling [28] draws from network analysis and graph theory, presenting various structural metrics and dividing them into the categories *size*, *density*, *partitionability*, *connector interplay*, *cyclicity*, and *concurrency*. Apart from the NOA metric for the calculation of a process's size, Mendling also defines Diameter (diam) for process models as the length of the longest path from a start node to an end node [28]. Various density metrics adapted by the same author, provide information regarding the relation between arcs and nodes in a model. The Coefficient of Connectivity (CNC) metric, earlier proposed by Latva-Koivisto [77] with the purpose of measuring the degree of complexity of a critical network, is used in regard to process models to calculate the ratio of arcs to nodes. A similar approach is used for the Density (Δ) metric, which "refers to the numbers of arcs divided by the maximum number of arcs for the same nodes". Additionally, the Average Connector Degree (ACD) in a process model demonstrates the number of nodes a connector is in average connected to and Maximum Connector Degree (MCD) the maximum number of nodes for a connector [28], [58]. A high value of density metrics is an indicator of complexity and low understandability of a model [51]. A number of metrics proposed by Mendling, regarding size and density, are displayed in Table 5.

Category	Metric	Description
Size	s_N	Number of nodes in a process graph
	Diameter	The length of the longest path from a start node to an end node
Density	Density (Δ)	The number of arcs divided by the number of the maximum number of arcs for the same number of nodes.
	CNC	The ratio of arcs to nodes
	ACD	The number of nodes a connector is in average connected to
	MCD	Maximum degree of a connector

Table 5: Complexity metrics deriving from graph theory [28]

An additional aspect of complexity, directly deriving from software complexity, is coupling. Coupling quantifies the interconnections between elements of a model. Researchers, inspired by the software metrics for coupling, have proposed a weighted Coupling metric (CP), adapted to business process models [78]. Essentially this metric involves the measurement of all types of connections (e.g., AND, OR and XOR) between the elements in a process models and expresses the ease of understanding and maintaining a model.

In [38], Vanderfeesten et al. define the Cross-Connectivity metric (CC), in an attempt to quantify how easy it is for a user to understand the interplay between all elements in a process model. This approach, though difficult to calculate, considers the cognitive effort required by a user to understand a model, as error-proneness in a model is directly affected by its understandability.

Based on previous research on process model complexity metrics by Rolón et al. [24], Reynoso et al. [29] formally define a set of complexity and understandability measures for business process models. Contrary to previous work, which mostly focuses on Event-driven process chains (EPCs), the metrics proposed in [29] are specifically defined for the Business Process Modeling Notation (BPMN) standard for business process modeling. By using basic measures to count each BPMN element in a model and combining them, more complex metrics to measure understandability and modifiability derive, e.g., the Connectivity Level between Activities metric (CLA), which is similar to the CNC metric previously discussed.

Likewise, Kluza and Nalepa [30] proposed the Durfee Square Metric (DSM) and Perfect Square Metric (PSM) specifically constructed for BPMN models, aiming to include, not only the number of elements in a process model, but also their variety in measuring its complexity. Both metrics are simple in their calculation and comprehensible enough for the modelers, according to the authors. More recent work also concentrates on models composed using the BPMN standard, which offers different artifacts and structures for the representation of business models and is widely used within the BPM field today [39], [47], [79].

It becomes apparent that a model's complexity cannot be directly determined by only one type of metric [47]. Mendling implies that metrics should be interpreted in relation to other metrics, by pointing out several limitations, e.g., that models with more activities can be more understandable or that the density metric should consider the size of a model so as to be sufficient [28]. Moreover,

Cardoso argues that the CFC metric should be used collaboratively with size metrics for better evaluation of a model's complexity [53]. Overall, only a few attempts of combining complexity metrics or defining thresholds for model classification appear in literature.

In accordance with the aforementioned approaches regarding the concept of complexity in business process modeling, it can be observed that the development of appropriate complexity measures is somewhat dependent on the researcher's view. The latter is further encouraged by the lack of strictness in the assimilated definition of process complexity. Essentially, the perception of each researcher, in reference to a model's complexity, determines the metrics selection and development. Table 6 summarizes several metrics for complexity measurement of business process models proposed in literature.

Metric	Purpose	Source
NOA, NOAC, NOAJS	Measure the activity complexity of a process model	[27]
CFC, HCM	Measure the control-flow complexity of process model	[27], [53]
CNC	Quantify the interrelationships between elements in a process model	[28], [77]
Diameter	Relates to the size of a process model	[28]
Density	Estimates how dense a process model is to measure error probability	[28]
ADC, MDC	Calculate the degree of connectivity between elements to measure error probability	[28], [58]
CP	Quantifies the ease of understanding and maintaining a process in a process model	[78]
CC	Represents the connectivity between elements to express the understandability of a process model	[38]
CLA	Measures the connectivity between activities in a BPMN model	[29]
DSM, PSM	Calculate the number and variety of elements in a BPMN model	[30]

Table 6: Summary of popular business process complexity metrics

3.2 Threshold Definition

Despite the variety of metrics proposed in literature over the last decade, only a small percentage of them are adequately validated [41], [73], [74]. At the same time, little work has been done regarding the definition of thresholds or reference values for the proposed metrics. The task of distinguishing between good or bad values for a given metric is, in most cases, performed by the experts of the field, based on their experience and established industry practices. Relative research, concerned with the assignment of reference values to quality and complexity metrics, is presented next.

Emphasizing the need for well-defined thresholds for complexity measures [80] Mendling et al. utilized logistic regression and an adaptation of the ROC curves method to determine and evaluate thresholds for certain structural metrics. For systematically extracting the thresholds and validating them a sample of 2003 EPCs was used. The derived thresholds are useful for the prediction of errors in process models and provide quantitative support to modeling guidelines proposed by the same authors [57].

In [46], Sánchez-González et al. empirically evaluate certain structural metrics, including the CFC and ADC metrics, and provide thresholds for classifying models in regard to understandability and modifiability. Subsequently, they developed the Gateway Complexity Indicator (GCI), a combination of the five independent structural metrics previously evaluated, into a weighted new metric, with the purpose of aiding in decision-making and providing guidelines for novice modelers. Later research highlights the definition of thresholds for complexity metrics as a significant mechanism of detecting non-suitable models, in regard to understandability and modifiability [81].

With the aim to improve quality characteristics of business process models, Fernandez – Ropero et al. [82] introduce a mechanisms that enables the detection of appropriate refactoring operators. These operators, obtained from relevant research, are used in conjunction with quality measures, to increase understandability and modifiability of models. A heuristic method of defining intervals for the selected measures is applied, in order to identify which refactoring operator needs to be activated. The idea of leveraging on quality metrics to facilitate the application of rules reveals great potential, however the values defined as thresholds by the authors are based solely on observation of real business processes and lack validation.

In their work Yahya et al. [49], extract threshold values for a vast amount of complexity metrics from literature, associating them with comprehensibility and modifiability of BP models. The aim of this study is to propose a framework for business process improvement, using complexity evaluation measures proposed in literature in conjunction with quality improvement techniques, deriving from literature as well. The definition of the thresholds is performed through experiments on 50 models extracted from the SOA-based Business Process Database, and aids both the identification of weaknesses for models, as well as the assessment of the framework itself.

An evaluation mechanism proposed in by Tsakalidis et al. also leverages on widely used complexity metrics from literature. The proposed mechanism regards complexity as a major obstacle for the transformation of business process to a Directed Acyclic Graph (DAG) and evaluates it with the support of threshold values for a number of metrics. These thresholds, if exceeded, reveal a model's need for Normalization to reduce complexity and support the transformation of the model to a DAG.

Generally, despite the substantial amount of research on complexity metrics and their definition, attempts to define reference or threshold values for these metrics are significantly lacking. The few approaches present in literature are commonly based on expert opinion. Furthermore, research on combining metrics to offer a thorough evaluation of quality and complexity is limited. Considering the above, the methods proposed in this thesis offer a holistic approach and aim at addressing complexity assessment in a way that enables the recognition of appropriate reference or threshold values.

3.3 Summary

In conclusion, there is a variety of metrics proposed in literature, associated with different aspects of a BP model (i.e., size, control-flow, structure), that offer ways to quantify a model's perceived complexity. However, it becomes apparent that a single value of one metric conveys limited meaning regarding the evaluation of a model's overall complexity. This task is particularly difficult, when relevant threshold values have not been defined for the given metric, which is commonly the case. Evidently, to provide a holistic evaluation for a BP model, based on the several aspects of complexity, a combination of metrics is required. Towards this end, the present research attempts to leverage proposed metrics and introduce a combination of them, that enables the identification of highly complex models. For this reason, cluster analysis is chosen, as the suitable method to

combine a set of the selected metrics and discover reference values. The focus of this study is to assess complexity through the scope of Redesign, meaning that the selection of the metrics is related to this end. The overview of our research is presented in the following chapter.

CHAPTER 4

Methodology

The present chapter offers an overview of the path followed to accomplish the objectives of this thesis. Initially, an overview of the steps taken is provided. Subsequently, the justification behind the selection of measures from literature is presented and the selected metrics are described and evaluated in detail. The rest of the chapter concerns the pre-processing of the database and the calculation and formation of the required data, offering also details regarding the parameters of the implemented algorithm.

4.1 Overview of Methodology

The first step is the selection of the metrics to be included in the analysis. Following the literature review on complexity measurement, the metrics that best serve the purpose of this thesis are selected. More details on the selection criteria and the metrics are discussed later. In order to evaluate BP complexity, a machine learning technique is employed, namely cluster analysis. As discussed in chapter 2, Clustering is an unsupervised method of grouping data based on similarity. The algorithm used for the current research is K-means, a popular method that groups data by calculating centroid values for a pre-determined number of clusters.

Figure 20 displays an overview of the methodology followed to extract reference values for the assessment of complexity. First, a Data Preprocessing step is required to develop the final dataset that will be used in the analysis. Next, the Clustering Phase follows, determining the centroid values that represent each cluster. Last, by analyzing the resulting clusters, i.e., through their respective centroid values, reference values for the selected complexity metrics derive, facilitating the identification of complex models.

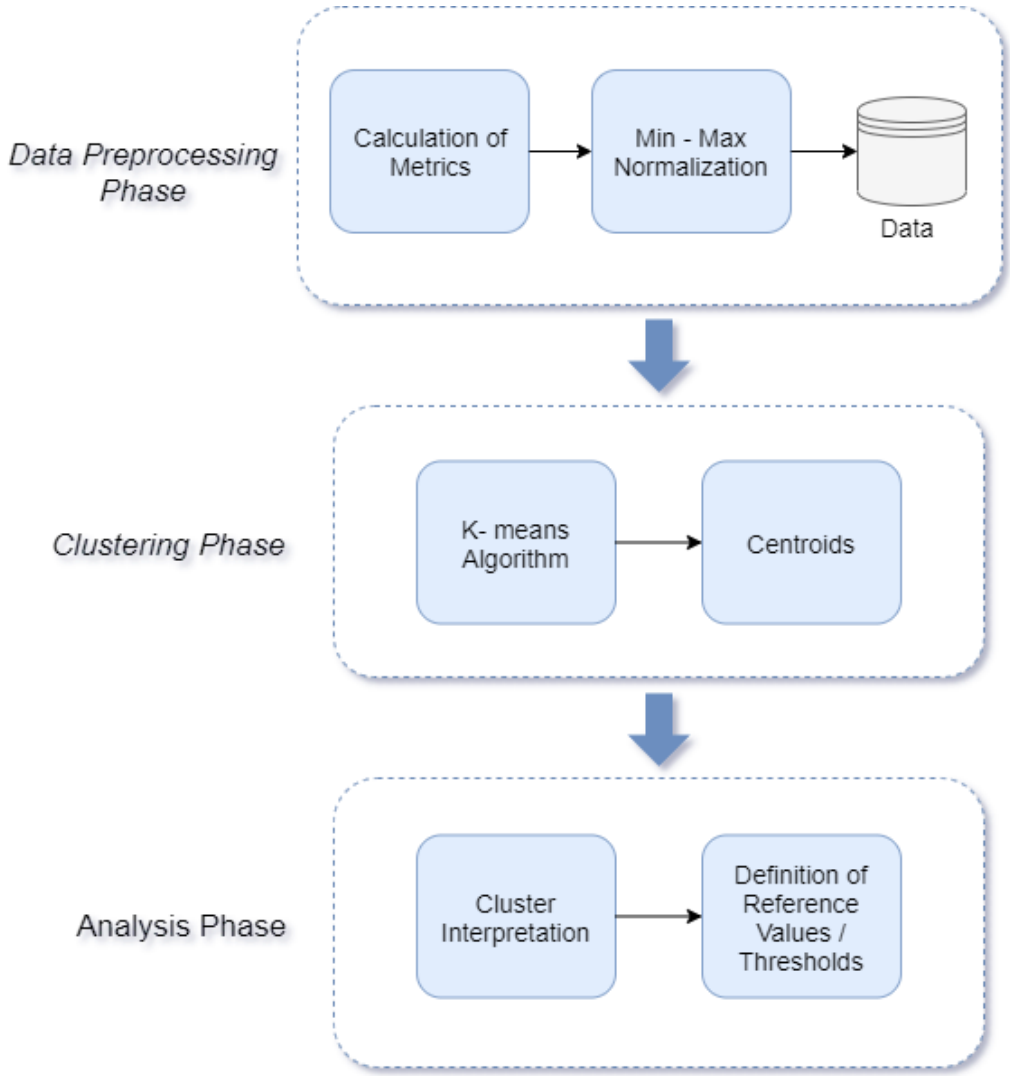

Figure 20: Methodology

4.2 Selection of Metrics

The complexity of a process model acts as a red flag for the process itself. Excessive complexity raises issues, regarding the understandability and maintainability of the process, which, most commonly, lead to higher costs, longer throughout times and low customer satisfaction; often, even simple processes happen to be represented by a complex model [53]. Avoiding such outcomes is top priority for stakeholders.

Business Process Redesign, as a key step in the Business Process Lifecycle, plays a vital role in addressing the above issues [26]. However, achieving process improvement entails lowering the overall complexity of the process, making it possible for Redesign heuristics to be applied. Increased

complexity warns for possible difficulty in the application of Redesign initiatives and, at the same time, reveals the need for implementing normalization practices. In this section, the measures selected, for the purpose of evaluating business process complexity to demonstrate the capability for Redesign, are described.

4.2.1 Selection Criteria

Given the main focus of this study, which is related to measuring the complexity of business processes for Redesign, it is only logical to concentrate on measures that are associated with this view of process modeling [83]. The core idea is to utilize already existing metrics in literature, that have been found to adequately communicate the degree of complexity a process entails.

The three aspects of a model's complexity, that have the strongest presence in research, are activity complexity, control flow complexity and structural complexity. Evidently, the three aspects combined constitute a holistic approach on complexity measurement, covering the most important elements of a model, as explained below.

Activity complexity conveys the important information of a process' size. The larger the process the more complex the model. Essentially, high values of activity complexity metrics reveal the need of simplifying a model through normalization, e.g., by the elimination of tasks that add zero value from the customers point of view [33]. In addition, activity metrics are easy to calculate and may prove even more insightful when examined in conjunction with control-flow metrics [53].

Control flow is one of the main factors influencing comprehensibility and, in extension, complexity of processes. The elements clearly associated with the control flow of a BPMN model are Activities, Sequence flows and Gateways. The impact of Gateways (especially of the loops created by them) on the cognitive effort of the analyst is considered to be significant and directly increases complexity [52].

Finally, well – structuredness of a model is a core modeling principal, which commands each split element to be matched by a joint element of the same type [57]. The importance of well – structured models for the avoidance of errors and overall quality of a model is also explained in [50]. According to Dumas [51], better structuring a model results in an undeniable increase in the number of elements in that model. Nevertheless, it decreases other, more descriptive metrics, that

relate to the models density and control flow complexity, ultimately improving the models understandability [51].

Considering the acclaimed impact of size, control flow and structural elements of a model on its overall complexity, as described above, it is deemed appropriate to focus on metrics that express these types of complexity. For these reasons, the metrics selected for measuring the total complexity of BPMN models are presented in Table 7.

Further characteristics of the above metrics compelled their selection for this study. Firstly, their popularity and adequate theoretical and empirical validation is well established by many studies [41], [44], [58], [77], [84]–[86]. Furthermore, these metrics are not restrictive regarding a modeling type, in the sense that they are defined in a generic way that covers most modeling standards and languages. Nevertheless, they are applicable to models represented in BPMN 2.0, which is the modeling standard this thesis investigates. Last, all selected metrics are defined by simple mathematical operations; hence, their calculation is a quite straightforward procedure. Following, a more detailed description of the metrics is presented.

