Obtaining Thresholds for the Effectiveness of Business Process Mining

[1] Ricardo Pérez-Castillo, [1] Laura Sánchez-González, [2] Mario Piattini, [1] Félix García and
[1] Ignacio García-Rodriguez de Guzmán
[1] University of Castilla-La Mancha
Paseo de la Universidad 4 13071, Ciudad Real, Spain
{ricardo.pdelcastillo, laura.sanchez, felix.garcia, ignacio.grodriguez}@uclm.es
[2] Alarcos Quality Center
Camino de Moledores s/n, 13051 Ciudad Real, Spain
mario.piattini@alarcosqualitycenter.com

Abstract
Business process mining is a powerful tool to retrieve the valuable business knowledge embedded in existing information systems. The effectiveness of this kind of proposal is usually evaluated using recall and precision, which respectively measure the completeness and exactness of the retrieved business processes. Since the effectiveness assessment of business process mining is a difficult and error-prone activity, the main hypothesis of this work studies the possibility of obtaining thresholds to determine when recall and precision values are appropriate. The business process mining technique under study is MARBLE, a model-driven framework to retrieve business processes from existing information systems. The Bender method was applied to obtain the thresholds of the recall and precision measures. The experimental data used as input were obtained from a set of 44 business processes retrieved with MARBLE through a family of case studies carried out over the last two years. The study provides thresholds for recall and precision measures, which facilitates the interpretation of their values by means of five linguistic labels that range from low to very high. As a result, recall must be high (with at least a medium precision above 0.56), and precision must also be high (with at least a low recall of 0.70) to ensure that business processes were recovered (by using MARBLE) with an effectiveness value above 0.65. The thresholds allowed us to ascertain with more confidence whether MARBLE can effectively mine business processes from existing information systems. In addition, the provided results can be used as reference values to compare MARBLE with other similar business process mining techniques.

Keywords: **Business Process, Process Mining, Thresholds, Bender Method, Case Study.**

I. INTRODUCTION

Business processes are an essential asset for organizations. They provide a means for mapping business objectives regarding how to carry out better operations, generating added value for customers [43]. Thereby, business processes improve the competitiveness level of the organizations that consider and manage them [8].

At this time, most business processes are automated by enterprise information systems [18]. Thus, business processes must sometimes be recovered from existing information systems for two main reasons. First, some organizations do not explicitly manage their business processes, and their existing information systems embed valuable business knowledge that is not present anywhere else [19]. Second, existing information systems evolve faster than business processes owing to uncontrolled software maintenance overtime, and obsolete and misaligned business processes raise an important concern for organizations [16].

The solution to retrieve business processes from existing information systems is known as 'business process mining' [41], and it is used by organizations to eradicate, or at least mitigate, these problems.

Many business process mining approaches and techniques exist in the literature [1]. Some business process mining proposals have been empirically validated in terms of their effectiveness, i.e., the degree of quality of the obtained business processes. So far, several measures have been used to measure the effectiveness of process mining techniques by comparing the discovered business processes with respect to the reference processes (i.e., business processes constructed in the design stage). Of the existing effectiveness measures, the most notable are *recall* and *precision*, two measures adapted from the research area of information retrieval, which have been used in a great number of empirical studies and experiments. Intuitively, precision indicates how exact a mined business process is, while recall indicates how complete a process is.

Despite the fact that recall and precision have been used in several empirical studies, there is no reference threshold to indicate which precision and recall values are appropriate for business process mining scenarios. For instance, some questions can arise such as *"How good is 0.67 precision?"* or *"Is that value an acceptable result?"* and these currently cannot be answered. This problem implies that the precision and recall values obtained in empirical studies cannot therefore be conveniently interpreted and this limits the usefulness of the information provided to assess the effectiveness of a particular business process mining technique.

This paper examines this challenge and obtains benchmark thresholds to interpret the precision and recall measure values. The thresholds were obtained by applying the Bender method [3] to the data extracted from a family of case studies in which MARBLE (*Modernization Approach for Recovering Business processes from LEgacy systems*), a business process mining technique proposed in previous works [29, 32], was validated. The family of case studies was carried out over the last two years and is composed of six industrial case studies involving six real-life information systems. As a result, the precision and recall measures were turned into indicators, by associating *decision criteria* to them [14]. In order to obtain the thresholds, the Bender method was applied to the data about MARBLE effectiveness extracted from the case studies.

