A Machine Learning-based Modelling Assistant for Improving Understandability of Business Process Models*

Rosa Velasquez[1[0000−0001−6817−1517]]

PROS Research Center, Universitat Politècnica de València, Spain
rvelasquez@pros.upv.es

Abstract. Context. Understandability is one of the most important quality criteria in business process models (BPMs). While the experimental study of the factors that affect understandability is an ongoing research, current initiatives are focused on a limited set of factors. An open challenge is to explore the relationships among several of these factors using automated statistical techniques. Machine Learning (ML) has been applied to generate statistical models, based on the combination of multiple factors, and to find relationships to predict indicator's values. **Objective.** This thesis addresses the design of a method to assess the understandability of BPMs based on ML in order to predict whether a model could be understandable. This method will be implemented in an assisted modelling tool. **Method.** Using the design science methodology, the research aims to identify the factors that influence the understandability, their relationship, and how to measure them. This way we can correlate these factors and know which of them most affect the comprehensibility of the BPMs. Our final target is to provide an automatic evaluation of understandability. **Results.** The expected contributions are 1. the design of an understandability automatic evaluation model and 2. an assisted modelling tool that incorporates the evaluation model to provide real-time guidance for more understandable models. **Conclusion.** We aim to demonstrate that ML techniques can be used to predict BPMs understandability automatically.

Keywords: Business process model · Process model understandability · Model Assistant · Machine learning.

1 Introduction

Business process modelling (BPM) plays an important role in the early development of information systems. Business process models typically use graphical

* This research proposal is supervised by Oscar Pastor at Research Center on Software Production Methods, Universitat Politècnica de València, Valencia, Spain, and co-supervised by Ignacio Panach, at Escola Tècnica Superior d'Enginyeria, Universitat de València, Spain.

notation to capture the tasks, events, states, and control flow logic that constitute a business process [23]. As with any other information system (IS) conceptual model, process model's quality affects both the efficiency and effectiveness of IS development. As the modelling activity is relevant and complex [20], it is important to ensure models' quality. An important quality criterion is understandability, that can be defined as how easy is to understand a model by its stakeholders and any readers [13]. Although there exists several definitions and interpretations of the understandability concept, this research focus on the definition of Reijers et al.[23], who defines the understandability of BPMs as **"the degree to which information contained in a process model can be easily understood by a reader of that model"**. Model understandability has been widely studied in BPM literature. The definition of quality frameworks to characterize and measure understandability [12, 6], the design and validation of instruments for understandability evaluation [15, 21], and the empirical evaluation of the effects of understandability factors [9, 19], are examples of ongoing research areas. The study of understandability factors is a challenging area as there are many of them; only a subset of them has been empirically assessed [24, 30, 17, 18, 17]. Hence, conducting empirical studies on the factors that influence the understandability of BPMs is a relevant need. A factor is defined as a condition or characteristic that actively contributes to the quality of the software [16]. Then, for this research, **a factor is a characteristic that can affect -positively or negatively- the understandability of the BPMs.** Some experimental studies for exploring the relationships among understandability factors has been conducted previously [9, 19]. However, these initiatives are focused on a limited set of factors. An open challenge is to explore the relationships among several of these factors using statistical techniques. In our opinion, we can use Machine Learning (ML) techniques to get relevant information based on any combination of multiples variables to examine patterns and relationships that permit prediction or assessment of the software quality [3]. The lack of qualified modellers has been identified as a pitfall that might affect the understandability of the model [25]. The use of a tool to support the modelling activity may decrease the gap between inexperienced modellers and the skills required to make understandable models. Nowadays, User Assistance Systems (UAS) are tools that help the user in accomplishing tasks. The goal of these systems is to enhance user experience and help during its use [14]. The main contribution of this thesis is to combine a prediction model based on ML with an assistant to support the modelling activity. This way, BPMs should be more understandable by the audience. In order to achieve this goal, it is necessary to 1. training a ML-based understandability evaluation model, and 2. integrating the evaluation model with a BPM supporting tool in order to provide assistance for modellers. So, we have two challenges, one for designing a platform for collecting data to train the ML-based model, and other to methodologically integrate this platform into BMPs modelling tools. To address these challenges, the design science methodology [26] will be used.

The remaining part of this paper is organized as follows: Section 2 provides an overview of the related work. Section 3 presents the motivation, goals and research hypotheses. Section 4 describes the research methodology that is used in our thesis. Section 5 presents our contributions and principal challenges. Finally, Section 6 concludes this paper.