Complexity Type	Metric
Activity	NOA
	NOAJS
Control flow	CFC
Structural	CNC
	Density

Table 7: Selected Metrics

4.2.2 Description and evaluation of selected metrics

The Number of Activities (NOA) and Number of Activities, Joints and Splits (NOAJS) activity complexity metrics, proposed by Cardoso et al., are inspired by the software complexity domain, as previously mentioned [27]. Both metrics stand as a representation of a process' size, by simply

counting the activities in a process. Their distinction is the fact that NOAJS includes all Split and Join constructs in a process, while NOA counts only the number of activities. It is important to note that NOAJS is selected instead of the NOAC metric, since it is proposed for not well-structured models, like most ones residing in the database to be examined. It is an undeniable fact that complexity rises when a process increases in size. As a result, high values of these metrics reveal a complex model. Frequently, though, the above metrics function, mainly collaboratively, as a point of reference for comparison purposes with other metrics [53].

Control-Flow Complexity (CFC) is regarded as a fundamental metric to evaluate a process model's complexity in relation to control flow. Cardoso introduced this metric in [42]; since, it has been widely studied and empirically validated in research [28], [53], [85]. The measure itself is inspired by McCabe's cyclomatic number [76] and aims to quantify the cognitive effort of understanding the multiple states of a process after the occurrence of each type of Split. Each formula computes the number of states that can be reached from each one of the three split types. The $CFC_{XOR-split}$, $CFC_{OR-split}$, and $CFC_{AND-split}$ functions are calculated as follows:

- $CFC_{XOR-split}(a)=fan-out(a)$, where 'a' is a XOR-split activity. The control-flow complexity of XOR-splits is determined by the number of branches that can be taken.
- $CFC_{OR-split}(a)=2^{fan-out(a)}-1$. The control-flow complexity of OR-splits is determined by the number of states that may arise from the execution of an OR-split construct.
- $CFC_{AND-split}(a)=1$. For an AND-split, the complexity is simply 1.

The higher the value of $CFC_{XOR\text{-}split}$, $CFC_{OR\text{-}split}$, and $CFC_{AND\text{-}split}$, the more complex is the design, as modelers have to consider all possible states between each control-flow construct and the associated outgoing transitions and activities. Mathematically, the Control-Flow Complexity metric is additive which implies that the CFC of all the split constructs should be added for the calculation. The absolute control flow complexity of a business process P is:

$$CFC_{abs}(P) = \left(\sum_{i \in (XOR-splits\ of\ P)} CFC_{XOR-split} i \right) + \left(\sum_{j \in (OR-splits\ of\ P)} CFC_{OR-split} j \right) + \left(\sum_{k \in (AND-splits\ of\ P)} CFC_{AND-split} k \right)$$

It should be noted that huge differences in CFC may occur for models with the same structure, but with different gateway labels, despite being equally understandable [28]. Preferably, the CFC metric should not be used in isolation to effectively evaluate the overall complexity of a business process, since it only analyzes a process from the control-flow point of view. A combinatorial approach with other complexity measures is considered more beneficial.

Structural metrics constitute the last type of measures included in this work. Coefficient of Network Connectivity (CNC) and Density, both exploit the notion of connectivity between elements to quantify structural complexity. Inspired by Graph theory, the given metrics explore the relation of the number of Arcs to the number of Nodes [77]. High values of these metrics reveal a dense model, which indicates increased complexity and error probability [28]. The CNC of a process model is a simple measure to calculate and understand and it is defined as the ratio of Arcs to Nodes.

$$CNC = \frac{Arcs}{Nodes}$$

The definition of Density follows the same rationale as the CNC. Essentially, Density refers to "the number of Arcs divided by the number of the maximum number of Arcs for the same number of Nodes" [28].

$$Density = \frac{Arcs}{Nodes * (Nodes - 1)}$$

The density metric constitutes a straightforward indicator of high complexity of a process model; however, it is strongly correlated with the size of a process [58]. Meaningful comparison exists between business process models with the same number of Nodes.

4.2.3 Calculation in a BPMN example

As previously mentioned, some of the selected metrics were introduced in a generic way that is applicable to many modeling languages. The BPMN standard comprises of a variety of elements, often absent in other modeling types, and, therefore, several problems may occur. The variety of elements present in a model, for instance, might confuse as to what should be considered a Node or an Arc. To address this issue, the calculation of the metrics on BPMN models is conducted under the following assumptions:

1. All the Gateways (Exclusive, Inclusive and Parallel) count as Joints or Splits, depending on their type.
2. All Activities, Gateways (Exclusive, Inclusive and Parallel) and Events (End, Start and Intermediate) count as Nodes.
3. All Sequence flow arrows count as Arcs.
4. Message flow arrows do not count as Arcs.

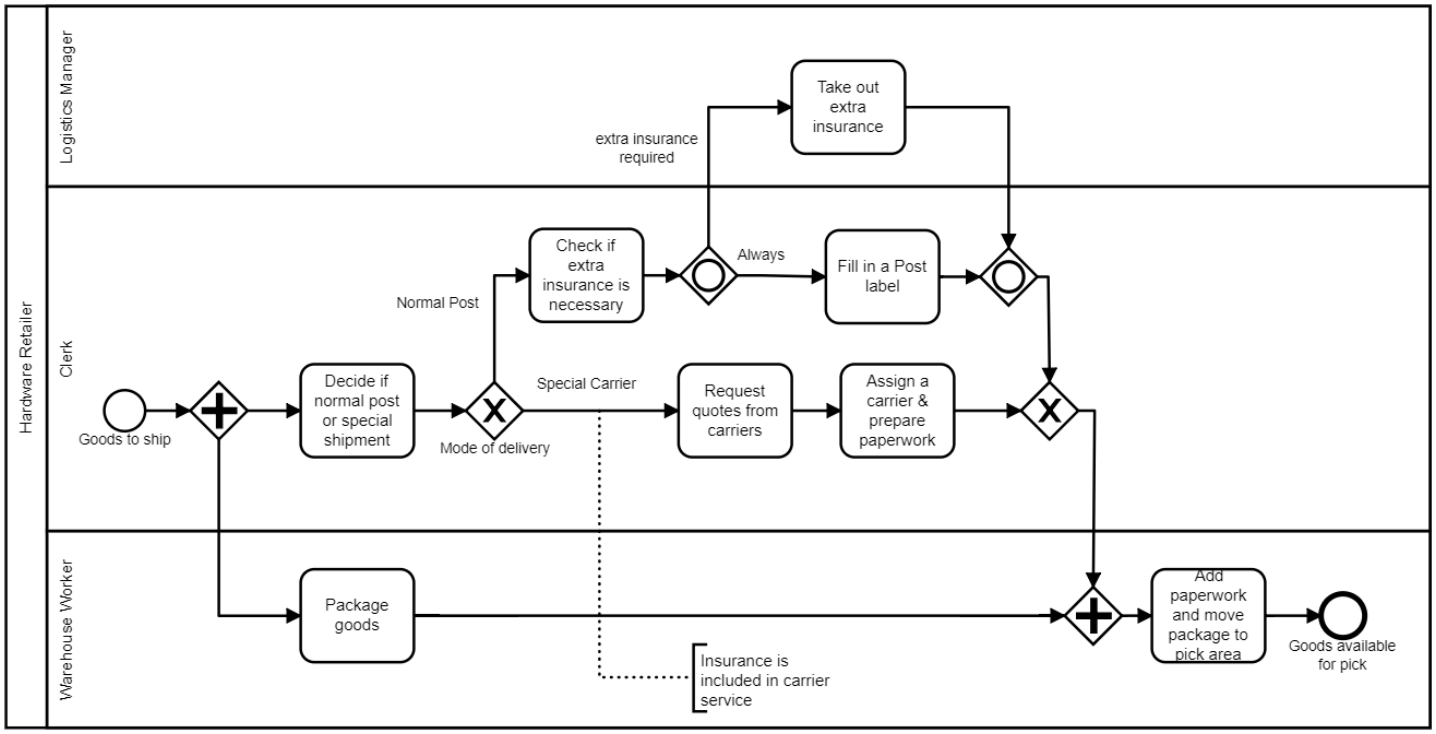

Figure 21: Hardware Shipment Process

Figure 21 shows an example of a BPMN model, that depicts a simple Hardware Shipment process from the SOA-based Business Process Database. Subsequently, examples of the calculation of each selected metric for the specific BPMN model are presented.

To calculate the *NOA* metric, a simple count of the Activity elements is required. In the same manner the *NOAJS* metric is calculated, adding the Joint and Split elements of the model.

$$NOA = 8$$

$$NOAJS = 14$$

The *CFC* metric is calculated by adding a value for all the different type of Split elements of the model.

$$CFC = CFC_{AND-split} + CFC_{OR-split} + CFC_{XOR-split} = 1 + (2^2 - 1) + 2 = 6$$

The *CNC* metric represents the ratio between Arcs and Nodes. The Arcs and Nodes are calculated by summing the elements as defined by the assumptions.

$$CNC = \frac{Arcs}{Nodes} = \frac{18}{16} = 1.125$$

Finally, *Density* is calculated by the below formula.

$$Density = \frac{Arcs}{Nodes * (Nodes - 1)} = \frac{18}{16 * (16 - 1)} = 0.075$$

Table 8 summarizes the values of the selected metrics.

Metric	*Value*
NOA	8
NOAJS	16
CFC	6
CNC	1.125
Density	0.075

Table 8: Calculated Values

This section justifies the selection of the metrics to be used in this research. The description of each metric is provided along with their mathematical calculation in a BPMN example. As previously discussed, standalone values mean very little for the evaluation of the complexity of a model. An assessment based on only one metric value bares little meaning, especially when threshold values have not been established for each metric. A more holistic approach for the measurement of complexity is needed that combines information about size, structure and control-flow. With the implementation of clustering in the set of selected values, an attempt is made to combine complexity measures and, ultimately, assign valuable meaning to reference values.

4.3 Data Pre-processing

The development of evaluation methods is based on the implementation of data analysis techniques, namely cluster analysis. For this purpose, data regarding the values of the complexity metrics selected in the previous section are required. This section describes the procedure followed to obtain the data necessary for the application of the clustering algorithm. First, the pre-processing of the SOA-based Business Process Database [87] is presented and, secondly, the calculation and standardization process of the data is explained.

4.3.1 SOA – based Business Process Database

In order to analyze and evaluate the notion of complexity for BPMN models, a repository of BPMN models is required. The SOA-based Business Process Database, comprised of 1000 business processes modeled in BPMN, constitutes a valuable source of models for the evaluation of our methodology.

Before the extraction of the data to be used for the implementation of the proposed method, a preprocessing of the database was deemed necessary. The data to be used should be extracted from actual, complete business processes, containing no errors and labeled in English. A thorough analysis of the database led to the selection of 87 business processes, aligned with the requirements set. The procedure followed to analyze the database and finalize the process models to be employed for the calculation of the metrics, ultimately resulting in the input dataset, is presented below.

A first glimpse at the SOA-based Business Process Database provides the results displayed in Table 9. The majority of the processes (530) contained in this database do not represent real business processes, but only examples of theoretical ones. On top of that, some of the processes would not open correctly with the used software, namely Camunda, because of either errors (17), or because they were created with a non-compatible modeling tool, namely Signavio (16). Moreover, 3 processes were not on the file at all and in their place were duplicates of other processes.

	Number of Processes
Not actual processes	530
Error	17
Not in file	3
Not Complete	155
Would Not Open (Signavio)	16
Non-English	67
English	212
SUM	1000

Table 9: SOA - based Business Process Database

From the remaining actual processes, 155 of them are not complete, meaning they incorporate tasks with not actual activities. The complete and actual processes in the database are described in a variety of languages, with the majority being in English (212). Spanish, French, German and Dutch are some of the languages used in the modeling of the rest of the processes (67). All the above can be found in Figure 22.

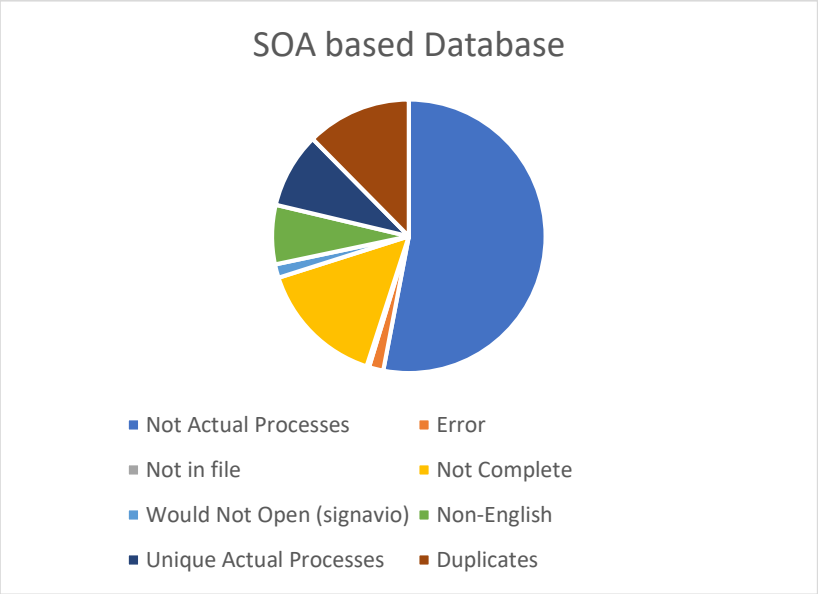

Figure 22: SOA-based Business Process Database

At this point the 212 complete processes in English and 14 translated complete processes are selected for further examination (226). By removing the duplicate processes only 87 unique actual processes (from the 1000 ones) remain. These processes were examined and given an appropriate title to convey their context, since the database only identified the processes with an index number. The 87 unique business processes (8.7 % of total processes in the database) that fit the criteria set from the author and are used in the creation of the dataset, are displayed in Appendix A.

4.3.2 Dataset Formation

Following the selection of the metrics and the establishment of the process pool to be used in the analysis, the development of the dataset ensued. The calculation of the value of each metric was performed according to the procedure described in section 4, for every one of the 87 valid BPMN

models in the given database. Figure 23 depicts a part of the initial dataset. It can be observed that the process index and title are presented first, followed by the values of each metric.

	A	B	C	D	E	F	G
1	PROCESS	TITLE	NOA	NOAJS	CFC	D	CNC
2	9	DEBUGGING	8	9	3	0.064	0.636
3	11	TRADING PROCESS	6	8	3	0.122	1.100
4	13	TROUBLE TICKET SYSTEM	6	8	4	0.100	1.000
5	15	DEMAND MEETING	7	10	4	0.083	1.000
6	17	HARDWARE SHIPMENT	8	14	6	0.075	1.125
7	18	VACATION REQUEST APPROVAL	3	5	4	0.143	1.000
8	19	APPLICATION HANDLING	4	7	9	0.071	1.000
9	20	CAB SERVICE	8	10	4	0.062	0.867
10	21	CAR RENTAL	6	10	5	0.090	1.077
11	23	BLOG POST PUBLISHMENT	11	12	2	0.063	1.000
12	29	SUPPLY PROCUREMENT	10	14	4	0.062	1.056
13	32	CUSTOMER ISSUE	2	2	0	0.250	0.750
14	33	BUG REPORT HANDLING	3	5	2	0.167	1.000
15	34	PURCHASE ORDER APPROVEMENT	4	6	2	0.143	1.000
16	41	PURCHASE WITH RFQ	5	8	3	0.100	1.000
17	42	UNSUCCESSFUL WIZARD PROCESS	5	6	2	0.143	1.000
18	44	PURCHASE ORDER DELIVERY CHECKING	3	5	3	0.109	1.091
19	47	PURCHASE REQUISITION	3	5	4	0.190	1.143
20	50	TEACHING	4	5	2	0.143	1.000
21	53	CHARITY	6	9	4	0.091	1.000
22	86	HOTEL AND FLIGHT RESERVATION	18	24	9	0.022	0.956
23	87	ISSUE DISCUSION AND VOTING	21	31	14	0.026	1.093
24	109	SOCIAL NETWORK CONECTION	4	5	2	0.068	0.750
25	110	SOCIAL MEDIA CAMPAIGN INTERACTION	9	10	7	0.077	0.923
26	111	THESIS PROPOSAL	14	15	7	0.040	0.870
27	113	JOB POSTING CREATION	5	7	2	0.125	1.000
28	114	TRAVEL BOOKING	9	12	3	0.059	0.941
29	116	SUPPLY ORDER	12	15	6	0.053	1.000

Figure 23: Example of Initial Data Values

For the formation of the final dataset required to perform the imminent analysis the adjustment of the values was necessary, given that the values of each metric are of different scale. Especially, in clustering analyses, data normalization is crucial in order to compare similarities between features based on certain distance measures. It has been proven that standardization of values prior to the implementation of clustering techniques leads to more efficient and accurate clusters [88], [89]. As a result, a normalization process, prior to any analysis method, is deemed obligatory.