The remainder of this paper is organized as follows. Section II provides the background about precision and recall measures. Section III presents related work. Section IV explains in detail the data obtained from the family of case studies which were used to obtain thresholds. Section V describes how the *Bender* method was applied to obtain the thresholds and the obtained results are presented. These thresholds are applied to interpret the data from the case studies, as illustrated in Section VI. Finally, Section VII discusses the main conclusions of this study as well as future work.

II. Information Retrieval Measures

Recall and precision are two measures raised from the information retrieval research field, and they have been used for many years [42]. In addition, both recall and precision measures have usually been used to evaluate the effectiveness of business process mining techniques or algorithms as well as other kinds of reverse engineering methods [37]. As a result, these measures not only evaluate the obtained business processes, but they evaluate the effectiveness of the method to recover business processes.

The evaluation of recall and precision consists of measuring the difference between the obtained and desired results. These measures provide a number between 0 and 1 without unity. Recall and precision measure the similarity between a mined business process *M* and a reference business process *R*. Recall indicates what proportion of *R* is present in *M* (i.e., how complete *M* is), while precision indicates what proportion of *M* matches *R* (i.e., how exact *M* is). On the one hand, recall is the number of relevant elements mined over the total number of existing relevant elements in a business process. On the other hand, precision is scored as the number of relevant elements mined over the total number of elements of a certain mined business process.

A particular element is considered *relevant* if that element faithfully represents a piece of business operation or business behavior of the organization in the real world. To evaluate these measures, various business process elements (e.g. activities, sequence flows, data objects, etc.) were used as the unit to analyse the relevance. Despite this variety, the business activity (or business task) is the most commonly used element.

As a consequence, the recall measure (1) is defined as the number of retrieved relevant tasks divided by the total number of relevant tasks (i.e., divided by the retrieved and non-retrieved relevant tasks). Moreover, a precision measure (2) is defined as the number of retrieved relevant tasks divided by the total number of retrieved (relevant and non-relevant) tasks.

$$RECALL = \frac{\{retrieved\ relevant\ tasks\}}{\{retrieved\ relevant\ tasks\} + \{non\ retrieved\ relevant\ tasks\}} \quad (1)$$

$$PRECISION = \frac{\{retrieved\ relevant\ tasks\}}{\{retrieved\ relevant\ tasks\} + \{retrieved\ non\ relevant\ tasks\}} \quad (2)$$

$$Fmeasure = \frac{(1+\alpha)\cdot Recall \cdot Precision}{\alpha \cdot \mathrm{Precision} + \mathrm{Recall}} \quad (3)$$

Although recall and precision are appropriate, there is an inverse relationship between both measures. As a result, these measures are rarely used in an isolated way, and recall and precision are therefore usually combined into a single measure known as the F-measure. The F-measure (3) consists of a weighted harmonic mean of recall and precision. The F-measure makes it possible to modify the importance of precision and recall by means of the α value. The α value is usually 1, which means that precision and recall have the same weight in the F-measure formula. This measure is interesting, because it gives a low result for business process mining techniques, which improves precision or recall exclusively to the detriment of the other [25].

All business process mining proposals usually try to maximize the F-measure by maximising both recall and precision measures, which is the desirable result. Indeed, the precision value should, ideally, always be 1 for any recall value but, according to [12], this is not possible in practice due to their inverse relationship (see Figure 1). Because of this, business process mining techniques obtain quite different results, for instance a high precision and a low recall or vice versa. For this reason, reference thresholds are needed to know when the effectiveness of a particular business process mining technique is actually good.

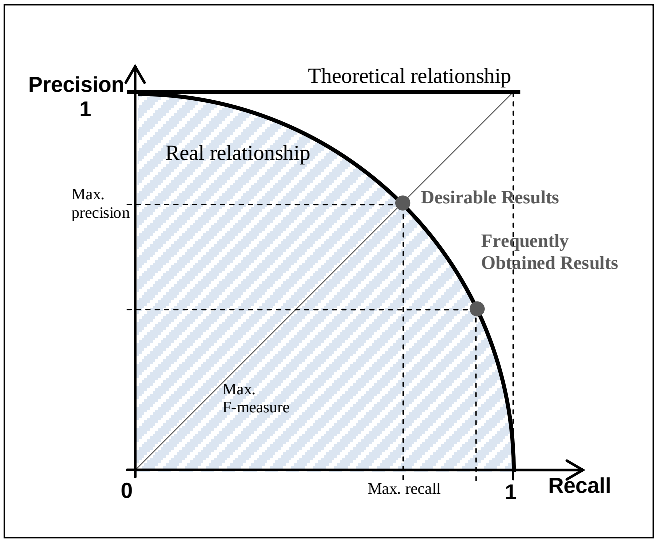

Figure 1. Relationship between precision and recall measures

III. Related Work

This section first presents some existing techniques for business process mining which use recall and precision measures to evaluate their effectiveness. Then, the section summarizes some threshold definition proposals.

