Error Metrics for Business Process Models

Jan Mendling[1] and Gustaf Neumann[1]

Vienna University of Economics and Business Administration
Augasse 2-6, 1090 Vienna, Austria
{jan.mendling|neumann}@wu-wien.ac.at

Abstract. Little research has been conducted so far on causes for errors in business process models. In this paper we investigate on how mainly domain independent factors such as the size or complexity of models influence errors observed in a wide range of existing business process models. In particular, we provide a set of six metrics presumably related to the comprehensibility of both the process model structure and the process state space, and discuss their capability to predict errors in the SAP reference model. The results show that already the three metrics *size*, *separability*, and *structuredness* suffice to achieve a high Nagelkerke R^2 value of 0.853 demonstrating a good predictive efficacy.

1 Introduction

Up to now there has been little research on why people introduce errors in business process models in practice. In a more general context, Simon [1] points to the limited cognitive capabilities and concludes that humans act rationally only to a limited extent. Related to modeling errors, this argument would imply that human modelers loose track of the interrelations of large and complex models due to their limited cognitive capabilities, and then introduce errors that they would not insert in a small model. A recent study provides first evidence for this hypothesis [2]. Before we can test such a hypothesis appropriately, we have to establish an understanding of which determinants drive error probability of process models and how we can measure them (cf. e.g. [3]).

In this context, the contribution of this paper is a newly developed set of metrics for measuring the error probability of business process models. Beyond the theoretical foundation of these metrics, we provide a first validation based on the EPCs of the SAP reference model [4]. By defining quality concepts like error probability in a measurable way, we contribute to the understanding of bad process design in general. Against this background, the remainder of the paper is structured as follows. In Section 2 we identify comprehensibility of a process model's structure and its state space as the key determinants for error probability. For each of both we define a set of metrics and discuss their theoretical impact on error probability. Section 3 provides a first evaluation of this set of metrics for predicting errors in the SAP reference model. Finally, Section 4 gives a summary and an outlook on future research. For related work, refer to [5].

2 Error Determinants for Business Process Models

Following the principles of measurement theory (see e.g. [3]), we consider the comprehensibility of the business process model as the main determinant for error probability. This is motivated by the assumption that the process models are constructed by human modelers and that their design is subject to bounded rationality [1]. The comprehensibility of any model by a person is determined by her familiarity with the real-world process and by the way the model elements are combined to represents real-world process. In this paper we only investigate into the second aspect. More precisely, we analyze (a) the *process model structure* and (b) the *process model state space*. For both these determinants, we identify a set of sub-determinants and discuss their impact on error probability. In the following, we consider a business process model to be a special kind of graph $G = (N, A)$ with at least three node types $N = T \cup S \cup J$, i.e., *tasks* T, *splits* S, *joins* J, and *control flow arcs* $A \subseteq N \times N$ to connect them. We use the generic term connectors $C = S \cup J$ for splits and joins collectively. Each connector has a *label* AND, OR, or XOR that gives its routing of merging semantics.

Size: Several papers point to size as an important factor for the comprehensibility of software and process models [5, 3]. While the size of software is frequently equated with lines of code, the size of a process model is often related to the number of nodes N of the process model. The metric S_N measures the number of nodes of the process model graph G. An increase in $S_N(G)$ should imply an increase in error probability of the overall model.

$$S_N(G) = |N|$$

Separability: Separability is closely related to the notion of a cut-vertex (or articulation point), i.e., a node whose deletion separates the process model into multiple components. We define the separability ratio Π as the number of cut-vertices to number of nodes. Cut-vertices can be found using depth-first search. An increase in $\Pi(G)$ should imply a decrease in error probability of the model.

$$\Pi(G) = \frac{|\{n \in N \mid n \text{ is cut-vertex}\}|}{|N| - 2}$$

Sequentiality: Sequentiality relates to the fact that sequences of consecutive tasks are the most simple building blocks of a process model. The sequentiality ratio Ξ relates arcs of a sequence to the total number of arcs. An increase in $\Xi(G)$ should imply a decrease in error probability of the overall model.

$$\Xi(G) = \frac{|\{a \in A \mid a \in (T \times T)\}|}{|A|}$$

Structuredness: Structuredness relates to how far a process model can be built by nesting blocks of matching join and split connectors (see e.g. [6]). The degree of structuredness can be determined by applying reduction rules and comparing

the size of the reduced model to the original size. The structuredness ratio Φ of the process graph is one minus the number of nodes in the reduced process graph divided by the number of nodes in the original process graph. An increase in $\Phi(G)$ should imply a decrease in error probability of the overall model.

$$\Phi_N = 1 - \frac{S_N(G')}{S_N(G)}$$

Cyclicity: Cyclic parts of a model are more difficult to understand than sequential parts. $|N_C|$ gives the number of nodes on some cycle and cyclicity CYC relates it to the total number of nodes. An increase in $CYC(G)$ should imply an increase in error probability of the overall model.