The type of normalization used in this study is feature scaling, which mainly aims at the normalization of the range of independent variables or features of data. Essentially, it constitutes a valid means to bring all values into the range of [0,1] and address the problem of the difference of

scale. The method utilized, namely Min-Max normalization, is a simple, yet efficient way of rescaling data, widely used as a preprocessing step in data mining [88].

The formula for the Min-Max normalization technique is:

$$x' = \frac{x - Min(x)}{Max(x) - Min(x)}$$

where x is an original value, x' is the normalized value and $Min(x)$, $Max(x)$ are the minimum and maximum values of the feature (i.e., metric), respectively.

The result of the above process is a normalized dataset of metric values, prepared for analysis, as partially displayed in Figure 24.

	A	B	C	D	E
1	***NOA***	***NOAJS***	***CFC***	**D**	**CNC**
2	0.158	0.137	0.188	0.203	0.097
3	0.105	0.118	0.188	0.454	0.515
4	0.105	0.118	0.250	0.359	0.425
5	0.132	0.157	0.250	0.288	0.425
6	0.158	0.235	0.375	0.252	0.538
7	0.026	0.059	0.250	0.542	0.425
8	0.053	0.098	0.563	0.237	0.425
9	0.158	0.157	0.250	0.196	0.305
10	0.105	0.157	0.313	0.315	0.495
11	0.237	0.196	0.125	0.199	0.425
12	0.211	0.235	0.250	0.197	0.475
13	0.000	0.000	0.000	1.000	0.199
14	0.026	0.059	0.125	0.644	0.425
15	0.053	0.078	0.125	0.542	0.425
16	0.079	0.118	0.188	0.359	0.425
17	0.079	0.078	0.125	0.542	0.425
18	0.026	0.059	0.188	0.398	0.507
19	0.026	0.059	0.250	0.746	0.554
20	0.053	0.059	0.125	0.542	0.425
21	0.105	0.137	0.250	0.320	0.425
22	0.421	0.431	0.563	0.024	0.385
23	0.500	0.569	0.875	0.043	0.509
24	0.053	0.059	0.125	0.223	0.199
25	0.184	0.157	0.438	0.260	0.356
26	0.316	0.255	0.438	0.100	0.307
27	0.079	0.098	0.125	0.466	0.425
28	0.184	0.196	0.188	0.183	0.372
29	0.263	0.255	0.375	0.156	0.425

Figure 24: Example of Normalized Metric Values

The steps of the complete process followed to obtain the final dataset used in the current research are displayed in Figure 25.

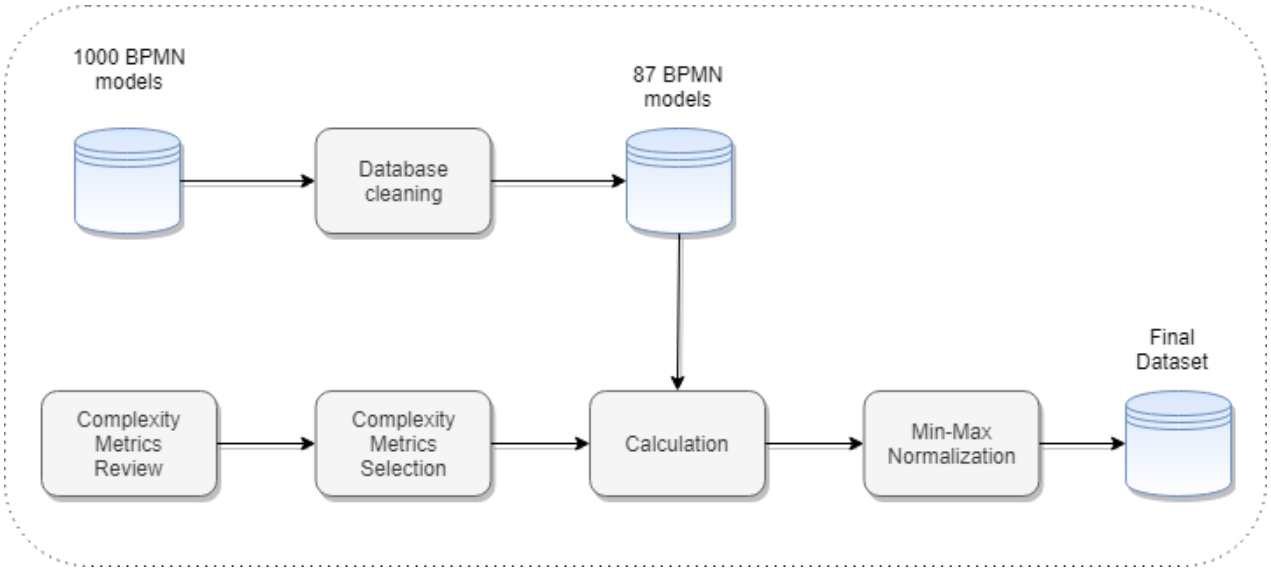

Figure 25: Formation of Final Dataset

4.4 Clustering Parameters

The development of the evaluation methods proposed in the next chapter, is achieved through Cluster analysis and, more specifically, using the K-means algorithm. This algorithm is one of the most popular ones for unsupervised learning problems and is found to deliver reliable results. Since the data available for analysis do not contain pre-existing labeling, which means the implementation of classification methods is not an option, clustering seems the obvious choice.

Several parameters require definition for the application of the K-means algorithm. The first and most important parameter to be established is the number of clusters for the data to be grouped in. Since the number of clusters are meant to partition the data to complexity categories, essentially representing the number of categories, that number is set to be three. Taking into account the number of instances in the dataset, more clusters would partition the data into very small groups that would not allow for trustworthy interpretation.

Another significant parameter for centroid-based clustering methods is the proximity measure. As discussed in chapter 2, there is a number of measures available for calculating distance between instances. In this case, the selected measure is Euclidean distance, a very popular method, commonly used as the default distance metric for many cluster analysis tools [90].

Lastly, the K-means algorithm requires an initialization method, meaning that the method of assigning the first three centroid values needs to be specified. Random initialization, during which the initial centroids are randomly placed in the Euclidean space, is chosen, since during the experiments no need for a more sophisticated initialization method was revealed. Table 10 summarizes the above.

Parameter	Choice
Method	Simple K-means
Number of Clusters	3
Distance Measure	Euclidean Distance
Initialization Method	Random

Table 10: K-means Parameter Setting

4.5 Summary

This chapter offers an overview of the steps followed for the development of the complexity assessment methods. Initially, the reasoning behind the selection of complexity metrics from literature is explained, the metrics are described and examples for their calculation in a BPMN model are provided. The selection of the complexity metrics directly serves the purpose of this thesis, which includes measuring complexity through the scope of Redesign. Next, the phase of Data Pre-processing is described in detail. The criteria for extracting models from the SOA-based Business Process Database are presented and, after thorough examination of the dataset, 87 models are obtained. Upon calculation of the metric values for each of the models, the final dataset to be analysed is formed. Last, the required parameters for the implementation of the K-means algorithm on the final dataset are defined. The next chapter details the steps followed, in order to

develop the two methods proposed for the evaluation of complexity for BPMN models using cluster analysis and presents the derived results.

CHAPTER 5

Evaluation of Business Process Complexity for Redesign

In this chapter, the methods of combining the selected complexity metrics and performing a cluster analysis, with the ultimate goal of evaluating a model's capability for Redesign, are presented in detail. Two approaches are followed to address this problem. First, a cluster analysis, using the metrics found in literature, is presented step by step. Next, a second approach is introduced, that includes the proposal of a weighted sum metric. The second approach is deemed necessary, in order to simplify the understanding of the model and the process of extracting meaningful thresholds. Finally, the procedure of assessing the complexity of a new process and recognizing its need or not for Normalization is described and explanatory examples are given.

5.1 First Approach: Cluster Analysis

As previously discussed, a cluster analysis is performed with the intend to identify business process models with high complexity, in potential need for Normalization. This section presents the findings derived from the implementation of the K-means algorithm on the dataset, containing the five metrics as attributes. The initial approach of the analysis is described in detail, explaining the reasons that, ultimately, lead to the need of an updated approach, requiring only three attributes.

5.1.1 Initial Clustering

After implementing the K – means algorithm using the parameters set on section 4, the following results emerged. The distribution of the 3 clusters formed is presented in Table 11. It can be observed that the first two clusters contain the majority of the instances (90%), with the third cluster including only 10% of the instances.

	Number of instances	Percentage of instances
Cluster 1	33	38%
Cluster 2	45	52%
Cluster 3	9	10%

Table 11: Initial Cluster Distribution

In Table 12, the centroid values of each cluster per feature are displayed. A first interpretation of the centroid values reveals a clear trend of low values belonging to the first cluster, slightly higher values belonging in the second one and, lastly, the highest values belonging to the third cluster. This distribution stands for almost all metrics, with the exception of Density. According to literature, higher values of the selected metrics are clear indicators of complexity, as previously established. Therefore, it is a safe assumption that Cluster 1 represents low complexity models, Cluster 2 moderate complexity models and Cluster 3 high complexity models.

A next step is a more in-depth analysis of the results and the impact each metric has on the proposed clustering method. An initial observation regarding the centroid values of NOA and NOAJS, is that they appear to be very similar.

	NOA	NOAJS	CFC	D	CNC
Centroid 1	0.091	0.091	0.115	0.430	0.393
Centroid 2	0.243	0.228	0.233	0.181	0.397
Centroid 3	0.529	0.540	0.667	0.073	0.559

Table 12: Centroid values for 5 Attributes

Indeed, the close correlation of the two metrics becomes apparent when the distribution of the clusters is visualized for this two metrics alone. Figure 26, that depicts the instances' metric values and distribution of the three clusters using different colors, reveals an almost linear relation

between the two values. The same conclusion arises upon examining the relation of the two metrics in comparison to another metric, namely CFC. By studying Figure 27 and Figure 28, the similarity of the two metrics (NOA-NOAJS) is once again established.

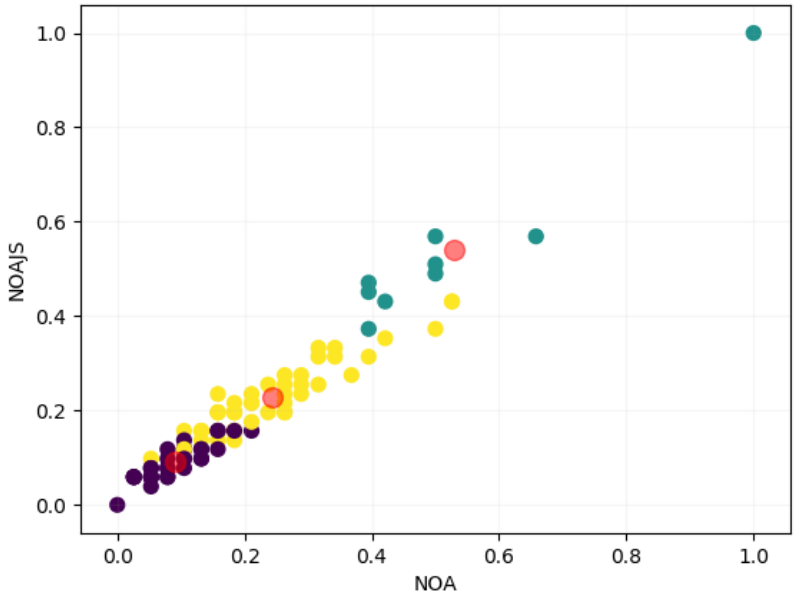

Figure 26:Cluster Distribution for NOA – NOAJS

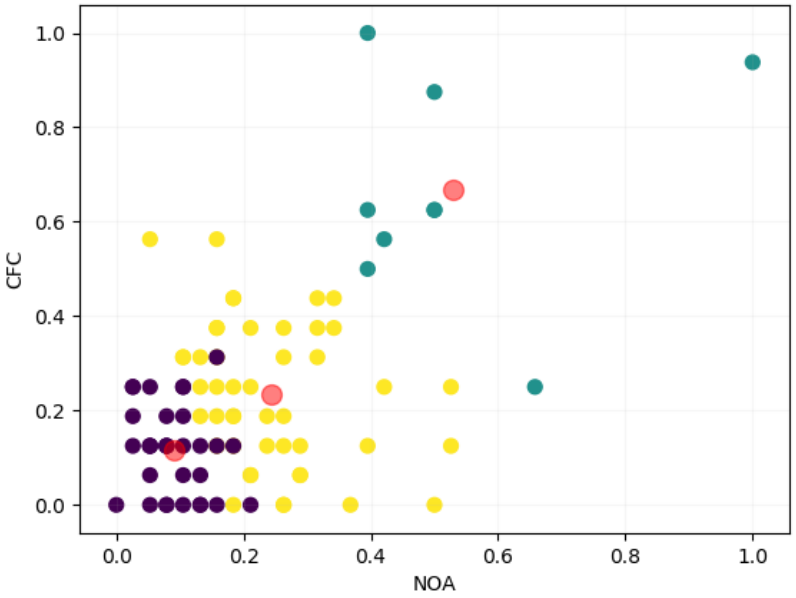

Figure 27:Cluster Distribution for NOA - CFC

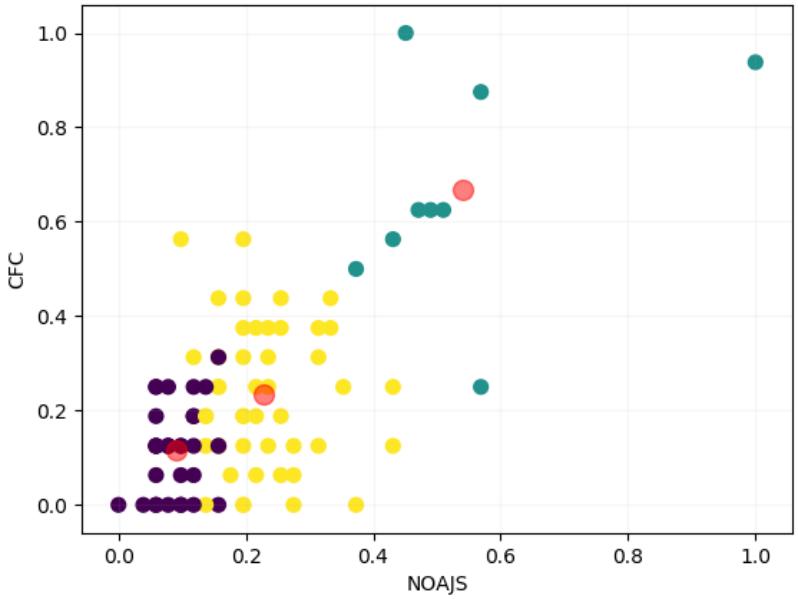

Figure 28:Cluster Distribution for NOAJS - CFC

All the above indicate that the two metrics (NOA-NOAJS) carry parallel information, which is not surprising, since their definitions are closely similar, with the difference that NOAJS also accounts for Joins and Splits. However, these types of elements, are commonly fewer than the activities in a diagram, which means that the effect of Joins and Splits exists, of course, but is not strong enough to distinct the two metrics for this cluster analysis. This deduction is reinforced by simply calculating the Pearson correlation between the two metrics, which is found to be 0.97, indicating very strong positive correlation.

As per the previous discussion, the use of both metrics as attributes in the cluster analysis is deemed unnecessary. However, the input of a model's size is a very important information, valuable to the present analysis. Therefore, we decide to exclude one of them, namely NOA, from the clustering process, while retaining NOAJS for the sheer reason that, apart from the information of size, it incorporates Join and Split elements in its calculation, mostly thought to contribute to the complexity of a process model [52].

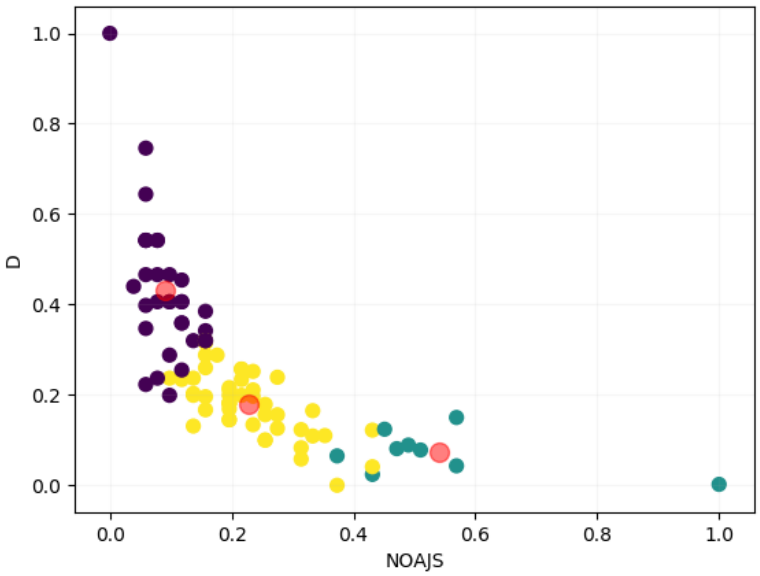

Figure 29: Cluster Distribution for NOAJS - D

Another outcome observed was the totally reverse effect of the Density (D) metric in the cluster analysis. As shown in Figure 29, high values of Density are associated with smaller sized models and lower Density values with larger models. This result strikes as a paradox, since complex models are expected to exhibit high values for both metrics. The model resulting from the cluster analysis groups together dense but small models, and the opposite. The same effect appears when we compare the clustering results between CFC and Density. Figure 30 demonstrates the association of low Density values with high CFC values and vice versa.