A. *Business Process Miners Precision and Recall*

There are many proposals for business process miners in the literature which use precision and recall measures to evaluate their effectiveness. *Lo and Khoo* [36] propose a QUARK (QUality Assurance framewoRK), which enables assessments of the effectiveness of specification miners based on finite state machines. This framework uses the k-tails algorithm [5], which considers all the sub-sequences of k length from each node to score precision and recall measures. This framework has reported an average precision of 0.10 and an average recall of 0.90. The disadvantage of this framework is that it uses the finite state machine format to represent the mined processes, instead of Business Process Modeling and Notation (BPMN) [28] or another standardized format.

Pérez-Castillo et al. [33] present a case study to compare the effectiveness of two related business process mining techniques. The first technique statically analyses the source code, and the second one analyses traces obtained during system execution. The study reports that the second technique has higher precision and recall values, which are respectively 0.70 and 0.77.

Lou et al. [20] propose a business process mining algorithm using event traces recorded during system execution. This work provides a set of experiments that reports a precision value around 0.50 and an average recall of 0.99. Following a similar approach, *Povzner et al.* [35] propose Autograph, a framework to extract data file signatures in workflows. This approach examines traces of file accesses, finds repeated and correlated accesses, and infers which files most likely belong to the same workflow. *Povzner* and colleagues carried out an experiment using benchmark programs and reached precision and recall values of 0.65 and 0.85 respectively.

Some approaches focus on recovering patterns concerning the source code architecture. For instance, *Asencio et al.* [2] provide a program analysis tool to automatically recognise the use of software design patterns found in object-oriented code. This framework is applied to an ACE system (Adaptive Communications Environment). The precision and recall values obtained in this evaluation are 0.68 and 0.71 respectively. Another study that follows the reverse engineering of design patterns from source code is proposed by *Philippow et al.* [34]. They provide a survey of pattern search methods, which have been applied to several systems. This work reports precision and recall values of 0.37 and 1.0 in some cases. *Yeh et al.* [45]carried out an empirical study of a miner, focusing on the recovery of object relationships from source code. This study reports an average precision of 0.74 and a recall of 0.81.

Other works address the search in business process model repositories, instead of recovery business process models from information systems. For instance, *Lucrédio et al.* [21] provide MOOGLE, a model search engine to recover several kinds of models (not only business process models) from large model repositories. This work presents the result of an experiment using MOOGLE, which provides precision and recall values of 0.24 and 0.53. In addition, *Lucrédio et al.* [21] take into account similar empirical studies such as [13, 15, 44] to fix a benchmark value of 0.50 for both recall and precision measures.

Despite the aforementioned proposals can consider different source datasets to asses precision and recall, TABLE I compares the precision and recall obtained in each work. Recall is higher than precision in all of the proposals. This trend is due to the fact that any reverse engineering method, such as business process mining, makes it possible to recover a significant amount of information. However, this does not mean that the appropriate information is automatically recovered. Indeed, some information may not be available to be retrieved by means of the business process mining technique. For this reason, obtaining higher precision is a key challenge.

TABLE I. Precision and recall obtained in the literature

	Precision	**Recall**
Lo and Khoo [36]	0.10	0.90
Pérez-Castillo et al.[33]	0.70	0.77
Lou et al.[20]	0.50	0.99
Povzner et al.[35]	0.65	0.85
Asencio et al.[2]	0.68	0.71
Philippow et al.[34]	0.37	1.00
Yeh et al.[45]	0.74	0.81
Lucrédio et al.[21]	0.24	0.53

These proposals focus on business process mining techniques or algorithms that use precision and recall measures to evaluate their effectiveness. However, a wide variety of measures are used to measure different features of business processes [39]. For instance, the control-flow complexity measure/(CFC) [7] is used to calculate the degree of complexity of process models from a decision-node perspective. The main idea is that the high complexity of process models produces difficulties in terms of understanding the model, because of an increase in errors, deadlocks, bottlenecks, etc. The use of a CFC measure allows designers to improve process models, reducing the time needed to read and understand processes in order to adapt them to new requirements. Other interesting measures include error probability measures [23]. This group of measures calculates the likelihood of finding errors in business process models, specifically 13 measures related to process model elements: number of nodes, diameter, density, depth, sequentiality, separability, cyclicity, average and max degree of connectors, mismatch connector, connector heterogeneity, connectivity coefficient and concurrency. However, this kind of measure focuses on mined business processes, but does not allow the effectiveness of the mining method to be measured.