2 Background and Related Work

In this section, we present the background on understandability of conceptual models, and introduce the specific business process modelling approach that serves as context for this research: Communication Analysis [7]. We also present related works about experimental evaluations of understandability factors, as well as applications of ML to improve conceptual modelling in software engineering.

2.1 Background

Understandability of Conceptual Models Lindland et al.[13] define conceptual model quality as a composition of syntactic, semantic, and pragmatic quality, being the goal of the latter *that the model has been understood*. Then, we can say that the pragmatic quality is related to understandability. Understandability is defined as the degree to which information contained in a process model can be easily understood by a reader of that model [23].

Understandability of BPMs has been addressed in different quality frameworks. Krogstie et al. in [12] propose the SEQUAL BPM framework, containing seven core quality levels, being understandability one of them. A more specific quality framework for process models' understandability is presented by Dikici et al. [6]. This framework defines two factor categories that affect understandability: (1) Process model factors and (2) Personal factors. Process model factors consider the own model characteristics, for instance, the visual layout of the model, or its structural complexity, which can be measured by the number of arcs or the distance among entities. Personal factors consider the audience characteristics, such as the modelling expertise and domain familiarity of the model users. Apart from these understandability factors, the framework also defines understandability indicators. The indicators serve to operationalize or quantify model understandability. These indicators are divided into Objectively Measured Understandability and Perceived Understandability. Objectively Measured Understandability is related to the effectiveness and efficiency of the user to solve a task using the model information. Perceived understandability is related to the self-reported perception of ease of use, usefulness and intention to use. This research considers understandability factors and indicators from the framework of Dikici et al.

Business Process Modelling with Communication Analysis BPMs can be modelled using different methods and notations, being Business Process Modelling Notation (BPMN)[1] one of the most popular. BPMN approaches process modelling by representing a flow of activities and actors involved in them. A different approach is provided by Communication Analysis (CA) [7], which aims to process modelling with a communication-centric approach. Using communicative event diagrams, CA represents the flow of interactions among actors. The system requirements to support the communication can be further specified using the communicative event template, and the detailed information that the actors send to each other can be specified in terms of message structures. This research will be focused on the understandability of BPMs modelled using CA, and considers the adaptation of the framework to assess CA models.

2.2 Related Work

Understandability of BPMs is a research field that has drawn the interest of many researchers. They have proposed new factors that impact this quality criteria.

One of the factors that have empirically demonstrated impact over understandability is modularity [24]; however, when analyzing modularity along with other factors such as model size and the experience level of participants, modularity showed a (less) significant effect [30]. This reinforces the importance of assessing multiple understandability factors at the same time.

Mendling et al. [18] experimentally studied model user characteristics (MVC) and model characteristics as factors that influence understandability, finding significant evidence over understandability for both of them. Studies on MVCs [17] allow identifying subdimensions such as familiarity, education, and knowledge. Nevertheless, this factor has been studied in isolation and the conclusion could not work when including other factors.

In 2011, Reijers and Mendling conducted in [23] a study for discovering the effects of multiple factors in the understandability of BPMs. The factors are found through a revision of the literature. They classify these factors into model factors, personal factors, and other factors. They restrict the investigation scope only to the first two categories. The research supports the findings found previously [18]: the influence of personal factors is bigger than model factors. Another conclusion is that the combination of both categories can accomplish a greater explanatory power. However, as the authors comment, incorporating other factors that have not been addressed in the literature becomes a relevant challenge -for instance, the model purpose and the environment used to present the model-.

Most of the previous studies investigate factors that influence the understandability in BPMs focused on the analysis of a few factors; for instance, personal characteristics, or model characteristics, or model modularity, among others. We consider this procedure too partial. A holistic conceptual characterization is required to determine both the set of relevant factors and the set of their relationships that explain how to measure understandability in detail.

[1] https://www.businessprocessglossary.com/12878/business-process-model-and-notation-bpmn

We believe that by studying the relationships among them, we can provide a larger explanatory power to the problem under investigation. Additionally, the sample size of previous studies only considers a few participants with similar characteristics and, the set of process models for experiments are between 2 and 12. Much more data are needed to accomplish sound studies and yield accurate conclusions. This leads us to the ML connection that we discuss below.

Software Engineering and ML have been used in software engineering to solve problems related to software quality. Specifically, these techniques have been successfully applied to **predict software quality** [29, 10, 1]. These researches first identify variables that impact the quality; second, they study their correlation; and finally, they predict the quality of software based on the previous analysis.