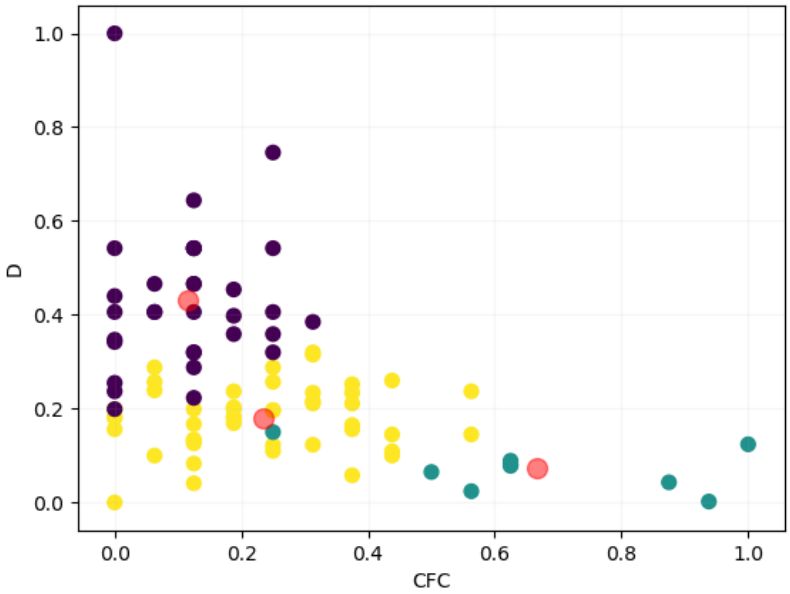

Figure 30: Cluster Distribution for CFC – Density

This can be explained by examining the mathematical formula of the metric. It is apparent that when the number of Nodes increases the denominator becomes larger in a way that it is impossible for the numerator (number of Arcs) to keep up, so the fraction decreases. This connection between size and density creates issues when comparing models of different sizes and does not serve the purpose of this analysis.

However, even in models with the same number of activities, Density is not performing as expected. By employing a simple example, we aim to discover how this metric works based on our assumptions. Figure 31 depicts two very simple theoretical processes. The first one is a simple sequential process with no gateways and NOAJS=3. The second one is a slightly more "complex model", with the same number of activities, that includes two OR-gateways, with NOAJS=5. One might expect the second model to have a higher Density value, since research supports that complex models have higher Density values. However, when the calculation is complete, taking into account the assumptions set in chapter 4, the simpler model has a Density value of 0.2, larger than the slightly more complex one, which has 0.16.

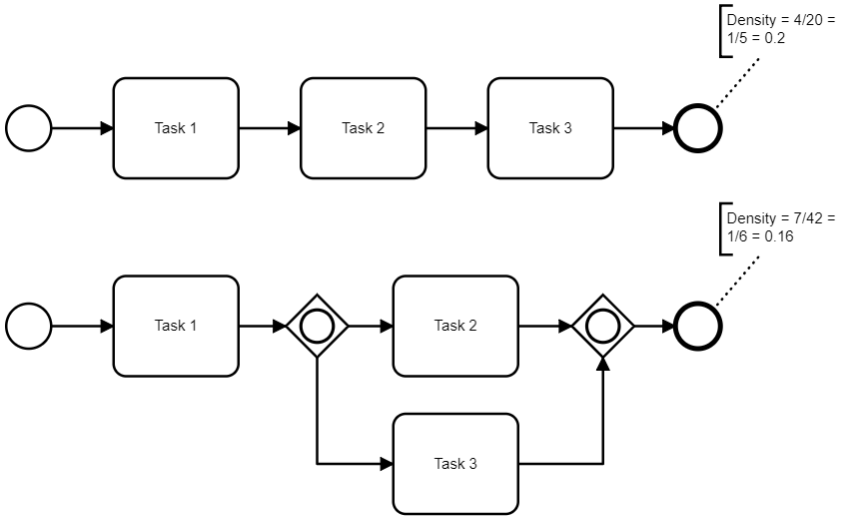

Figure 31: Example for Density calculation

These are strong indications that Density does not work as expected in literature, which may result from the assumptions set in the previous chapter. These assumptions are required in order to adapt the metrics in BPMN models. In addition, literature suggests that this metric should be used to compare models of the same size, i.e., with the same number of activities [28]. In our case, the number of Nodes used in the calculation does not match the number of activities a process contains, because of the adaptation of the metrics to BPMN. Considering that our goal is to classify the clusters in terms of complexity and the Density metric does not contribute to this end, it is also excluded from the analysis.

The last two Figures, Figure 30 and Figure 31, obtained by the previous analysis, concern the CNC metric in relation to NOAJS and CFC respectively. In both of the images the clusters are clearly divided, offering a clear picture of the distribution. All in all, NOAJS and CFC seem to have the greatest influence on the present clustering results, with their interpretation totally agreeing with existing literature. Although the CNC metric appears to have the smallest standard deviation among all metrics, which is partly expected considering the mathematical definition of the metric, the information it adds to the analysis is deemed significant. Consequently, the three metrics are considered capable of yielding clear and accurate results in accordance with the purpose of this analysis and the relevant literature. The next step is for an updated cluster analysis, including only the necessary metrics, that have been found to properly serve the aim of this analysis.

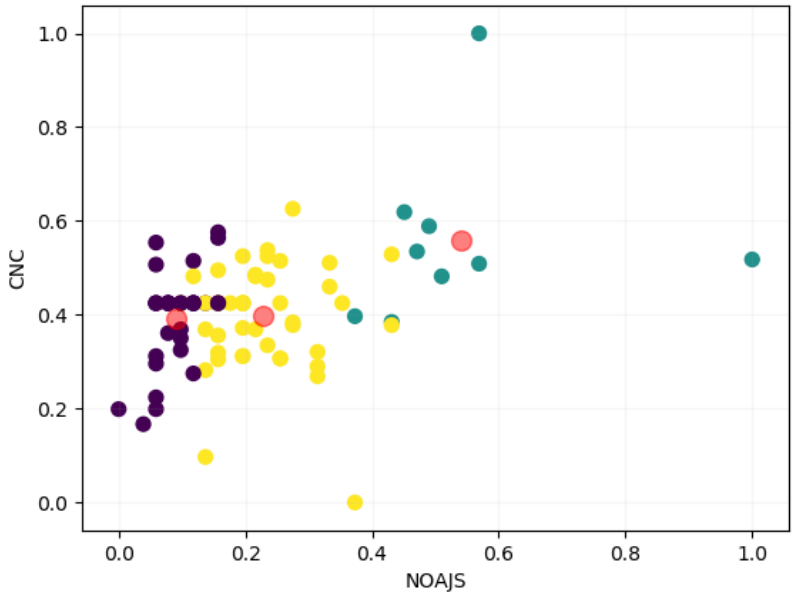

Figure 32: Cluster Distribution for NOAJS-CNC

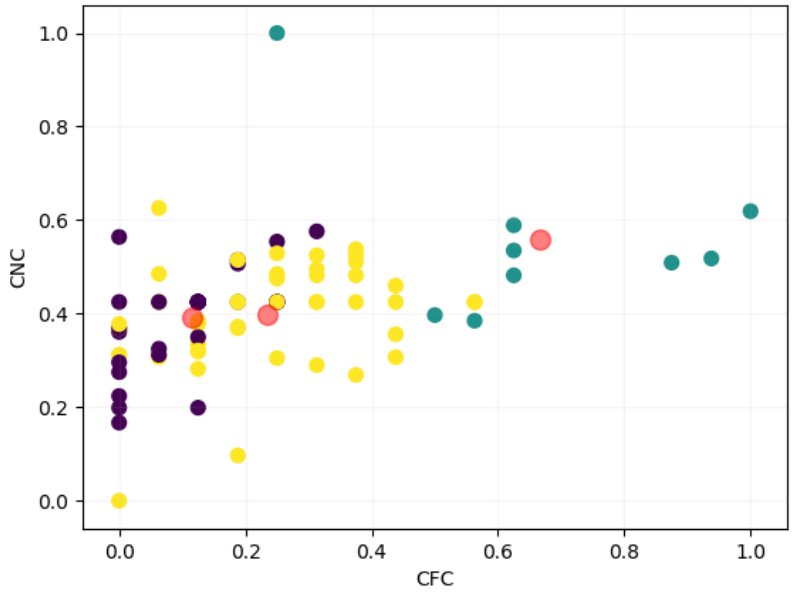

Figure 33: Cluster Distribution for CFC - CNC

5.1.2 Updated Clustering

After the above preliminary analysis, we proceed in implementing a new cluster analysis, driven by three metrics; namely NOAJS, CFC and CNC. The model to result from this new approach is expected to be more accurate and efficient for the purpose of evaluating complexity. The three metrics cover adequately all three aspects of complexity, namely activity, control-flow and structure, that we regard important for the aim of the present thesis.

Table 13 presents the distribution of the three clusters emerging from the analysis, which appear similar to the one from the initial clustering. The percentage owned by Cluster 1 and Cluster 2 comes close to 90% of the instances, while only 9.2% belong in Cluster 3.

	Number of instances	Percentage of instances
Cluster 1	48	55.2%
Cluster 2	31	35.6%
Cluster 3	8	9.2%

Table 13: New Cluster Distribution

The centroid values act as the representatives of each cluster. This means that each cluster consist of instances similar to their respective centroids. By examining Table 14 it becomes apparent that Cluster 1 includes low values for all three metrics. This means that low sized models, with limited control flow complexity and simple structure are categorized in that cluster. Cluster 2 involves slightly increased values for all the three metrics compared to Cluster 1. The values reveal that moderate sized models, with mediocre control flow complexity and acceptable structure belong to this cluster. Finally, the third cluster is found to contain the highest values for the three metrics, meaning it includes larger models, with increased control flow complexity and more complicated structure. Thus, we deduce that Cluster 3 contains models with higher complexity.

	NOAJS	CFC	CNC
Centroid 1	0.149	0.093	0.363
Centroid 2	0.213	0.339	0.447
Centroid 3	0.561	0.688	0.580

Table 14: Centroid Values for 3 Attributes

Centroid values displayed in Table 14 act as reference values, that represent a group of instances with similar feature values. By investigating these features, comparing the clusters to one another becomes feasible and, considering the performed literature review for complexity metrics, a pattern is revealed. Leveraging the latter, the possibility to assign meaning to each cluster regarding complexity is facilitated. Table 15 describes each cluster in terms of meaning and color representation for the remaining analysis.

	Complexity	Color
Cluster 1	Low	Purple
Cluster 2	Moderate	Yellow
Cluster 3	High	Green

Table 15: Cluster Description

The distribution of the cluster is displayed in Figure 34, Figure 35 and Figure 36 for each combination of metrics. It is clearly observed that the clusters are indeed easily distinguished, which is an indication of efficient clustering. The previous conclusion, regarding the assignment of meaning to each cluster in terms of complexity, is reinforced by the visualization of the clusters.

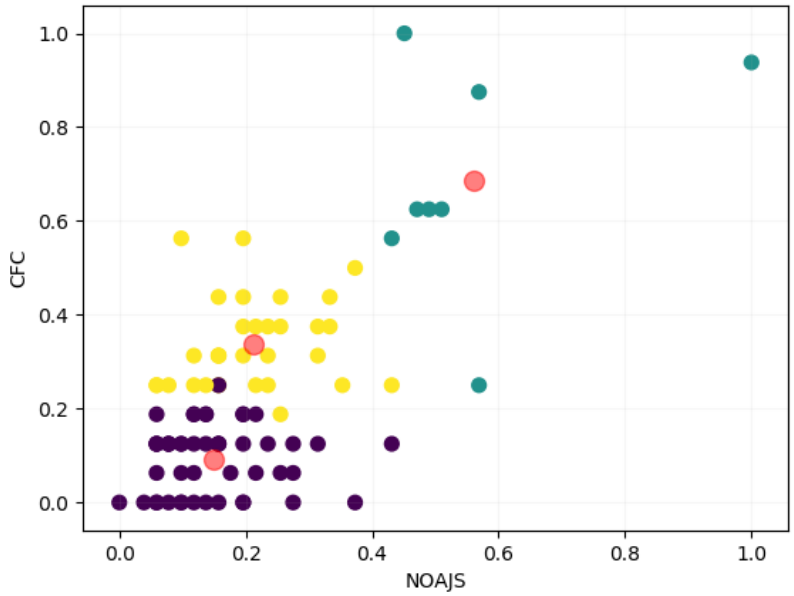

Figure 34: Updated Clustering - Cluster Distribution for NOAJS - CFC

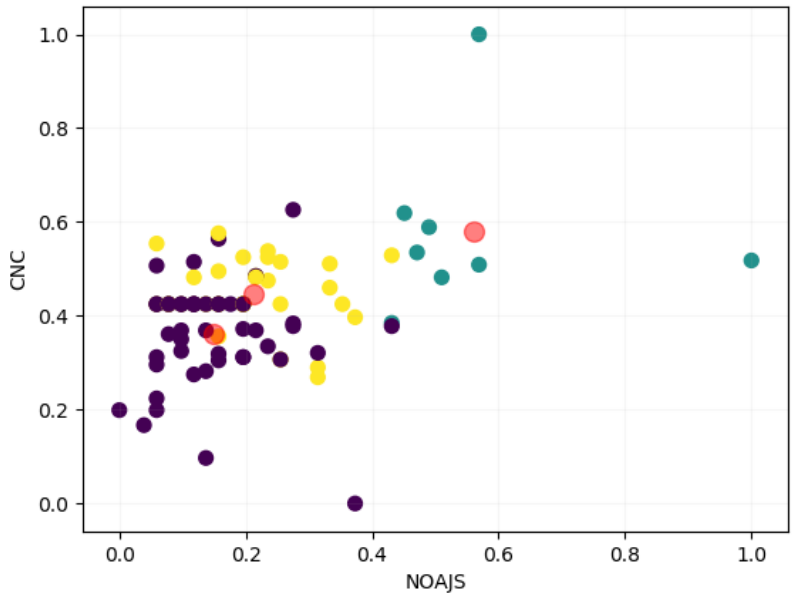

Figure 35: Updated Clustering - Cluster Distribution for NOAJS - CNC

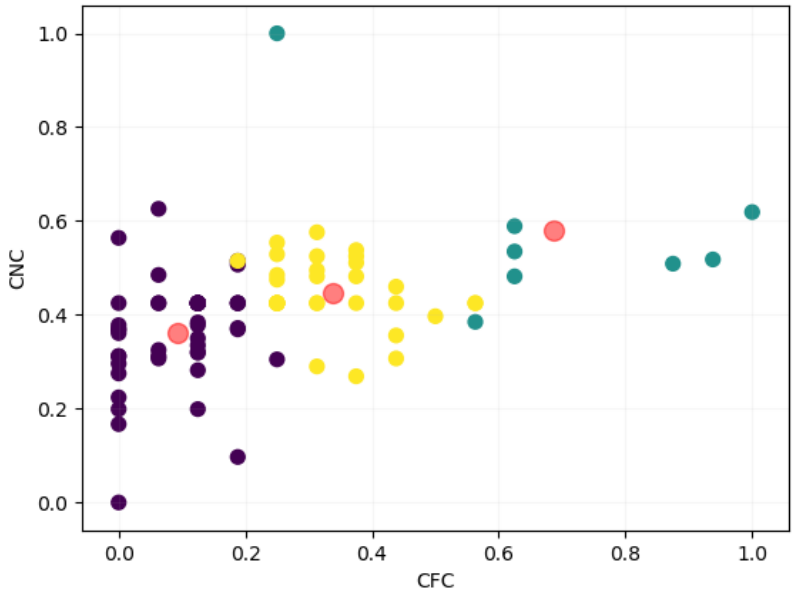

Figure 36:Updated Clustering - Cluster Distribution for CFC -CNC

The fact that three features are used in the creation of the model, facilitates a 3-dimensional illustration of the clusters. Figure 37 displays the distribution of each cluster in the 3-dimensional space and allows for a closer examination. Evidently, the clusters appear to be clearly distinguished, especially Cluster 3, which includes instances with high values on each feature. Additional perspectives of the same distribution are available in Figures 38 and 39, for better comprehension.