B. *Threshold Definition Proposals*

It is possible to find several proposals about the definition of thresholds in the literature, especially in the medical field. Unfortunately, for calculation of threshold values no standard method is available. Some authors base threshold values on experience, which is the case with *McCabe* [22], *Nejmeh* [26] and *Coleman et al.* [10]. Others adapt some medical methods for software engineering, for example *Shatnawi* [40], who uses the Bender method [3] in order to obtain thresholds for object-oriented measures. In that work, the author defines thresholds to study the relationship between some object-oriented measures (e.g., Coupling Between Objects (CBO), Response For Class (RFC), Weighted Methods Complexity (WMC), Depth of Inheritance Hierarchy (DIT)) as well as error-severity categories, concluding by demonstrating the utility of the Bender method for this purpose. On the other hand, *Berlarbi et al.* [4] extract thresholds in order to predict which classes are likely to contain faults and test threshold effects in subsets of the *Chidamber and Kemerer* measures [11]. However, these results are only valid for the measures used by the authors, since other models may potentially produce different results.

IV. DATA COLLECTION

This section presents the data we used in this study to obtain the benchmark thresholds by applying the Bender method. The case studies were carried out over the last two years with six different information systems to validate the effectiveness of MARBLE. To improve the rigor and validity of the case studies, the guidelines proposed by *Runeson* and *Höst* [38] and *Brereton el al.* [6] were followed. The following sub-sections summarise how the stages of the protocol were followed.

A. *The Object of Study*

The *object of study* is MARBLE [29], which is a model-driven framework to recover business process models from existing information systems. MARBLE considers different models at different abstraction levels and a set of model transformations between them. A supporting tool is additionally provided in order to automate the technique and facilitate its adoption.

Specifically, MARBLE defines four kinds of models at four different abstraction levels. *L0* is the lowest level of abstraction since it represents the existing information system in the real world as a set of different software artifacts (e.g. source code, database, documentation, etc). *L1* contains different platform-specific models (PSM) depicting the different software artifacts of the LIS. *L2* integrates all the specific *L1* models into a platform-independent model (PIM), which is represented according to the Knowledge Discovery Metamodel (KDM) [17]. KDM enables the representation and management of the knowledge extracted by means of reverse engineering from all the different software artifacts of legacy information systems. Finally, *L3* depicts the discovered business processes, which are represented according to the BPMN metamodel [28].

Moreover, MARBLE defines three model transformations between the four levels (see Figure 2). First, the L0-to-L1 transformation obtains PSM models from each legacy software artifact using a specific metamodel for each artifact. Traditional reverse engineering techniques such as static analysis or dynamic analysis are used to extract the needed knowledge. Second, the L1-to-L2 transformation (see Figure 2) consists of a set of model transformations (e.g. implemented using the QVT (Query/View/Transformation) language [27]) to obtain a KDM model built from the PSM models at L1. Finally, the L2-to-L3 transformation obtains the current business process model. This transformation is based on a set of business patterns [30], which define the transformation rules between levels L2 and L3. In addition, this last transformation can be implemented through QVT rules [31].

MARBLE assesses precision, recall and F-measure by considering task as relevant when they are retrieved in the same order, with the same related sequence flows and data objects [29]. This evaluation is carried out by business experts once business processes were retrieved.

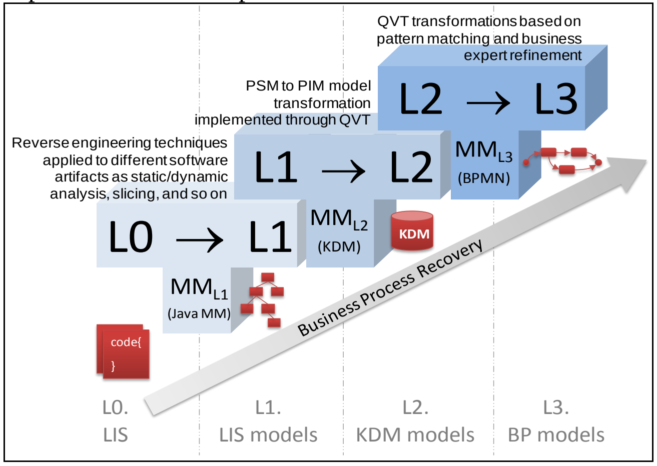
QVT transformations based on pattern matching and business expert refinement
PSM to PIM model transformation implemented through QVT
Reverse engineering techniques applied to different software artifacts as static/dynamic analysis, slicing, and so on
L2 → L3
L1 → L2
L0 → L1
MM_{L3} (BPMN)
MM_{L2} (KDM)
MM_{L1} (Java MM)
KDM
code{ }
Business Process Recovery
L0. LIS
L1. LIS models
L2. KDM models
L3. BP models