Other initiatives have used ML techniques in the modelling activity to improve the quality of the models. These approaches can be classified into two directions: Repair model and Recommender System. In the repair model [2, 22], the approach is to identify and resolve issues of the model automatically. However, proposed tools assess the quality of already designed models and do not prevent potential pitfalls during the modelling activity. In recommender systems [8, 11], the approach is to recommend model fragments to add to the main model. These recommendations are based on an assistant function. Their results show that the supporting tool increases modelling speed and decreases the complexity of the modelling activity. Nevertheless, the cognitive load to integrate fragments of models can counterbalance this gain (re-use fragments of the model) because the modeller must understand and analyze these new fragments. On the other hand, these recommendations only assure the syntactic and semantic quality, but they do not guarantee that their integration affects understandability positively.

In summary, ML has been historically applied in software engineering to predict and assess tasks related to quality through integrates and search relationship among variables. Also has been applied in modelling activity to repairing models or recommending model improvements. However, these approaches do not face the problem under investigation. In this context, we propose an assisted modelling tool based on ML that improves the understandability of BPMs through the integration of multiple factors.

3 Research Definition

The **motivation** for this thesis is to help business process modellers to design more understandable models. To achieve this, we propose to research the factors that influence understandability of BPMs in an experimental, integrated manner, to better predict BPMs' understandability indicators, and transfer this knowledge into a modelling tool. The **main objective** of this thesis is to design an assisted modelling tool to guide the modeller in building understandable BPMs. The assistant will incorporate an automatic understandability evaluation model based on ML. The evaluation model will take as input the business process models audience's personal factors as well as model factors, and will predict the value of understandability indicators for the assessed business process model.

Based on the above, the specific research objectives (SRO) are:

- SRO 1: Analyze and select the factors that influence the understandability of BPMs together with the metrics and instruments used to measure them. This way we can establish an instrument to collect information to train the ML-based model.
- SRO 2: Design an automatic understandability evaluation model of the BPMs by applying ML techniques to correlate the understandability factors and indicators.
- SRO 3: Design and implement a method for an assisted modelling tool for BPM by integrating the automatic understandability evaluation model.
- SRO 4: Validate the assisted modelling tool through an empirical experiment, and generate evidence about model users' improvements on understandability and modellers satisfaction.

In order to assess the impact of both the automatic evaluation model of understandability and the assistant, we define three hypotheses:

- H_1: There is correlation among factors that influence the understandability of BPMs and the understandability indicators.
- H_2: The understandability of BPMs can be automatically evaluated using ML techniques.
- H_3: The BPMs produced with the assisted modelling tool that integrates an automatic evaluation ML model have better understandability than the BPMs produced without the assisted modelling tool.

4 Research Method

The research method is Design Science [26], which guides the design and investigation of artefacts in the context of stakeholders. In this research, the social context is the understandability of BPM. The artefact is a modelling assistant based on ML techniques. This research considers three phases of the method's engineering cycle: Problem investigation, Treatment design, and Treatment validation. The treatment implementation phase is out of the scope of this thesis. Each phase of this methodology is described in next subsections.

4.1 Problem Investigation

The focus for this phase is on describing the state of the art of the existing factors that affect model understanding and how these factors have been assessed in other researches. It also covers the identification of the measurement instruments or guidelines both for understandability factors and indicators. To achieve this goal, literature reviews on understandability factors, indicators, instruments and tools will be performed. This process will be conclude with a proposal of a consolidated instrument adapted to get information about the factors and understandability level of the model.

The expected result of this phase is a report about factors that impact understandability and indicators to measure understandability.

4.2 Treatment design

In this phase, there are two research objectives: The design of the automatic evaluation model of understandability (SRO 2) and the design of an assistant to support the modelling activity (SRO 3).

Regarding SRO 2, the Cross Industry Standard Process for Data Mining (CRISP-DM) method will be applied [27], given its detailed guidance for developing ML models. We are aware that identifying and integrating multiple understandability factors is a difficult task. Hence, the selection of the characteristics of the data set is a relevant activity that is supported by the CRISP-DM [27] methodology, specifically in the "Select Data (Rationale for Inclusion/Exclusion of characteristics)" phase. On the other hand, the data stored are not understandable for most ML techniques, because ML requires feature vectors as inputs. The feature vectors are the observable quantities that are the input for an ML [3]. Hence, we propose an ontological encoding to convert the data stored in feature vectors. Data will be collected from undergraduate students in business modelling activities. More than 1000 observations are expected to be collected during the research.

Concerning SRO 3, the approach is to design a initial version of an assistant to support modelling activity following the Situational Method Engineering [4] methodology. The goal is to design a method to integrate the understandability evaluation model into real-time, tool-supported modelling activities, with the purpose of guiding the modeller in building understandable BPMs, based on customizable factors, e.g. audience's background and purpose, among others.