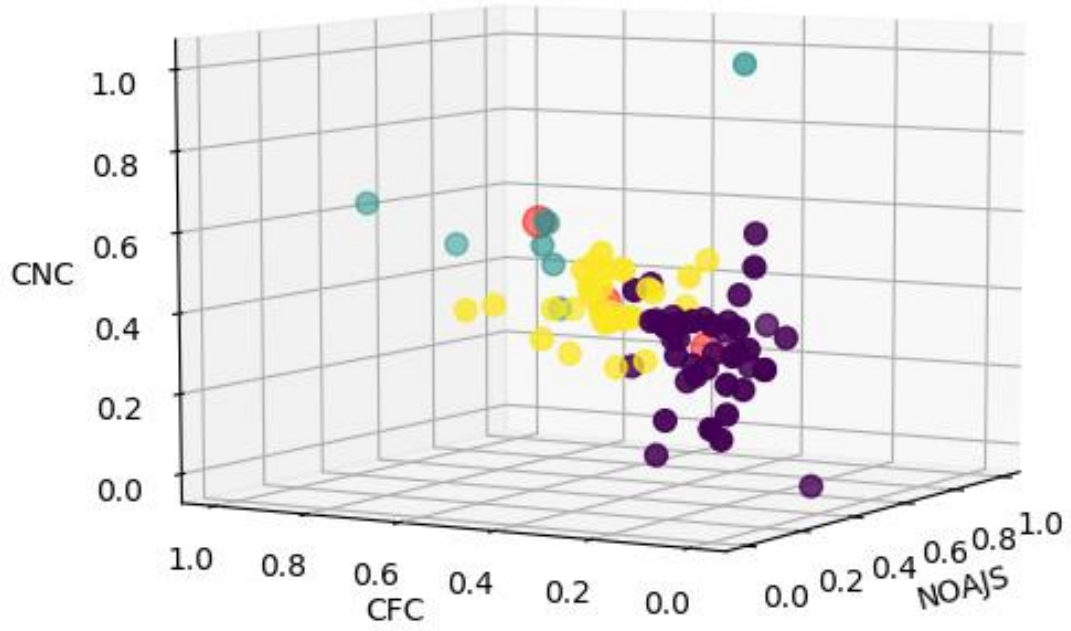

Figure 37: 3D illustration

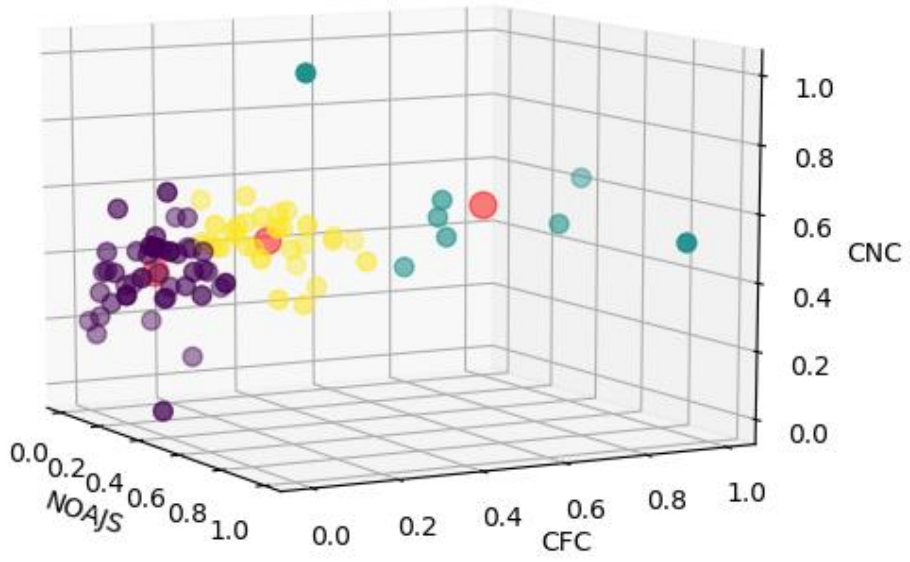

Figure 38: 3D illustration - 2nd perspective

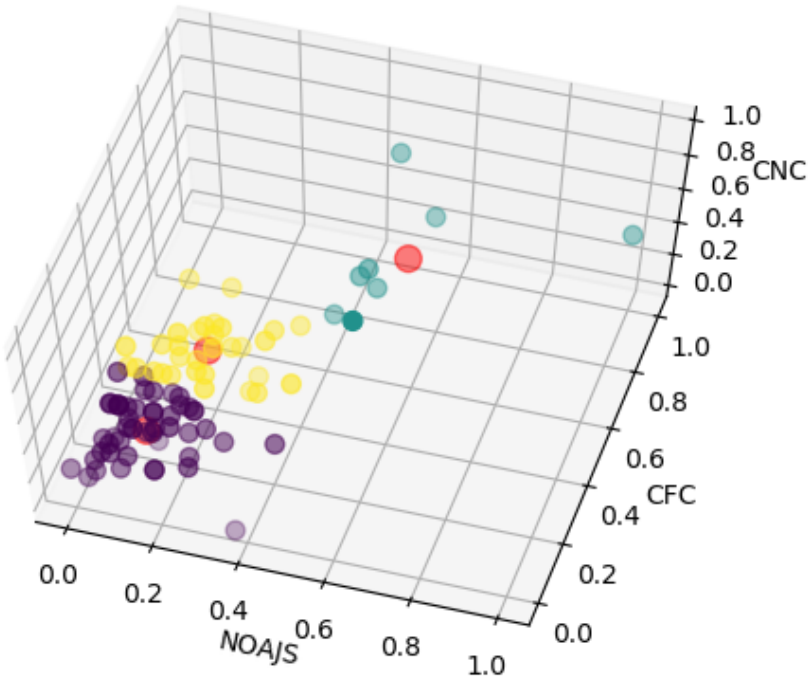

Figure 39: 3D illustration - 3nd perspective

5.2 Second Approach: Proposal of a Weighted Sum Metric

In an attempt to obtain a more flexible way of measuring the complexity level of a BP model, a second approach is proposed in this section. Drawing from related research on complexity metrics, a method to combine NOAJS, CFC and CNC into a single weighted sum measure is introduced. This constitutes a different way to evaluate complexity, using information extracted from all three of the selected metrics and uniting them into one single measure through a sum. In this way the extraction of threshold values is supported. In addition, this approach allows for the weighting of each metric, depending on the perspective of each user. Basically, it is a way to assign priority numbers that translate to specific weights for each value.

The method selected for combining the metrics into one weighted sum is introduced in [83] and is presented in detail below. It is basically a method to assign priorities to metric values before combining them. Let $p_1, p_2 \dots p_n$ be the priorities, in the form of numbers, that are assigned to n metrics and $w_1, w_2 \dots w_n$ their weight. The weight of each metric is calculated by the formula:

$$w_i = \frac{p_i}{\sum_{i=1}^{n} p_i}$$

If $v_1, v_2 \dots v_n$ are the values of each of the n metrics, then:

$$WS = \sum_{i=1}^{n} w_i v_i$$

This constitutes a generic form of the weighting method used that can be applied to any number of metric values. In our study the selected metrics are 3, which means that $n = 3$. Examples of the calculation of the weighted sum metric are presented in the next section.

The described approach enables the extraction of threshold values, compared to the first approach, which introduces reference values for each complexity level. Distinguishing between complexity levels is accomplished again through a cluster analysis. This time, however, the centroids are represented by a one-dimensional vector. This means that they can be depicted on a line. therefore, the mean point of the intervals between centroids per two acts as a separator for the clusters. This approach on threshold definition is further explained later, upon the application of the method. An overview of the steps followed for the weighted sum method is available on Figure 40.

Figure 40: Steps of the Weighted Sum Method

Assigning priority to metrics offers the potential for researchers to customize the assessment method, according to their perspective on complexity evaluation. Essentially, the weighting of the features prior to the implementation of the algorithm allows for the creation of a brand new model. Two scenarios are analysed next in order to demonstrate this fact. The first scenario is a simple combination of the metrics, with equal weighting for each one. Thresholds are extracted, based on this method that facilitates the separation of the complexity levels. The second scenario considers the weighting parameter, by lowering the priority of a metric, to demonstrate the impact this option has on the threshold values. Obviously, there are multiple options when it comes to

assigning priority to metrics, that can serve the purpose of each researcher. In this particular example, the priority of the control-flow metric, i.e., CFC, is significantly higher, while the priority of the size metric, i.e., NOAJS, is lowered relatively to the other metrics.

5.2.1 Scenario 1: Equal priorities

The first Scenario to be examined concerns equal weighting for the three metrics. This translates to each metric having a priority of 1 ($p_{1,2,3} = 1$), which in turn means each metric weights approximately 0.3333, according to the next calculation.

$$w_{1,2,3} = \frac{1}{3} = 0.3333$$

Following the previously defined procedure, the weighted sum measure is calculated for all instances. For example, for the first instance with values 0.137, 0.188, 0.097 the weighted sum is calculated by the below formula. Some of the weighted sum values are displayed in Figure 41 as an example.

$$WS = \sum_{i=1}^{3} w_i v_i = 0.3333 * 0.137 + 0.3333 * 0.188 + 0.3333 * 0.097 = 0.1405$$

	A	B	C	D
1	***NOAJS***	***CFC***	**CNC**	**WEIGHTED SUM**
2	0.137	0.188	0.097	0.1405
3	0.118	0.188	0.515	0.2735
4	0.118	0.250	0.425	0.2643
5	0.157	0.250	0.425	0.2773
6	0.235	0.375	0.538	0.3828
7	0.059	0.250	0.425	0.2446
8	0.098	0.563	0.425	0.3619
9	0.157	0.250	0.305	0.2372
10	0.157	0.313	0.495	0.3213
11	0.196	0.125	0.425	0.2487
12	0.235	0.250	0.475	0.3202
13	0.000	0.000	0.199	0.0664
14	0.059	0.125	0.425	0.2030
15	0.078	0.125	0.425	0.2095
16	0.118	0.188	0.425	0.2434
17	0.078	0.125	0.425	0.2095
18	0.059	0.188	0.507	0.2512
19	0.059	0.250	0.554	0.2877
20	0.059	0.125	0.425	0.2030
21	0.137	0.250	0.425	0.2708
22	0.431	0.563	0.385	0.4596
23	0.569	0.875	0.509	0.6509
24	0.059	0.125	0.199	0.1277
25	0.157	0.438	0.356	0.3167
26	0.255	0.438	0.307	0.3332
27	0.098	0.125	0.425	0.2161
28	0.196	0.188	0.372	0.2519
29	0.255	0.375	0.425	0.3517

Figure 41: Scenario 1 - Weighted Sum

At this point, a cluster analysis on the one-dimensional array of the weighted sum feature is possible. The result is illustrated in Figure 42. It can be easily observed that the instances and their distance from the centroids can be represented on a straight line, including values from 0 to 1. According to K-means the cluster distribution is the one presented in Table 16. It appears that only a small percentage (8%) of instances are grouped in Cluster 3, while the majority belongs in Clusters 1 and 2. Figure 42 depicts the distribution of the clusters in relation to the instances. This approach offers similar results as the first one, however the visualization of the cluster distribution allows for extracting easily comprehensible threshold values.

	Number of instances	Percentage of instances
Cluster 1	47	54%
Cluster 2	33	38%
Cluster 3	7	8%

Table 16: Scenario 1 - Cluster Distribution

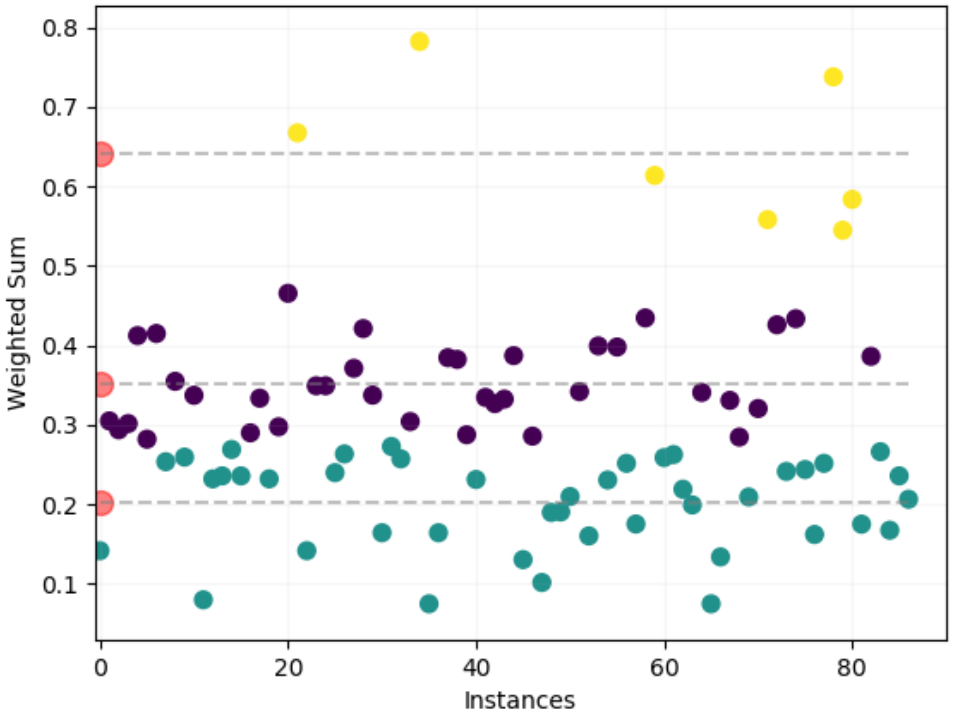

Figure 42: Scenario 1 - Clustering visualization

The centroid values provided by the analysis are displayed in Table 17. Essentially, these values constitute reference points on a straight line. In this way, comparing the distance between centroids and assigning an instance to a cluster becomes a more straightforward task, since the defined radius of a cluster is easily distinguished. Given the weighted sum of a future process model, the latter will be assigned to the cluster having a centroid value closest to its weighted sum.

	Weighted Sum
Centroid 1	0.197
Centroid 2	0.339
Centroid 3	0.630

Table 17: Scenario 1 - Centroid Values

Figure 43 illustrates the clusters and their centroid values on a straight line. Based on the definition of the K-means algorithm and the parameters used in this case, the cluster assignment is determined by the Euclidean distance. Basically, this means that the points that separate the clusters, and act as threshold values, can be decided through a simple calculation of the means of the formed intervals between the cluster centroids. These values are present in Table 18.

Centroids	Mean of Intervals
Centroid 1 – Centroid 2	0.268
Centroid 2 – Centroid 3	0.484

Table 18: Scenario 1 - Means of formed Intervals

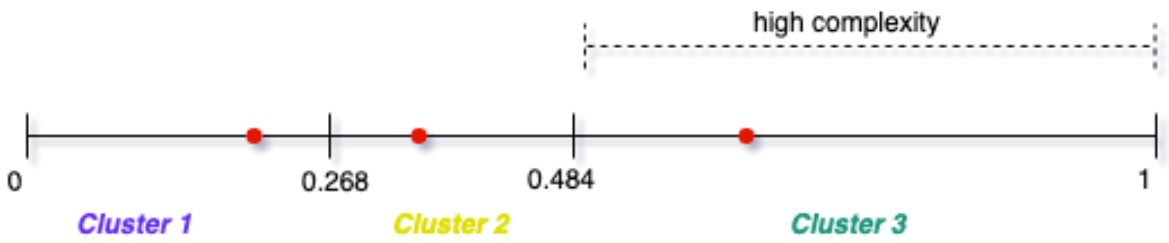

Figure 43: Scenario 1 - Visualization of Centroids and Thresholds

Pertaining to the related work on complexity evaluation and the definition of the used metrics, it is possible to draw conclusions regarding the type of models each cluster represents. Judging by the centroid values of the clusters, a correlation between these values and complexity is obvious.

Taking into account the previously defined threshold values, Table 19 describes the complexity level represented by each cluster and the threshold values that separate them.