Figure 2. MARBLE Overview

B. *Design*

While the *object of study* is MARBLE, the *purpose of this study* is the evaluation of its effectiveness. For this purpose, the family of case studies defines the following research question: *"Can MARBLE effectively recover business processes from existing information systems?"*

Each case study in the family follows the *embedded case study* design. This means that each case study consists of a single case, (i.e. it focuses on a single information system), but in addition, the study considers several analysis units within the case. The analysis units are the different business processes retrieved from each information system, which represent the independent variable in each study.

Each study consists of the analysis of all the retrieved business processes in order to answer the research question. Therefore, the family of six case studies uses recall, precision and F-measure (c.f. Section II) to provide quantitative answers to the proposed research question.

C. *Case Selection*

Cases under study cannot be randomly selected. Thus, the case selection protocol defines a list of four criteria to choose suitable cases to be studied. The criterion *C1* guarantees that the existing information system selected is an enterprise system that supports the business operation of an organisation. *C2* ensures that the selected system is truly a legacy information system. This criterion considers the number of system modifications that altered the business processes. *C3* ensures that the system is not a *toy program*, since it defines a threshold of 20,000 lines of source code. Finally, *C4* guarantees that the system is based on the Java platform, since the supporting tool was developed for Java-based systems. After evaluating, according to these criteria, six information systems were selected from eleven available systems, given by partner companies. TABLE II shows the name of each system, a brief description, the type of architecture, and the size in thousands of lines of source code.

TABLE II. LEGACY INFORMATION SYSTEMS UNDER STUDY

Id	Name	Description	Architecture	Size (KLOC)
S1	AELG-Members	Supports the administration of an organization of authors	Desktop application	23.5
S2	Tap CRM	A sales force automation tool for sales management	Web application	49.6
S3	VillasanteLab	Manages a laboratory in the water and waste industry	Web application	28.8
S4	XuntaEadmin	Supports the electronic administration of a Spanish regional ministry	E-government system	320.2
S5	SIXA	A learning management system	Web application	140.6
S6	CHES	A computer-based health evaluation system for the oncology area	Desktop application	619.8

D. *Collected Data*

TABLE III provides the data collected during the execution of the case studies. TABLE III shows: (i) the business process identifier; (ii) the case study identifier; (iii) the retrieved business process name; (iv) the precision (v) recall, and (vi) F-measure values for each retrieved business process.

These data constitute the input necessary for the application of the Bender method in order to obtain the goodness thresholds for both recall and precision measures (see Section V).

E. *Analysis and Interpretation*

The collected data were analysed to obtain evidence chains from data to answer the research question (c.f. sub-section B). In order to answer the proposed question, Figure 3 shows the box chart for recall and precision measures. The means of the distributions of the precision measure (0.51, 0.53, 0.66, 0.64, 0.54 and 0.63) are always lower than the recall means (0.77, 0.83, 0.91, 0.65, 0.66 and 0.64). Higher recall values mean that the proposed technique retrieves very complete processes (i.e., it retrieves most of the tasks from the real-world business processes). Nevertheless, the lower precision value means that the business process mining technique is imprecise (i.e., the number of retrieved non-relevant tasks is very high) (see Figure 3). These recall and precision values respectively provide F-measure means for each case study of 0.61, 0.64, 0.76, 0.63, 0.59 and 0.63.