$$CYC_N = \frac{|N_C|}{|N|}$$

Parallelism: Modelers have to keep track of concurrent paths that need to be synchronized. AND- and OR-splits introduce new threads of control such that the number of control tokens potentially increases by the number of the output degree minus one. The Token Split TS metric counts these newly introduced tokens. An increase in $TS(G)$ should imply an increase in error probability of the overall model.

$$TS(G) = \sum_{c \in C_{or} \cup C_{and}} d_{out}(n) - 1$$

3 Evaluation of the Metrics

In this section, we utilize the SAP reference model in order to evaluate the metrics proposed in the previous section for their capability to predict errors. A relaxed soundness analysis (see [7, 2]) of this model collection revealed that 34 of the about 600 EPC business process models (see [8]) have errors. Since the dependent variable is binary (error yes/no), we use a logistic regression (logit) model (see e.g. [9]) where a positive coefficient increases and negative one decreases error probability.

Depending on the metrics defined before, we calculated multivariate logit models to predict error probability. The model with three variables was the largest that satisfied various goodness tests, i.e., the Hosmer&Lemeshow test is 0.216 which is significantly greater that 0.05 and all Wald statistics indicate that the coefficients are significantly different from zero. The Nagelkerke R^2 as a coefficient of determination describes which fraction of the variability is explained. With an R^2 of 0.853 we achieve a high rate of explanation by the metrics. Furthermore, it is interesting to note that the three *variables of the model confirm the tendency* that was postulated in the hypothesis, i.e. error probability increases with size and decreases with higher separability or structuredness.

4 Contribution and Limitations

In this paper we explored in how far errors in a business process model can be predicted by the help of suitable metrics. Based on the general hypothesis that error probability is determined by the comprehensibility of the process model structure and the process state space, we identify a set of six metrics for predicting error probability. Each of these metrics is discussed in terms of its motivation, calculation, and its theoretical impact on error probability. We use the sample of the SAP reference model to evaluate the metrics in a multivariate logit model. The results are statistically significant and show that already three variables (*size*, *separability*, and *structuredness*) suffice to achieve a high Nagelkerke R^2 value of 0.853. Furthermore, the *hypothetical direction of their impact on error probability is confirmed*. This is a considerable improvement compared to the previous analysis where count metrics yielded a Nagelkerke R^2 value of less than 0.35 [2]. Still, our evaluation has some limitations that we aim to address in future research. First, relaxed soundness is not able to find all problematic parts of a process model. We are currently working on a soundness notion for EPCs and a respective verification approach in order to get more precise information about errors in an EPC. Second, there is no research available that discusses the representativeness of the SAP reference model as a process model collection. Therefore, future research will have to evaluate the metrics against other samples in order to establish them as error predictors.

References

1. Simon, H.A.: Sciences of the Artificial. 3rd edn. The MIT Press (1996)
2. Mendling, J., Moser, M., Neumann, G., Verbeek, H., Dongen, B., Aalst, W.: A Quantitative Analysis of Faulty EPCs in the SAP Reference Model. BPM Center Report BPM-06-08, BPMCenter.org (2006)
3. Fenton, N.E., Pfleeger, S.L.: Software Metrics. A Rigorous and Practical Approach. PWS, Boston (1997)
4. Keller, G., Teufel, T.: SAP(R) R/3 Process Oriented Implementation: Iterative Process Prototyping. Addison-Wesley (1998)
5. Mendling, J., Neumann, G.: Error metrics for business process models. Technical Report JM-2006-12-03, Vienna Univ. of Econ. and Business Administration (2006)
6. Kiepuszewski, B., Hofstede, A., Bussler, C.: On structured workflow modelling. In Wangler, B., Bergman, L., eds.: Advanced Information Systems Engineering, 12th International Conference CAiSE 2000. Volume 1789 of Lecture Notes in Computer Science., Springer (2000) 431–445
7. Dehnert, J., Rittgen, P.: Relaxed Soundness of Business Processes. In Dittrick, K.R., Geppert, A., Norrie, M.C., eds.: Proceedings of the 13th International Conference on Advanced Information Systems Engineering. Volume 2068 of Lecture Notes in Computer Science., Interlaken, Springer (2001) 151–170
8. Keller, G., Nüttgens, M., Scheer, A.W.: Semantische Prozessmodellierung auf der Grundlage "Ereignisgesteuerter Prozessketten (EPK)". Heft 89, Institut für Wirtschaftsinformatik, Saarbrücken, Germany (1992)
9. Hair, jr., J.F., Anderson, R.E., Tatham, R.L., Black, W.C.: Multivariate Data Analysis. 5th edition edn. Prentice-Hall International, Inc. (1998)