	Thresholds	Complexity
Cluster 1	weighted sum value in (0, 0.268]	Low
Cluster 2	weighted sum value in (0.268, 0.484]	Moderate
Cluster 3	weighted sum value in (0.484, 1]	High

Table 19: Scenario 1 - Threshold values for complexity levels

5.2.2 Scenario 2: Low NOAJS Priority, High CFC Priority

This particular approach of a weighted sum measure offers the option of assigning priority to the selected metrics and developing a customized model for assessing complexity. Let's assume that the purpose of an analysis dictates that a model's size should not influence the evaluation of complexity, while the control flow complexity of a model, i.e., the amount of gateway splits, should be mainly considered. This approach allows for the assignment of priorities accordingly, e.g., $p_1 = 1, p_2 = 10, p_3 = 5$, which means that NOAJS has a very low priority, CFC has a high priority, while CNC a moderate one. Following the assignment, the weights of the metrics are calculated according to the method explained previously. For this example, the weights are as follows: $w_1 = 0.0625, w_2 = 0.625, w_3 = 0.3125$, and are used in the calculation. The weighted sum measure is calculated for all instances according to the following example. Given the first instance with values 0.137, 0.188, 0.097, the weighted sum is calculated by the formula below:

$$WS = \sum_{i=1}^{3} w_i v_i = 0.0625 * 0.137 + 0.625 * 0.188 + 0.3125 * 0.097 = 0.1560$$

To further exemplify the way our data are now formed, we quote Figure 44.

Following the same process as in Scenario 1, a cluster analysis is performed on the new weighted sum feature. Next, the instances are grouped to the three clusters again using the K-means algorithm. Their distribution, that does not differentiate dramatically from the previous clustering,

can be found in Table 20. Cluster 1 and Cluster 2 include the majority of the instances, while a small percentage belongs to Cluster 3. The centroids of the formed clusters are presented in Table 21.

	A	B	C	D
1	***NOAJS***	***CFC***	**CNC**	**WEIGHTED SUM**
2	0.137	0.188	0.097	0.1560
3	0.118	0.188	0.515	0.2856
4	0.118	0.250	0.425	0.2965
5	0.157	0.250	0.425	0.2989
6	0.235	0.375	0.538	0.4172
7	0.059	0.250	0.425	0.2928
8	0.098	0.563	0.425	0.4905
9	0.157	0.250	0.305	0.2613
10	0.157	0.313	0.495	0.3597
11	0.196	0.125	0.425	0.2232
12	0.235	0.250	0.475	0.3195
13	0.000	0.000	0.199	0.0623
14	0.059	0.125	0.425	0.2147
15	0.078	0.125	0.425	0.2159
16	0.118	0.188	0.425	0.2574
17	0.078	0.125	0.425	0.2159
18	0.059	0.188	0.507	0.2794
19	0.059	0.250	0.554	0.3331
20	0.059	0.125	0.425	0.2147
21	0.137	0.250	0.425	0.2977
22	0.431	0.563	0.385	0.4988
23	0.569	0.875	0.509	0.7415
24	0.059	0.125	0.199	0.1441
25	0.157	0.438	0.356	0.3944
26	0.255	0.438	0.307	0.3854
27	0.098	0.125	0.425	0.2171
28	0.196	0.188	0.372	0.2457

Figure 44: Scenario 2 - Weighted Sum

	Number of instances	Percentage of instances
Cluster 1	42	48%
Cluster 2	35	40%
Cluster 3	10	12%

Table 20: Scenario 2 - Cluster Distribution

	Weighted Sum
Centroid 1	0.1681
Centroid 2	0.3408
Centroid 3	0.6154

Table 21: Scenario 2 - Centroid Values

Additionally, the instances and their distance from the centroids can be represented in a straight line, including values from 0 to 1. As discussed in the previous section, the one-dimensional representation allows for the definition of thresholds that separate the clusters. These thresholds are signified by the means of the intervals formed, between the centroids. The calculated thresholds are displayed in Table 22. Figure 45 illustrates the clusters and the threshold values that separate them, along with the centroids, in a straight line.

Centroids	Mean of Intervals
C1 - C2	0.254
C2 - C3	0.478

Table 22: Scenario 2 - Means of Intervals

Figure 45: Scenario 2 - Visualization of Centroids and Thresholds

According to related research on complexity, higher values of the selected metrics reveal higher complexity. In this case, assigning a complexity level to each cluster is justified. Table 23 reveals the complexity represented by each cluster and the extracted thresholds. Compared to the results derived from the cluster analysis performed in Scenario 1, there is no significant change in the thresholds values. The difference between the two Scenarios regarding the threshold definition are trivial. However, the new model developed is expected to have an impact on the classification of new models, since the values of the metrics need to be weighted first to be able to compare them to those specific thresholds. More details regarding the assessment process are available in the next section.

	Thresholds	Complexity
Cluster 1	weighted sum value in [0, 0.254]	Low
Cluster 2	weighted sum value in (0.254, 0.478]	Moderate
Cluster 3	weighted sum value in (0.478, 1]	High

Table 23: Scenario 2 - Threshold values for complexity levels

5.3 Assessing the complexity of a new model

The two approaches presented to evaluate complexity of BPMN models can be used for assessing the complexity of future models. Both methods aim to determine when a model is considered highly complex, which in turn allows for deciding on a model's capability for Redesign. Highly complex models, when identified, are most likely in need of Normalization. Through Normalization someone can achieve a model's complexity reduction and, therefore such models can be eligible for Redesign consideration. In the following, we present the process of assessing the complexity of a new model, using both methods proposed earlier.

5.3.1 Assessment Process using Euclidean Distance

According to the first method, a cluster analysis is performed to group the instances to three clusters. These clusters are represented by centroids, defined by specific coordinates. The model created is able to categorize a new instance by calculating its Euclidean distance from each centroid, thus deciding which cluster better represents that instance.

In this particular case, the development of the model aims at categorizing instances based on their complexity. Taking into account related research on complexity metrics, it becomes possible to recognize the level of complexity each cluster represents. For this reason, the clusters correspond to a complexity level, with Cluster 1 denoting low complexity, Cluster 2 moderate complexity and Cluster 3 high complexity, as previously established. High Complexity reduces the capability of processes for Redesign, thus highly complex models are in need of Normalization to reduce excessive complexity.

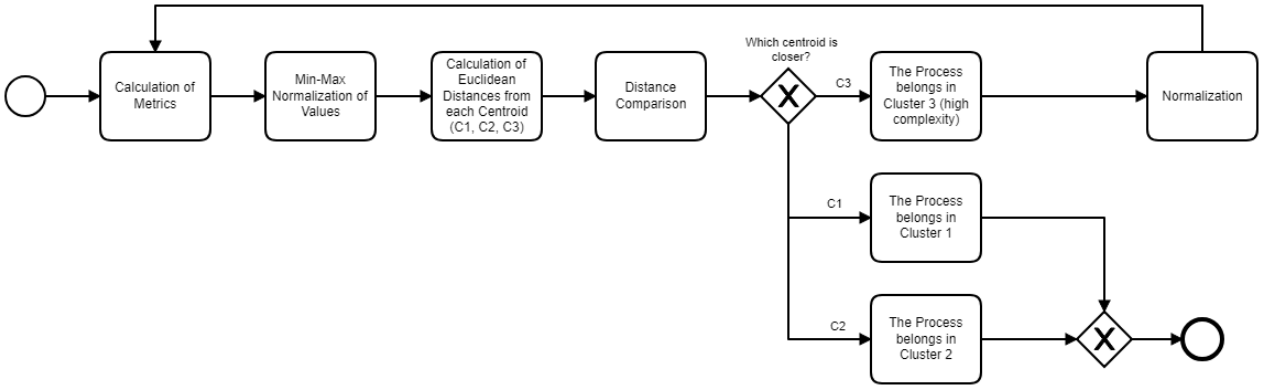

Figure 46 : BPMN model for Assessing Complexity using Euclidean Distance

The process of assessing a model's complexity using the cluster analysis is displayed in Figure 46. The steps of this process are:

1. Calculation of metrics
2. Min-Max normalization of the metric values
3. Calculation of Euclidean distance from each centroid (C1, C2, C3) derived from the cluster analysis
4. Comparison of the distances to decide which centroid is closer
5. Assign model to the most similar cluster
6. Model assigned to highly complex cluster → Normalization

5.3.2 Assessment Process using Weighted Sum Thresholds

The second method proposed for evaluating BP complexity involves the introduction of a weighted sum metric. Compared to the cluster analysis, the advantage of this method lies on two important facts. Firstly, this method leverages a weighting mechanism, that enables the assigning of priorities

to the features, i.e., metrics. In this way, the model is customizable, considering that it offers the capability for the researcher to change the priorities of each metric according to the purpose of the analysis. For instance, one might not consider size as an unimportant factor for a complexity analysis and prefers to assign higher priority to other metrics, regarding gateway complexity and structure. This method offers the option of customizing the model to meet certain requirements. Furthermore, by prioritizing metrics and inspecting the impact the different priorities have on complexity, the identification of the problematic metrics becomes feasible.

Secondly, this method facilitates the extraction of specific threshold values. The first approach, which includes cluster analysis offers reference values, in the form of centroids, that assign a process model to a cluster. This method, while leveraging cluster analysis, at the same time enables the visualization of the centroid values in one straight line. This way the definition of threshold values becomes possible, allowing a straightforward comparison.

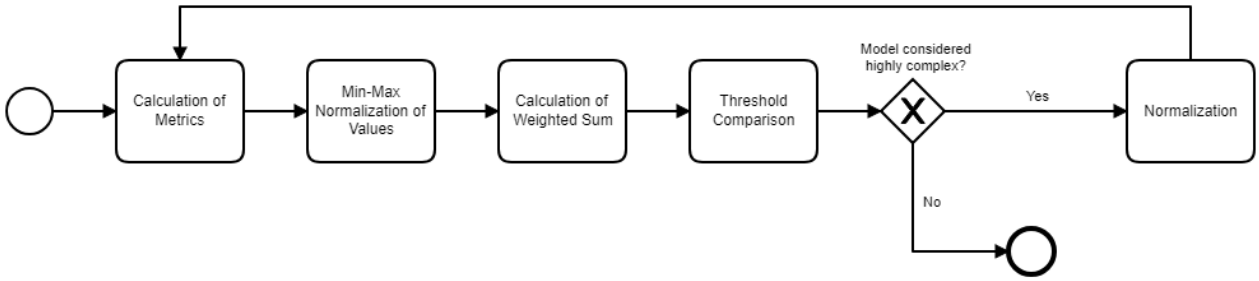

Figure 47: BPMN model for Assessing Complexity using Weighted Sum Thresholds

The general process of assessing complexity using the weighted sum method is described in Figure 47. The steps of this process are as follows:

1. Calculation of metrics
2. Min-Max normalization of the metric values
3. Calculation of weighted sum, based on the priorities used during threshold definition
4. Comparison of the weighted sum to the defined thresholds
5. Assign complexity level to model
6. Model assigned to highly complex cluster → Normalization

5.3.3 Examples from Literature

This section features the application of the methods developed to examples deriving from relevant literature. The assessment of their complexity is performed according to the processes explained in the previous section. The first step of the assessment involves the calculation of the metric values and the standardization process, followed by the complexity evaluation according to the developed methods. Table 24 contains the calculated values for each one of the processes taken from literature.

Process	NOAJS	CFC	CNC
Evaluate Quote Process [91]	14	9	1.087
Property Valuation [92]	20	5	1.077
Healthcare Scenario [93]	26	7	1.063
Medical Assessment [94]	18	9	1.238
Loan Application [95]	7	4	0.929
Bank Account Opening [96]	22	12	1.08
Baking Workflow Process [97]	14	4	1.055
Emergency Ward of a Hospital [98]	19	9	1.190
Auction [99]	23	12	1.048
Admission Process [100]	19	6	1

Table 24: Metric Values for Example Processes

The next step is the application of the standardization process to scale the values. As previously mentioned, the process of standardization is considered of great importance, especially for the implementation of distance-based clustering algorithms [88], [89]. Min-Max normalization is performed on the metric values in order to scale them to values from 0 to 1. The normalization follows the same procedure as the one explained in chapter 4. The Min and Max values of each metric are in accordance to the dataset, with $Min(NOAJS) = 2$, $Max(NOAJS) = 53$, $Min(CFC) = 0$, $Max(CFC) = 16$, $Min(CNC) = 0.529$ and $Max(CNC) = 1.636$.

For instance, for the Evaluate Quote Process the calculation is as follows:

$$NOAJS' = \frac{NOAJS - Min(NOAJS)}{Max(NOAJS) - Min(NOAJS)} = \frac{14 - 2}{53 - 2} = 0.2353$$

$$CFC' = \frac{CFC - Min(CFC)}{Max(CFC) - Min(CFC)} = \frac{9 - 0}{16 - 0} = 0.5625$$

$$CNC' = \frac{CNC - Min(CNC)}{Max(CNC) - Min(CNC)} = \frac{1.087 - 0.529}{1.636 - 0.529} = 0.5040$$

The calculation of the normalized values for each one of the processes is performed, according to the above example. Table 25 displays the normalized values of the metrics for each one of the literature examples.

Process	NOAJS	CFC	CNC
Evaluate Quote Process [91]	0.2353	0.5625	0.5040
Property Valuation [92]	0.3529	0.3125	0.4950
Healthcare Scenario [93]	0.4706	0.4375	0.4819
Medical Assessment [94]	0.3137	0.5625	0.6406
Loan Application [95]	0.0980	0.2500	0.3609
Bank Account Opening [96]	0.3922	0.7500	0.4977
Baking Workflow Process [97]	0.2353	0.2500	0.4757
Emergency Ward of a Hospital [98]	0.3333	0.5625	0.5975
Auction [99]	0.4118	0.7500	0.4695
Admission Process[100]	0.3333	0.3750	0.4255

Table 25: Normalised Metric Values for Example Processes

Cluster analysis:

The next step for this approach is to determine which centroid best represents a model. Thus, the calculation of the Euclidean distances from each centroid is required.

Centroid 1 (0.149, 0.093, 0.363):

$$Distance1 = \sqrt{(0.149 - 0.2353)^2 + (0.093 - 0.5625)^2 + (0.363 - 0.5040)^2} = 0.498$$

Centroid 2 (0.213, 0.339, 0.447):

$$Distance2 = \sqrt{(0.213 - 0.2353)^2 + (0.339 - 0.5625)^2 + (0.447 - 0.5040)^2} = 0.232$$

Centroid 3 (0.561, 0.688, 0.580):

$$Distance3 = \sqrt{(0.561 - 0.2353)^2 + (0.688 - 0.5625)^2 + (0.580 - 0.5040)^2} = 0.357$$

The closest distance is from centroid 2, so the process models is assigned to Cluster 2. Thus, this process model is classified as of Moderate complexity, since it is similar to centroid 2, which represents moderate complexity BP models. In the same way, a complexity level is assigned to every example. Table 26 displays the assessment results for all processes.

Weighted Sum – Scenario 1:

This approach involves equal priority for all metrics. According to this method the next step is to calculate the Weighted sum considering the priorities assigned during the development of the model. Since the priorities are equal the weights are the same for each metric, which means $w_{1,2,3} = 0.3333$. For example, for the Evaluate Quote Process:

$$WS = \sum_{i=1}^{3} w_i v_i = 0.3333 * 0.2353 + 0.3333 * 0.5625 + 0.3333 * 0.5040 = 0.4339$$

Comparing the WS with the thresholds extracted for this method and displayed in Figure 26, it becomes apparent that this process belongs in Cluster 2, defined by the range (0.268, 0.484]. Since this cluster represents models of moderate complexity, this complexity level is assigned to this process.

Weighted Sum – Scenario 2:

The second Scenario considers higher priority for the CFC metric and lower for the NOAJS metric. The calculation of the Weighted sum accounts for the priorities assigned during the development of the model. The weights for each metric shape differently, in this specific example the weights are $w_1 = 0.0625, w_2 = 0.625, w_3 = 0.3125$. For instance, for the Evaluate Quote Process:

$$WS = \sum_{i=1}^{3} w_i v_i = 0.0625 * 0.2353 + 0.625 * 0.5625 + 0.3125 * 0.5040 = 0.5237$$

The result of the WS indicates that, when compared to the established thresholds in Figure 42, the process model belongs in Cluster 3. This Cluster is defined by the range (0.478, 1] and represents highly complex models, therefore high complexity level is assigned to the process. It appears that increasing the priority of the CFC metric impacted the result of the assessment, since the complexity assessed as moderate in Scenario 1, is regarded high in Scenario 2. This is expected, considering the value of that metric (CFC = 9), which is quite high compared to the size of the process (NOAJS = 14). This simple example demonstrates the potential of identifying the problematic areas of models, by analyzing the influence of each metric on the overall complexity. In the examined case, it is possible to deduce that this process has an increased control-flow complexity, that can be attributed to many gateways. A possible course of action for reducing complexity should involve the reduction of gateway splits, as revealed by the analysis.