TABLE III. RESULTS OBTAINED FOR THE FAMILY OF CASE STUDIES

Subject ID	Study ID	Business Process Model Name	Recall	Precision	F-Measure
1	1	Categories Management	0.769	0.580	0.661
2	1	Author Management	0.888	0.504	0.643
3	1	Reporting	0.667	0.444	0.533
4	2	Security Management	0.882	0.435	0.583
5	2	Administration	0.788	0.578	0.667
6	2	Chemical Analysis Management	0.911	0.638	0.750
7	2	Chemical Calibration Management	0.824	0.718	0.767
8	2	User Management	0.842	0.250	0.386
9	2	Chemical Dilution Management	0.667	0.410	0.508
10	2	Reporting	0.864	0.717	0.784
11	2	District Management	0.900	0.500	0.643
12	3	Administration	0.873	0.515	0.648
13	3	Social Protection Floor Mgmt.	0.914	0.660	0.766
14	3	Document Management	0.885	0.697	0.780
15	3	Rental House Management	0.961	0.762	0.850
16	3	Renovation Document Registration	0.983	0.828	0.899
17	3	House Applicant Management	0.955	0.892	0.922
18	3	Personal File Management	0.892	0.552	0.682
19	3	Rural House Management	0.840	0.534	0.653
20	3	Developer Management	0.853	0.596	0.702
21	3	Renovation Management	0.968	0.689	0.805
22	3	Emancipation Grant Management	0.899	0.508	0.649
23	3	Second-Hand House Management	0.894	0.724	0.800
24	4	Schedule Management	0.820	0.702	0.756
25	4	Qualification Management	0.918	0.767	0.836
26	4	Class Group Management	0.564	0.667	0.611
27	4	Evaluation	0.204	0.636	0.309
28	4	Remark Management	0.771	0.851	0.809
29	4	Reporting	0.731	0.667	0.698
30	4	Notice Management	0.581	0.556	0.568
31	4	Permission Management	0.571	0.511	0.539
32	4	Student Management	0.529	0.486	0.507
33	4	Subject Management	0.529	0.429	0.474
34	4	Teacher Management	0.932	0.786	0.853
35	5	Provider Management	0.697	0.390	0.500
36	5	Product Management	0.500	0.667	0.571
37	5	Event Management	0.676	0.532	0.595
38	5	Sending Management	0.725	0.578	0.643
39	5	Reporting	0.643	0.545	0.590
40	5	Relationship Management	0.698	0.545	0.612
41	6	Patient Admission	0,400	0,400	0,400
42	6	Data Collection Management	1.000	0.400	0.571
43	6	Data Analysis Management	0.800	1.000	0.889
44	6	Patient Stay Management	0.333	0.750	0.462
		Mean	0,763	0,604	0,656
		Standard Deviation	0,180	0,156	0,145
		Median	0.579	0.822	0.649

The obtained results are aligned with other business process mining proposals (cf. Section III), which achieve recall values higher than precision values. The lower precision is due to two main reasons.

First, all reverse engineering techniques, such as business process mining techniques, lose specific knowledge when the information is represented in a higher abstraction level. For the particular business process mining technique, the low precision is a consequence of the significant number of non-relevant tasks. In turn, this is due to the fact that some tasks are basically obtained from the source code and are related to the technical nature, and do not therefore represent any piece of the embedded business knowledge. Second, an inverse relationship exists between both recall and precision measures. Taking into account this circumstance, automated business process mining techniques achieve high recall more easily than high precision.

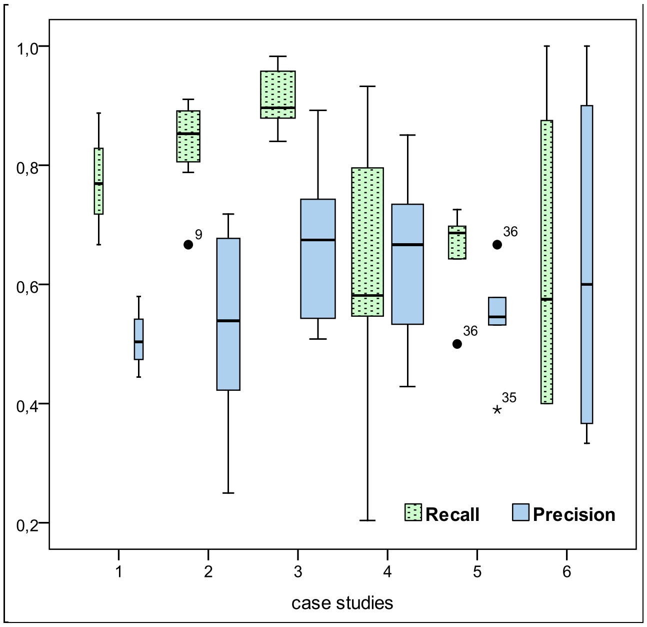

Figure 3. Box chart of recall and precision for each case study

Regardless, the aforementioned analysis of the obtained data was not sufficient to answer the research question. If we consider the reference value of 0.5 as used by *Lucrédio et al.* [21], the research question could be positively answered, since recall and precision are above 0.5. However, that reference value was selected heuristically, which limits the reliability of the conclusions. For this reason, in the present work, goodness thresholds for recall and precision were achieved by applying the Bender method.