The expected results of this phase is a method that integrates the factors that influence understandability and a tool to support the modelling activity. Whereas obtained data are limited due to the scope of the experiments, we will hope that this data is enough to generate a first version of the artefact, that will be prepared to continuously gather more data for further refinement of the predictive model.

4.3 Treatment validation

This phase presents the validation of the assistant to support the modelling activity (SRO 4). The main idea is to test the research hypotheses.

In response to H_1, the approach is to elaborate an experiment to validate the effect of the factors on understandability. We will present the findings detected by ML techniques to professional modellers and contrast if the detected relationships are aligned with their perspective.

To assess H_2, we will test different automatic evaluation models generated with different ML techniques, following the guidelines provides by CRISP-DM, in order to perform an experimental comparison among them in terms of accuracy and response time. The accuracy is the degree to which an automatic evaluation model predicts the understandability of a conceptual model that has already been manually rated.

Finally, in response to H_3, the approach is to provide empirical evidence of the quality, efficiency, and modellers' satisfaction with the assistant. First, an initial experiment will be conducted through the technology acceptance model (TAM) [5] to evaluate assistant acceptance. Second, regarding the understandability improvement on the BPMs, a second experiment will be designed according to the guidelines of Wohlin et al. [28]. This validation will focus on the comparison between the proposed assistant and the modelling activity without an assistant. The principal idea to include validation of H_3 is to validate a proof of concept where we will transferring what has been learned (with ML) to an artefact (assisted modelling tool). However, this hypothesis is independent of H_1 and H_2.

The expected results of this phase is a comparative study related to ML techniques to measure the understandability of BPMs and experiments conducted to test the modelling assistant.

5 Contributions and Challenges

This thesis incorporates ML to support the conceptual modelling activity, specifically to improve understandability. The contribution of this thesis is twofold. The first contribution provides an automatic evaluation model that predicts the level of understandability for the BPMs. This model will be based on the relationship detected with ML techniques. The second contribution provides empirical evidence about the integration of the automatic evaluation model of understandability into an initial version of an assisted modelling tool.

The main challenges of the proposed research are data collection and the appropriate representation of the understandability factors. With regard to data collection, many well-known ML algorithms depend on data to learn how assessing or predicting a problem. Therefore, the first challenge consists in adapting the software that uses Communication Analysis method[2] to get data about factors. This tool will be applied in ongoing business modelling courses with undergraduate students, with a minimum estimation of 1000 observations to be collected. Concerning representation of the factors, it is needed to adequately code the understandability factors from the Dikici et al. framework, specially the ones related to the model, in order to generate a feature vector as input for ML training. A possible solution could be encoding the model through ontologies.

6 Conclusions

BPMs' understandability is defined as the degree to which information contained in a process model can be easily understood by a reader of that model. The

[2] https://communicationalanalysismodeller.netlify.app/

understandability can be impacted or influenced -negatively or positively - by multiples factors, and hence, exploring the relationships among these factors is a relevant need. This research proposes helping in the generation of understandable process models through an assisted modelling tool, based on an ML-based understandability evaluation model. Our goal is to use ML techniques to identify patterns and correlations among the factors that impact BMPs' understandability indicators and predict the understandability level of BPMs. We will use the design science methodology to address this research, by first selecting and identifying the factors that influence the understandability. Second, we will capture the information about these factors in an experiment. Third, we will analyze the results applying ML techniques and extract the correlation among these factors. Fourth, we will develop an evaluation model to predict the level of understandability. Fifth, we will integrate the evaluation model into an assisted modelling tool. Finally, we will validate the evaluation model and the modelling assistant through practical case studies.

To date, there are no presentations or dissemination of any results about this research yet, since the author has recently begun this research.