	Complexity Level		
Process	Cluster Analysis	Weighted Sum – Scenario 1	Weighted Sum – Scenario 2
Evaluate Quote Process [91]	Moderate	Moderate	High
Property Valuation [92]	Moderate	Moderate	Moderate
Healthcare Scenario [93]	Moderate	Moderate	Moderate
Medical Assessment [94]	High	High	High
Loan Application [95]	Low	Low	Moderate
Bank Account Opening [96]	High	High	High
Baking Workflow Process [97]	Moderate	Moderate	Moderate
Emergency Ward of a Hospital [98]	High	High	High
Auction [99]	High	High	High
Admission Process [100]	Moderate	Moderate	Moderate

Table 26: Assessment Results

In Table 26 the results of the assessment for the 10 process models examined are presented. It appears that the assessment produces the same results using cluster analysis and the weighted sum measure with equal priorities. Essentially these two methods are developed in a way that it is expected to produce similar results, with the difference that the weighted sum method facilitates the extraction of thresholds, allowing for a more straightforward comparison. However, this is not the case for the methods that consider different priorities. It is easily observed, that between the two Scenarios the complexity levels of two processes shift from moderate to high and from low to moderate respectively. This outcome relates to the increased priority granted to the CFC metric, which in turn raises the weighted sum measure accordingly.

5.4 Summary

For the purpose of measuring end evaluating Business Process Complexity, two methods are proposed in this chapter. The first method includes a cluster analysis of selected features, i.e., complexity metrics, to determine a model that efficiently categorizes the instances into clusters and offers reference values, in the form of centroids, for future categorization. By combining the clustering results, centroid values and literature findings, it becomes possible to distinguish clusters based on their complexity level. The second approach evaluates complexity through a weighted sum metric, that combines the selected metrics to one single measure. This approach offers the added option of assigning priority to metrics, that proves useful for highlighting the impact of each metric to the perceived complexity of a model. Additionally, it facilitates the extraction of exact thresholds for the combined measure, in contrast to the first method that offers reference values.

CHAPTER 6

Discussion & Conclusions

6.1 Thesis Overview

The aim of this Thesis, as stated in Chapter 1, involves the development of methods able to categorize a Business Process model regarding its complexity. The main purpose of these methods is to efficiently assess a model's complexity in relation to its capability for Redesign. Following this reasoning, these methods also reveal the need for addressing excessive complexity by utilizing normalization techniques.

For making this thesis objective feasible, the selection of suitable metrics that quantify the notion of complexity is required. Several such metrics have been proposed in literature over the years. Through the scope of Redesign, five of them are initially selected to be combined in an attempt to cover all recognized aspects of complexity. These metrics refer to activity complexity, control flow complexity and structural complexity and specific criteria are considered for their selection. These criteria include (1) the generic definition of the metrics, that allow them to be adapted to the BPMN standard, (2) their adequate empirical validation and (3) their popularity in related research. Following an initial analysis, three of these metrics are found to serve the purpose of this study and are used for the development of the assessment methods.

Taking into account that the database used as a source for the BP models contains models represented in BPMN, the calculation formulas of the selected metrics need to be adapted to this specific technique. For this reason, making assumptions regarding the calculations is considered essential. Apart from the calculation of the metric values, them being in the same scale is also a necessity; in this way the impact of each metric on the proposed methods can be balanced. Applying the Min-Max Normalization technique results in a dataset that is appropriate for the imminent cluster analyses.

Leveraging the complexity metrics selected from literature, along with the formation of the appropriate dataset, the development of the methods becomes feasible. Cluster analysis is regarded as a suitable method to divide the models into categories in relation to their complexity, considering the fact the initial dataset does not include labels. The first approach involves the utilization of cluster analysis on the selected metrics in order to identify the different complexity categories of the BPMN models. The second approach introduces a priority assignment mechanism that facilitates the combination of the selected metrics to one single measure. Essentially, a weighted sum is calculated shifting the problem from three dimensions to one, thus enabling the extraction of threshold values.

The trained models developed are tested on ten BPMN models from literature to further exemplify their usage and, moreover, examine their behavior. The assessment processes are followed in detail for the three scenarios, i.e., cluster analysis, equally weighted sum and unequally weighted sum. Each example is categorized according to a complexity level, based on either the reference values or the thresholds defined. The results indicate that the proposed approaches achieve mostly similar categorizations. However, it should be noted that the unbalanced prioritizing of metrics is found to alter, in some cases, the results of the assessment. These modifications are substantially based on the priorities set.

6.2 Research Contribution

This thesis main contribution concerns the development of methods, able to evaluate the complexity of BPMN models. The scope of the research focuses on recognizing the Redesign capabilities of a BPMN model. In cases of high complexity, a Business Process' capability for redesign is limited and complexity reducing mechanisms, i.e., Normalization, need to be employed. For this reason, the proposed methods categorize the models in relation to their complexity, attempting to identify highly complex ones.

The first method proposes the implementation of a clustering technique, namely the K-means algorithm, on a dataset comprising of selected complexity metrics from literature. The values of the selected metrics for BPMN models, calculated and normalized, form the dataset required for the implementation of the method. The models are grouped into three clusters, each one represented by a centroid, according to the algorithm. Upon inspecting the centroids for each cluster, in conjunction with related work on complexity measurement, we deduce that each centroid, hence

each cluster, represents a complexity level. The complexity levels derive from the values of the centroids, since it is acknowledged that higher values for the selected metrics translate into high complexity. This method provides reference values, in the form of centroids, that allow the assessment of future BPMN models. When assessing the complexity of a new model, the proximity from each reference value determines how the model will be categorized.

The second method introduces the combination of the selected complexity metrics to one single weighted sum. Essentially, this approach limits the dimensions of the data from three to one, while retaining the information provided by the metrics. Through the application of the K-means algorithm on the one weighted sum measure considering an equal weighting for all metrics, a clustering is formed. Following the same reasoning as before, each cluster signifies a complexity level, according to its representative, i.e., the centroid value. In this case, the categorization of a new BPMN model to a complexity level is performed through the definition of threshold values, since the one-dimensional nature of the data allows for a representation on a single line. The categories, i.e., clusters, are separated by the means of the intervals formed between two adjacent centroids and, in essence, act as thresholds. Therefore, this method constitutes a simpler and more straightforward way of assessing complexity.

Nevertheless, this is not the only advantage of the weighted sum measure proposed. Through this method, the assignment of priorities to the metrics is possible. By altering the priorities, the weights of each metric adjust accordingly, meaning that the value of a metric impacts, less or more, the final results. The implementation of a clustering on an unequally weighted sum, enables the definition of new thresholds, creating a new model for assessing complexity. This additional capability offered by this method may be useful to customize the approach a researcher follows on complexity and to configure the model to their specific purpose.

However, assigning priorities to the metrics exhibits another very important capability. By adjusting the priorities of the metrics and analyzing the impact of that modification on their perceived complexity, the identification of problematic areas is facilitated. For instance, if increasing the priority of a metric also increases the complexity of a process model, forcing it to change category, this constitutes a clear indication of this metric's impact on the overall complexity. In general, this method, through highlighting specific metrics, facilitates the identification of the particular aspects

of a model affecting complexity. This approach is considered very useful for revealing the parts of a process model in need of normalization.

Additional contributions of this research include, firstly, the detailed review of the complexity metrics proposed in literature and the selection of the appropriate ones for representing complexity regarding a process's capability for transformation and redesign. Secondly, another contribution involves the adaptation of the selected complexity metrics from their generic form to the BPMN standard, by defining specific assumptions for their calculation. Last, the cleaning of the SOA-Business Process Database and the formation of the dataset, containing the calculated values of the selected metrics, which was required for the extraction of the reference values and thresholds, constitutes an added contribution.

6.3 Research Limitations & Future Work

The methods proposed are considered an efficient initial approach on complexity assessment of business process models, using cluster analysis. However, there are several limitations regarding the application of the methods. Firstly, the developed methods leverage a number of metrics from literature, following a literature review on complexity measurements. From the multitude of defined metrics only five of them are utilized for the premise of this thesis. Essentially, the selection criteria, set by the author for the selection of specific metrics, largely depend on the interpretation of complexity.

In addition, the development of the methods involves the utilization of a data mining technique, particularly the implementation of the K-means algorithm for clustering. Such learning algorithms produce a model based on instances of data; hence they are largely dependent on the dataset provided. The results of the algorithms, the reference values and threshold values defined, are influenced in a significant degree by the nature of the dataset. In this case, the dataset used is a relatively small one, consisting of 87 Business process models expressed in BPMN, which constitutes another limitation of the current research.

An additional limitation regards the number of categories used to describe complexity. This number, i.e., three complexity categories, corresponds to the number of clusters to be formed by the algorithm and is an input parameter provided by the author, according to the purpose of the analysis. The number of instances in the dataset and the purpose of the analysis, along with related

work on complexity, led to the selection of this particular parameter value. Further partitioning of the complexity levels, for example considering four or five complexity levels, would yield different results.

Future work on complexity assessment could be directed towards tackling these limitations. The application of the proposed methods on a larger dataset, containing business process models representative of industry standards, could provide a more accurate assessment of complexity. In this way, the definition of reference values and thresholds would obtain a more established standing.

Moreover, the complexity metrics considered for the present research could expand, including additional metrics that quantify complexity from many different aspects. The selection of the appropriate metrics is closely related to the purpose of each analysis. Given that the aim of this research focuses on the redesign capability of a process, the suitable metrics were selected by the author. The same methods could apply to a different set of metrics, pertinent to the context of each analysis. On top of that, the introduction of the weighting mechanism allows for experimentation with regard to prioritizing metrics and highlighting different aspects of complexity.

6.4 Conclusions

The present thesis develops two distinct methods of assessing the complexity of BPMN models regarding their redesign capability, that leverage established complexity metrics. Both methods utilize a clustering technique, to group the models into clusters that correspond to complexity categories. The first approach on complexity assessment regards a cluster analysis on a selected set of metrics, that facilitates the identification of highly complex models. The clusters, each represented by their centroid, are interpreted to a complexity level, namely low, moderate or high, based on related literature. The assessment of the complexity of new BPMN models is achieved through the establishment of the centroids as reference values for each category, i.e., cluster.

The second approach introduces a combination of the selected metrics to one single weighted sum measure. This approach offers the advantage of a more straightforward complexity assessment, through the definition of specific threshold values, to categorize future models. Additionally, this approach provides the possibility of prioritizing among metrics, by changing the influence each metric has on the proposed measure. Thereby, the effect of each metric on the overall complexity

can be clarified, revealing at the same time, the aspects of the model having the strongest impact on complexity. Thus, apart from a way to assess complexity, this method also acts as a means to determine which parts of a process model suffer from increased complexity and require normalization.

References

[1] M. Weske, *Business Process Management: Concepts, Languages, Architectures*. Springer Science & Business Media, 2012.

[2] K. Vergidis, "Business process optimisation using an evolutionary multi-objective framework," 2008.

[3] M. Hammer and J. Champy, "A manifesto for business revolution," *Reengineering the Corporation*, 1993.

[4] T. H. Davenport, *Process innovation: reengineering work through information technology*. Harvard Business Press, 1993.

[5] K. D. Swenson, M. von Rosing, M. Von Rosing, H. Von Scheel, and A. W. Scheer, *Phase 4: What Is Business Process Management?* 2015.

[6] A. Lindsay, D. Downs, and K. Lunn, "Business processes—attempts to find a definition," *Information and software technology*, vol. 45, no. 15, pp. 1015–1019, 2003.

[7] J. Recker, M. Rosemann, M. Indulska, and P. Green, "Business process modeling-a comparative analysis," *Journal of the Association for Information Systems*, vol. 10, no. 4, p. 1, 2009.

[8] W. M. Van der Aalst, "Business process management: a comprehensive survey," *ISRN Software Engineering*, vol. 2013, 2013.

[9] G. M. Giaglis, "A taxonomy of business process modeling and information systems modeling techniques," *International Journal of Flexible Manufacturing Systems*, vol. 13, no. 2, pp. 209–228, 2001.

[10] K. Vergidis, A. Tiwari, and B. Majeed, "Business process analysis and optimization: Beyond reengineering," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 38, no. 1, pp. 69–82, 2007.

[11] W. M. Van der Aalst, "Formalization and verification of event-driven process chains," *Information and Software technology*, vol. 41, no. 10, pp. 639–650, 1999.

[12] J. Mendling, "Event-driven process chains (epc)," in *Metrics for process models*, Springer, 2008, pp. 17–57.

[13] K. Figl, J. Mendling, and M. Strembeck, "Towards a usability assessment of process modeling languages," in *8th GI-Workshop Geschäftsprozessmanagement mit Ereignisgesteuerten Prozessketten (EPK), CEUR-WS*, 2009, vol. 554, pp. 138–156.

[14] O. A. Specification, "Omg unified modeling language (omg uml), superstructure, v2. 1.2," *Object Management Group*, vol. 70, 2007.

[15] P. Wohed, W. M. van der Aalst, M. Dumas, A. H. ter Hofstede, and N. Russell, "Pattern-based analysis of UML activity diagrams," *Beta, Research School for Operations Management and Logistics, Eindhoven*, 2004.

[16] M. Dumas and A. H. Ter Hofstede, "UML activity diagrams as a workflow specification language," in *International conference on the unified modeling language*, 2001, pp. 76–90.

[17] C. Menzel and R. J. Mayer, "The IDEF family of languages," in *Handbook on architectures of information systems*, Springer, 1998, pp. 209–241.

[18] "IDEFØ – Function Modeling Method – IDEF." [Online]. Available: http://www.idef.com/idefo-function_modeling_method/. [Accessed: 10-Oct-2019].

[19] O. S. Noran, "Business modelling: UML vs. IDEF," *School of Computing and Information Technology, Griffith University*, 2000.
[20] V. Bosilj-Vuksic, G. M. Giaglis, and V. Hlupic, "IDEF diagrams and petri nets for business process modeling: suitability, efficacy, and complementary use," in *Enterprise information systems II*, Springer, 2001, pp. 143–148.
[21] S. A. White, "Introduction to BPMN," *Ibm Cooperation*, vol. 2, no. 0, p. 0, 2004.
[22] "About the Business Process Model And Notation Specification Version 2.0." [Online]. Available: https://www.omg.org/spec/BPMN/2.0. [Accessed: 21-Oct-2019].
[23] S. A. White and C. Bock, *BPMN 2.0 Handbook Second Edition: Methods, Concepts, Case Studies and Standards in Business Process Management Notation*. Future Strategies Inc., 2011.
[24] E. Rolón, F. Ruiz, F. García, and M. Piattini, "Applying software metrics to evaluate business process models," *CLEI-Electronic Journal*, vol. 9, no. 1, 2006.
[25] M. Chinosi and A. Trombetta, "BPMN: An introduction to the standard," *Computer Standards & Interfaces*, vol. 34, no. 1, pp. 124–134, 2012.
[26] M. Dumas, M. La Rosa, J. Mendling, and H. A. Reijers, "Process Redesign," in *Fundamentals of Business Process Management*, Springer, 2018, pp. 297–339.
[27] J. Cardoso, J. Mendling, G. Neumann, and H. A. Reijers, "A discourse on complexity of process models," in *International Conference on Business Process Management*, 2006, pp. 117–128.
[28] J. Mendling, *Metrics for process models: empirical foundations of verification, error prediction, and guidelines for correctness*, vol. 6. Springer Science & Business Media, 2008.
[29] L. Reynoso, E. Rolón, M. Genero, F. García, F. Ruiz, and M. Piattini, "Formal definition of measures for BPMN models," in *International Workshop on Software Measurement*, 2009, pp. 285–306.
[30] K. Kluza and G. J. Nalepa, "Proposal of square metrics for measuring business process model complexity," in *2012 Federated Conference on Computer Science and Information Systems (FedCSIS)*, 2012, pp. 919–922.
[31] J. Mendling, M. Dumas, M. La Rosa, and H. A. Reijers, "Structuring Business Process Management," in *The Art of Structuring*, Springer, 2019, pp. 203–211.
[32] N. Abdi, B. Zarei, J. Vaisy, and B. Parvin, "Innovation models and business process redesign," *International Business and Management*, vol. 3, no. 2, pp. 147–152, 2011.
[33] H. A. Reijers and S. L. Mansar, "Best practices in business process redesign: an overview and qualitative evaluation of successful redesign heuristics," *Omega*, vol. 33, no. 4, pp. 283–306, 2005.
[34] R. J. Vanwersch *et al.*, "A critical evaluation and framework of business process improvement methods," *Business & Information Systems Engineering*, vol. 58, no. 1, pp. 43–53, 2016.
[35] Y. Yu, A. Pelaez, and K. R. Lang, "Designing and evaluating business process models: an experimental approach," *Information Systems and e-Business Management*, vol. 14, no. 4, pp. 767–789, 2016.
[36] N. Kock, J. Verville, A. Danesh-Pajou, and D. DeLuca, "Communication flow orientation in business process modeling and its effect on redesign success: Results from a field study," *Decision Support Systems*, vol. 46, no. 2, pp. 562–575, 2009.
[37] G. Tsakalidis, K. Vergidis, G. Kougka, and A. Gounaris, "Eligibility of BPMN models for business process redesign," *Information*, vol. 10, no. 7, p. 225, 2019.
[38] I. Vanderfeesten, H. A. Reijers, J. Mendling, W. M. van der Aalst, and J. Cardoso, "On a quest for good process models: The cross-connectivity metric," in *International Conference on Advanced Information Systems Engineering*, 2008, pp. 480–494.