V. THRESHOLD DEFINITION

A. The Bender Method

In this work, we used the Bender method to extract thresholds for both recall and precision measures. Thresholds help us to assess the numerical results of these measures, creating the probability of finding adequate results. This method was developed to discover thresholds in epidemiological studies. The Bender method assumes that the risk of an event happening is constant below a specific value (i.e., the threshold) and increases according to a logistic equation. The Bender method defines a "Value of an Acceptable Risk Level" (VARL), defined as (4).

$$VARL = \frac{1}{\beta}\cdot\left(ln\left(\frac{p_0}{1-p_0}-\alpha\right)\right) \qquad (4)$$

In equation (4), p_0 is the probability of an event occurring. This value is indicated by the author and can vary from 0 to 1. For example, p_0 = 0.6 indicates that there is a probability of 0.6 for considering recall or precision as appropriate. On the other hand, α and β are coefficients of a logistic regression equation, shaped as (5).

$$y = \alpha + \beta\cdot measureX \qquad (5)$$

The independent variable in the logistic regression equation is the measure or measures (denominated as *measureX* in (5)), which we want to extract their thresholds. The dependent variable must be a binary variable and is usually called *goodness*. This dependent variable indicates if the independent variables are considered good, taking into account some factors such as error degree, suitability, and so on.

Sample size in regression analysis is extremely important and it can determinate if the study has a significant effect. An approximation of sample size can be calculated though four parameters as indicated in [9]: alpha probability level, number of predictors in the linear model, anticipated effect size and desired statistical power level. An adequate sample size ensures more reliable results.

The Bender method works by extracting different VARLs with the specific values of$p_{0,2}, p_{0,4}, p_{0,6}, p_{0,8}$. These four values of probability divide the measure distributions into five groups, which have associated linguistic labels for an accurate interpretation, ranging from "very low" to "very high". The application of the Bender method in this work is described in the next section.

B. Application of the Bender Method to Obtain Precision and Recall Thresholds

First, to apply the Bender method, the dependent and independent variables must be selected. The independent variables are the different measures analysed in this work, i.e., recall and precision. The dependent variable is a binary variable created specifically for this purpose. This variable is called 'goodness' and indicates whether both the recall and the precision are considered good, taking in account the F-measure value. As described in Section II, the F-measure evaluates the relationship between recall and precision, because the idea is to obtain their maximum value at the same time. For this reason, we considered that in previously analysed data (c.f. Section IV), the retrieved business process which obtains an F-measure higher than the median of its own distribution is considered valid. Thereby, we consider a binary variable denominated as *goodness* to be 1 and 0 when the F-measure value is respectively under or above the median. For example, for a specific value of the F-measure of 0.75 (see subject 6 in TABLE III), the recall is 0.91 and the precision is 0.63. We considered these values as *appropriate* because the F-measure is higher than the median (0.65). In that particular case, the value of the goodness measure is 1. On the other hand, the F-measure is 0.38 when the recall is 0.84 and the precision 0.25 (see subject 8 in TABLE III). Although recall has a high value (which can indicate whether the retrieval is good), the precision value is too low to consider the entire retrieval process as *appropriate*. As a consequence, the *goodness* value is 0 (i.e., not appropriate).

Although the new measure goodness involves an assessment of the measures (recall and precision) the statistical method selected (Bender method) is considered more powerful by the way it not only indicates limit values. The selected method also investigates a threshold effect by the association of probabilities of considering the retrieval process acceptable or not.

A logistic regression analysis was applied to the experimental data provided in TABLE III. The minimum sample size for the experimental data has been calculated with the following configuration according to [9]:

- Alpha probability level, also known as the p-value, is 0.05 to ensure statistical significance.
- Number of predictors in the linear model. In this case, a regression equation for each variable is needed, so the number of predictors is 1.
- The anticipated effect size (f^2) is 0.2 which is a medium value, since by convention effect sizes of 0.02, 0.15 and 0.35 are considered small, medium and large.
- The value of desired statistical power level is 0.8, since by convention this value is always greater than or equal to 0.8.

As a result, a sample size greater than or equal to 41 is adequate in order to obtain a meaningful regression analysis. In our case, 44 data entries are available (see TABLE III).

After applying a logistic regression analysis to the experimental data, we obtained α and β coefficients, which can be used to obtain VARLs values (see TABLE IV).