References

1. Al-Jamimi, H.A., Ahmed, M.: Machine learning-based software quality prediction models: State of the art. In: 2013 International Conference on Information Science and Applications (ICISA). pp. 1–4 (2013)
2. Barriga, A., Rutle, A., Heldal, R.: Personalized and automatic model repairing using reinforcement learning. In: 2019 ACM/IEEE 22nd International Conference on Model Driven Engineering Languages and Systems Companion (MODELS-C). pp. 175–181 (2019)
3. Baştanlar, Y., Özuysal, M.: Introduction to Machine Learning. miRNomics: MicroRNA Biology and Computational Analysis pp. 105–128 (2014)
4. Bucher, T., Klesse, M., Kurpjuweit, S., Winter, R.: Situational Method Engineering. In: Situational Method Engineering: Fundamentals and Experiences. pp. 33–48. Springer, Boston, MA (2007)
5. Davis, F.D.: Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly pp. 319–340 (1989)
6. Dikici, A., Turetken, O., Demirors, O.: Factors influencing the understandability of process models: A systematic literature review. Information and Software Technology pp. 112–129 (2018)
7. España, S., González, A., Pastor, Ó.: Communication analysis: a requirements engineering method for information systems. In: International Conference on Advanced Information Systems Engineering. pp. 530–545. Springer (2009)
8. Fellmann, M., Zarvić, N., Thomas, O.: Business Processes Modelling Assistance by Recommender Functionalities: A First Evaluation from Potential Users. In: Perspectives in Business Informatics Research, pp. 79–92. Springer (2017)
9. Gabryelczyk, R., Jurczuk, A.: Does Experience Matter? Factors Affecting the Understandability of the Business Process Modelling Notation. Procedia Engineering **182** (2017)
10. Ganesan, K., Khoshgoftaar, T., Allen, E.: Case-based software quality prediction. International Journal of Software Engineering and Knowledge Engineering **10**, 139–152 (2000)

11. Koschmider, A., Oberweis, A.: Designing Business Processes with a Recommendation-Based Editor. Handbook on Business Process Management 1 pp. 299–312 (2010)
12. Krogstie, J.: SEQUAL Specialized for Business Process Models. Quality in Business Process Modeling pp. 103–138 (2016)
13. Lindland, O.I., Sindre, G., Solvberg, A.: Understanding quality in conceptual modeling. IEEE Software pp. 42–49 (1994)
14. Maedche, A., Morana, S., Schacht, S., Werth, D., Krumeich, J.: Advanced User Assistance Systems. Business & Information Systems Engineering **58** (2016)
15. Maes, A., Poels, G.: Evaluating quality of conceptual modelling scripts based on user perceptions. Data & Knowledge Engineering pp. 701–724 (2007)
16. McCall, J.A., Richards, P.K., Walters, G.F.: Factors in Software Quality. Volume I. Concepts and Definitions of Software Quality. Tech. rep., General Electric co Sunnyvale ca (1977)
17. Mendling, J., Recker, J., Reijers, H.A., Leopold, H.: An Empirical Review of the Connection Between Model Viewer Characteristics and the Comprehension of Conceptual Process Models. Information Systems Frontiers **21**, 1111–1135 (2019)
18. Mendling, J., Reijers, H.A., Cardoso, J.: What Makes Process Models Understandable? In: Business Process Management, pp. 48–63. Springer Berlin Heidelberg (2007)
19. Mendling, J., Strembeck, M., Recker, J.: Factors of process model comprehension—Findings from a series of experiments. Decision Support Systems pp. 195–206 (2012)
20. Moody, D.L.: Theoretical and practical issues in evaluating the quality of conceptual models: current state and future directions. Data & Knowledge Engineering (2005)
21. Moody, D.L., Sindre, G., Brasethvik, T., Sølvberg, A.: Evaluating the Quality of Process Models: Empirical Testing of a Quality Framework. In: Conceptual Modeling — ER 2002 (2003)
22. Pinna Puissant, J., Van Der Straeten, R., Mens, T.: Resolving model inconsistencies using automated regression planning. Software & Systems Modeling pp. 461–481 (2015)
23. Reijers, H.A., Mendling, J.: A Study Into the Factors That Influence the Understandability of Business Process Models. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans pp. 449–462 (2011)
24. Reijers, H., Mendling, J.: Modularity in Process Models: Review and Effects. In: Business Process Management. pp. 20–35. Springer, Berlin, Heidelberg (2008)
25. Rosemann, M.: Potential pitfalls of process modeling: Part A. Business Process Management Journal pp. 377–384 (2006)
26. Wieringa, R.J.: Design Science Methodology for Information Systems and Software Engineering. Springer-Verlag (2014)
27. Wirth, R., Hipp, J.: CRISP-DM: Towards a Standard Process Model for Data Mining. In: Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining (2000)
28. Wohlin, C., Runeson, P., Höst, M., Ohlsson, M.C., Regnell, B., Wesslén, A.: Experimentation in Software Engineering. Springer Berlin Heidelberg (2012)
29. Zhang, D., Tsai, J.J.: Machine Learning and Software Engineering. Software Quality Journal pp. 87–119 (2003)
30. Zugal, S., Pinggera, J., Weber, B., Mendling, J., Reijers, H.A.: Assessing the Impact of Hierarchy on Model Understandability – A Cognitive Perspective. In: Models in Software Engineering, pp. 123–133. Springer Berlin Heidelberg (2012)