[39] F. Corradini *et al.*, "A guidelines framework for understandable BPMN models," *Data & Knowledge Engineering*, vol. 113, pp. 129–154, 2018.
[40] I. M.-M. de Oca, M. Snoeck, H. A. Reijers, and A. Rodríguez-Morffi, "A systematic literature review of studies on business process modeling quality," *Information and Software Technology*, vol. 58, pp. 187–205, 2015.
[41] G. Polančič and B. Cegnar, "Complexity metrics for process models – A systematic literature review," *Computer Standards & Interfaces*, vol. 51, pp. 104–117, Mar. 2017.
[42] J. Cardoso, "Control-flow complexity measurement of processes and Weyuker's properties," in *6th International Enformatika Conference*, 2005, vol. 8, pp. 213–218.
[43] A. Geraci *et al.*, *IEEE standard computer dictionary: Compilation of IEEE standard computer glossaries*. IEEE Press, 1991.
[44] J. Mendling, H. A. Reijers, and J. Cardoso, "What makes process models understandable?," in *International Conference on Business Process Management*, 2007, pp. 48–63.
[45] L. Sánchez-González, F. García, J. Mendling, F. Ruiz, and M. Piattini, "Prediction of business process model quality based on structural metrics," in *International Conference on Conceptual Modeling*, 2010, pp. 458–463.
[46] L. Sánchez-González, F. García, F. Ruiz, and J. Mendling, "Quality indicators for business process models from a gateway complexity perspective," *Information and Software Technology*, vol. 54, no. 11, pp. 1159–1174, 2012.
[47] K. Kluza, G. J. Nalepa, and J. Lisiecki, "Square complexity metrics for business process models," in *Advances in Business ICT*, Springer, 2014, pp. 89–107.
[48] A. Dikici, O. Turetken, and O. Demirors, "Factors influencing the understandability of process models: A systematic literature review," *Information and Software Technology*, vol. 93, pp. 112–129, 2018.
[49] F. Yahya, K. Boukadi, and H. Ben-Abdallah, "Improving the quality of Business Process Models: Lesson learned from the State of the Art," *Business Process Management Journal*, Dec. 2018.
[50] R. Laue and J. Mendling, "Structuredness and its significance for correctness of process models," *Information Systems and E-Business Management*, vol. 8, no. 3, pp. 287–307, 2010.
[51] M. Dumas, M. La Rosa, J. Mendling, R. Mäesalu, H. A. Reijers, and N. Semenenko, "Understanding business process models: the costs and benefits of structuredness," in *International Conference on Advanced Information Systems Engineering*, 2012, pp. 31–46.
[52] K. Figl and R. Laue, "Influence factors for local comprehensibility of process models," *International Journal of Human-Computer Studies*, vol. 82, pp. 96–110, 2015.
[53] J. Cardoso, "Business process control-flow complexity: Metric, evaluation, and validation," *International Journal of Web Services Research (IJWSR)*, vol. 5, no. 2, pp. 49–76, 2008.
[54] H. Leopold, J. Mendling, and O. Günther, "Learning from quality issues of BPMN models from industry," *IEEE Software*, vol. 33, no. 4, pp. 26–33, 2015.
[55] M. La Rosa, P. Wohed, J. Mendling, A. H. Ter Hofstede, H. A. Reijers, and W. M. van der Aalst, "Managing process model complexity via abstract syntax modifications," *IEEE Transactions on Industrial Informatics*, vol. 7, no. 4, pp. 614–629, 2011.
[56] K. Kluza and K. Kaczor, "Overview of BPMN model equivalences: towards normalization of BPMN diagrams," in *8th Workshop on Knowledge Engineering and Software Engineering (KESE2012) at the at the biennial European Conference on Artificial Intelligence (ECAI 2012): August*, 2012, vol. 28, pp. 38–45.
[57] J. Mendling, H. A. Reijers, and W. M. van der Aalst, "Seven process modeling guidelines (7PMG)," *Information and Software Technology*, vol. 52, no. 2, pp. 127–136, 2010.

[58] J. Mendling, G. Neumann, and W. Van Der Aalst, "Understanding the occurrence of errors in process models based on metrics," in *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, 2007, pp. 113–130.
[59] M. La Rosa *et al.*, "APROMORE: An advanced process model repository," *Expert Systems with Applications*, vol. 38, no. 6, pp. 7029–7040, 2011.
[60] W. Khlif, H. Ben-Abdallah, and N. E. Ben Ayed, "A methodology for the semantic and structural restructuring of BPMN models," *Business Process Management Journal*, vol. 23, no. 1, pp. 16–46, 2017.
[61] W. Khlif, N. E. B. Ayed, and H. Ben-Abdallah, "EVARES: A Quality-driven Refactoring Method for Business Process Models.," in *ICEIS (3)*, 2017, pp. 409–416.
[62] P.-N. Tan, M. Steinbach, and V. Kumar, "Data mining cluster analysis: basic concepts and algorithms," *Introduction to data mining*, pp. 487–533, 2013.
[63] V. Estivill-Castro, "Why so many clustering algorithms: a position paper.," *SIGKDD explorations*, vol. 4, no. 1, pp. 65–75, 2002.
[64] A. Amelio and A. Tagarelli, "Data Mining: Clustering," *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, p. 437, 2018.
[65] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to data mining*. Pearson Education India, 2016.
[66] Xin-She Yang, *Introduction to Algorithms for Data Mining and Machine Learning*. Elsevier, 2019.
[67] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern recognition letters*, vol. 31, no. 8, pp. 651–666, 2010.
[68] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1967, vol. 1, pp. 281–297.
[69] S. Lloyd, "Least squares quantization in PCM," *IEEE transactions on information theory*, vol. 28, no. 2, pp. 129–137, 1982.
[70] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, 2007, pp. 1027–1035.
[71] J. M. Pena, J. A. Lozano, and P. Larranaga, "An empirical comparison of four initialization methods for the k-means algorithm," *Pattern recognition letters*, vol. 20, no. 10, pp. 1027–1040, 1999.
[72] P. Berkhin, "A survey of clustering data mining techniques," in *Grouping multidimensional data*, Springer, 2006, pp. 25–71.
[73] G. Muketha, "A survey of business processes complexity metrics," 2010.
[74] L. Sánchez González, F. García Rubio, F. Ruiz González, and M. Piattini Velthuis, "Measurement in business processes: a systematic review," *Business Process Management Journal*, vol. 16, no. 1, pp. 114–134, 2010.
[75] C. Jones, *Programming Productivity*, 1st edition. New York: McGraw-Hill College, 1986.
[76] T. J. McCabe, "A complexity measure," *IEEE Transactions on software Engineering*, no. 4, pp. 308–320, 1976.
[77] A. M. Latva-Koivisto, "Finding a complexity measure for business process models," *Helsinki University of Technology, Systems Analysis Laboratory*, 2001.
[78] I. T. Vanderfeesten, J. S. Cardoso, and H. A. Reijers, "A weighted coupling metric for business process models.," in *CAiSE Forum*, 2007, vol. 247, pp. 41–44.

[79] K. Kluza, “Measuring complexity of business process models integrated with rules,” in *International Conference on Artificial Intelligence and Soft Computing*, 2015, pp. 649–659.
[80] J. Mendling, L. Sánchez-González, F. García, and M. La Rosa, “Thresholds for error probability measures of business process models,” *Journal of Systems and Software*, vol. 85, no. 5, pp. 1188–1197, 2012.
[81] L. Sánchez-González, F. García, F. Ruiz, and M. Piattini, “A case study about the improvement of business process models driven by indicators,” *Softw Syst Model*, vol. 16, no. 3, pp. 759–788, Jul. 2017.
[82] M. Fernández-Ropero, R. Pérez-Castillo, I. Caballero, and M. Piattini, “Quality-driven business process refactoring,” in *International Conference on Business Information Systems (ICBIS 2012)*, 2012, pp. 960–966.
[83] L. Makni, W. Khlif, N. Zaaboub Haddar, and H. Ben-Abdallah, “A tool for evaluationg the quality of business process models,” *INFORMATIK 2010–Business Process and Service Science–Proceedings of ISSS and BPSC*, 2010.
[84] J. Cardoso, “Evaluating the process control-flow complexity measure,” in *IEEE International Conference on Web Services (ICWS’05)*, 2005.
[85] E. Rolón, J. Cardoso, F. García, F. Ruiz, and M. Piattini, “Analysis and validation of control-flow complexity measures with bpmn process models,” in *Enterprise, Business-Process and Information Systems Modeling*, Springer, 2009, pp. 58–70.
[86] L. Sánchez-González, F. García, J. Mendling, and F. Ruiz, “Quality assessment of business process models based on thresholds,” in *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, 2010, pp. 78–95.
[87] “BPOSCTeam SOA-Based Business Process Database.” [Online]. Available: https://sites.google.com/site/bposcteam2015/ressources. [Accessed: 10-Oct-2019].
[88] L. Al Shalabi, Z. Shaaban, and B. Kasasbeh, “Data mining: A preprocessing engine,” *Journal of Computer Science*, vol. 2, no. 9, pp. 735–739, 2006.
[89] I. B. Mohamad and D. Usman, “Standardization and its effects on K-means clustering algorithm,” *Research Journal of Applied Sciences, Engineering and Technology*, vol. 6, no. 17, pp. 3299–3303, 2013.
[90] A. Kassambara, *Practical guide to cluster analysis in R: Unsupervised machine learning*, vol. 1. STHDA, 2017.
[91] J. Kolar, L. Dockal, and T. Pitner, “A Dynamic Approach to Process Design: A Pattern for Extending the Flexibility of Process Models,” in *IFIP Working Conference on The Practice of Enterprise Modeling*, 2013, pp. 176–190.
[92] U. Kannengiesser and L. Zhu, “Towards Concise Architectures for Flexible Business Processes,” *Advanced Engineering Informatics*, pp. 61–75, 2011.
[93] D. Knuplesch, M. Reichert, W. Fdhila, and S. Rinderle-Ma, “On enabling compliance of cross-organizational business processes,” in *Business Process Management*, Springer, 2013, pp. 146–154.
[94] L. T. Herbert, R. Sharp, and M. R. Hansen, “Specification, Verification and Optimisation of Business Processes: A Unified Framework,” 2014.
[95] A. Rogge-Solti and M. Weske, “Prediction of business process durations using non-Markovian stochastic Petri nets,” *Information Systems*, vol. 54, pp. 1–14, 2015.
[96] P. Poizat, G. Salaün, and A. Krishna, “Checking Business Process Evolution,” presented at the International Workshop on Formal Aspects of Component Software (pp. 36-53), Besançon, France, 2016.

[97] L. Herbert, Z. N. L. Hansen, P. Jacobsen, and P. Cunha, "Evolutionary optimization of production materials workflow processes," *Procedia CIRP*, vol. 25, pp. 53–60, 2014.

[98] F. Mannhardt, M. de Leoni, H. A. Reijers, and W. M. van der Aalst, "Data-driven process discovery-revealing conditional infrequent behavior from event logs," in *International Conference on Advanced Information Systems Engineering*, 2017, pp. 545–560.

[99] S. Fan, H. Zhimin, V. C. Storey, and J. L. Zhao, "A process ontology based approach to easing semantic ambiguity in business process modeling," *Data & Knowledge Engineering*, no. 102, pp. 57–77, 2016.

[100] Z. A. Bukhsh, M. van Sinderen, K. Sikkel, and D. A. Quartel, "Understanding Modeling Requirements of Unstructured Business Processes.," in *ICE-B*, 2017, pp. 17–27.

Appendix A - Selected BPMN models from SOA-based Database

In this Chapter information about the BPMN models used for the presented analysis is available. These models are selected from the SOA-based Business Process Database based on the criteria described in chapter 4 and are titled according to their context. Table 27 features the index number of each process in the SOA-based Business Process Database, followed by the Title of each process.

NUM	PROCESS	TITLE
1	9	DEBUGGING
2	11	TRADING PROCESS
3	13	TROUBLE TICKET SYSTEM
4	15	DEMAND MEETING
5	17	HARDWARE SHIPMENT
6	18	VACATION REQUEST APPROVAL
7	19	APPLICATION HANDLING
8	20	CAB SERVICE
9	21	CAR RENTAL
10	23	BLOG POST PUBLISHMENT
11	29	SUPPLY PROCUREMENT
12	32	CUSTOMER ISSUE
13	33	BUG REPORT HANDLING
14	34	PURCHASE ORDER APPROVEMENT
15	41	PURCHASE WITH RFQ
16	42	UNSUCCESSFUL WIZARD PROCESS
17	44	PURCHASE ORDER DELIVERY CHECKING
18	47	PURCHASE REQUISITION
19	50	TEACHING
20	53	CHARITY
21	86	HOTEL AND FLIGHT RESERVATION
22	87	ISSUE DISCUSION AND VOTING
23	109	SOCIAL NETWORK CONECTION
24	110	SOCIAL MEDIA CAMPAIGN INTERACTION
25	111	THESIS PROPOSAL
26	113	JOB POSTING CREATION
27	114	TRAVEL BOOKING
28	116	SUPPLY ORDER
29	120	PURCHASE REQUEST
30	122	ACCOUNT OPENING
31	123	JOB APPLICATION PROCESS

32	124	ISSUE DISCUSSION MODERATION
33	125	BOOK ORDER ONLINE
34	126	PATIENT ADMISSION TO HEALTH AGENCY
35	131	PRODUCT DEVELOPMENT AND LAUNCH
36	136	PAPER PUBLISHMENT
37	138	INSURANCE POLICY ISSUANCE
38	139	CREDIT CARD TRANSACTION
39	144	JOB VACANCY APPROVAL
40	147	QUOTATION MANAGEMENT SYSTEM
41	148	PROJECT ASSIGNMENT TO CONTRACTOR
42	153	FLIGHT AND HOTEL BOOKING
43	156	RESTOCKING WAREHOUSE
44	157	LOAN APPROVAL
45	167	MEDIA CONTENT PUBLISHMENT
46	169	POLICY APPROVAL
47	171	CREDIT APPLICATION
48	175	ORDER HANDLING
49	177	SOFTWARE PRODUCT DEVELOPMENT CYCLE
50	180	MAINTENANCE PROBLEM REPORT
51	181	UPDATE DEFAULTING CUSTOMERS' TABLE
52	183	MORTAGE APPLICATION
53	184	CALENDAR SCHEDUALING
54	188	ELECTRICAL DESIGN DEVELOPMENT
55	196	CUSTOMIZED PC PURCHASE
56	198	FLIGHT CHECK IN PROCEDURE
57	209	MARKETING CAMPAIGN
58	210	CLOUD-BASED HEALTH MONITORING
59	213	ADVERTISEMENT ON WEBSITES
60	214	NEW EMPLOYEE PROCEDURES
61	215	COURSE STRUCTURE SURVEY
62	220	BAG PRODUCTION
63	221	STRATEGIC PLANNING
64	222	BPM INTEGRATION PROJECT
65	232	ERROR HANDLING RULES UPDATE
66	234	PATIENT RECORD UPDATE
67	235	PATIENT EXAMINATION
68	241	DATA CENTER DEPLOYMENT
69	501	PIZZA ORDER
70	557	ORDER TO CASH
71	575	EMLOYEE RECRUITMENT PROCESS
72	578	EMPLOYEE SELECTION AND RECRUITMENT
73	581	IT HELP DESK
74	592	NOBEL NOMINATION PRECEDURE
75	595	VIP CUSTOMER SOFTWARE SUPPORT

76	603	STUDENT ASSIGNMENT
77	605	REWARD DELIVERY
78	612	MAP CREATION
79	791	TRIAL SERVICE AGREEMENT
80	794	AGILE SOFTWARE DEVELOPMENT
81	798	COURSE PREPARATION
82	843	ONLINE PURCHASING
83	878	HIRRING PROCESS
84	895	CINEMA TICKET PURCHASING APP
85	929	MEETING PREPARATION
86	931	SOFTWARE DEVELOPMENT
87	993	ARTWORK DESIGN

Table 27: Selected Business Processes from the SOA-based Database