TABLE IV. α AND β COEFFICIENTS

Results	**Precision**	**Recall**
α	-6.77	-12.46
β	11.28	15.73
VARL$p_{0,2}$	0.47	0.70
VARL$p_{0,4}$	0.56	0.76
VARL$p_{0,6}$	0.63	0.81
VARLp_{08}	0.72	0.88

TABLE IV depicts the threshold values, which can be interpreted as follows: “There is a probability of 0.2 for considering the retrieval of a business process acceptable if the measure precision is 0.47 and recall is 0.70”.

We assigned linguistic labels to these different groups in order to provide an interpretation for the numerical results, with the following results:

For recall:

- if recall ≤ 0.70 →*very low recall*
- if 0.70 < recall ≤ 0.76 → *low recall*
- if 0.76 < recall ≤ 0.81 →*medium recall*
- if 0.81 < recall ≤ 0.88 →*high recall*
- if recall > 0.88 →*very high recall*

For precision:

- if precision ≤ 0.47 →*very low precision*
- if 0.47 < precision ≤ 0.56 →*low precision*
- if 0.56 < precision ≤ 0.63 →*medium precision*
- if 0.63 < precision ≤ 0.72 →*high precision*
- if precision> 0.72 →*very high precision*

As noted in the earlier sections, precision and recall should be evaluated in conjunction, represented as the F-measure. The idea is to maximize both measures, because when good retrieval occurs, precision and recall have high values. However, case studies reveal that maximising both values involves high costs, and it is difficult to obtain high values for both measures. This leads to combining the precision and recall results to obtain a general evaluation of the retrieval process. TABLE V provides the range of F-measure values for each pair of combined labels of both recall and precision measures.

In TABLE V, it is possible to observe the evaluation of the F-measure through the combination of minimum and maximum precision and recall values for each threshold. Each cell represents the minimum and maximum value of the F-measure that can be achieved with the respective levels of recall and precision. The best choice is when a high F-measure is achieved. TABLE V shows some highlighted cells that respectively indicate the recall and precision thresholds that provide F-measures above 0.65, 0.70 and 0.75. We consider these combinations as the appropriate values for the effectiveness of MARBLE. Thereby, the effectiveness of MARBLE is considered as *acceptable* in two minimal cases: (i) if precision is at least medium when recall is high; or (ii) if recall is at least low when precision is high. As a result, for MARBLE to be considered an effective mining technique, there must be at least (i) a precision above 0.56 in combination with a recall above 0.81; or (ii) a recall above 0.70 in combination with a precision above 0.63. Despite these acceptable values, the most appropriate effectiveness would be achieved when the recall is at least high and precision is very high (see highlighted cells in TABLE V). This means that the best values for recall and precision must be above 0.81 and 0.72, respectively.

TABLE V. ASSESSMENT OF F-MEASURE

		Recall				
		Very low	Low	Medium	High	Very high
Precision	Very low	(0, 0.56)	(0, 0.58)	(0, 0.59)	(0, 0.61)	(0, 0.64)
	Low	(0, 0.62)	(0.56, 0.64)	(0.58, 0.66)	(0.59, 0.68)	(0.61, 0.72)
	Medium	(0, 0.66)	0.62, 0.69)	(0.64, 0.71)	(0.66, 0.73)	(0.68, 0.77)
	High	(0, 0.71)	(0.66, 0.74)	(0.69, 0.76)	(0.71, 0.79)	(0.73, 0.84)
	Very high	(0, 0.82)	(0.71, 0.84)	(0.74, 0.90)	(0.76, 0.94)	(0.79, 1)

F-measure Intervals	≥0.65	≥0.70	≥0.75

VI. DISCUSSION OF THE RESULTS AND IMPLICATIONS

This section discusses the thresholds obtained by applying the Bender method. First, the thresholds for recall and precision are used to evaluate the input data from the family of case studies in Section IV. Finally, the threats to the validity of this study are presented.

A. *Interpretation of Experimental Data*

The thresholds obtained for the precision and recall measures were used to provide a more confident answer to the research question put forth in Section IV.

If we compare the total mean for precision and recall (see TABLE III) with the thresholds obtained, we observe that both recall and precision measures are *medium*, since the mean of the recall is 0.763 (between 0.76 and 0.81) and the mean of the precision is 0.604 (between 0.56 and 0.63). As a result, the answer to the research question is more finely tuned with this additional information. Now, it is possible to state that the proposed technique retrieves business processes with a medium effectiveness, which means that it does not have the most appropriate performance.

In order to evaluate which case studies reported an appropriate effectiveness, the obtained thresholds can also be applied to the mean of the recall and precision of each case study. TABLE VI shows the interpretation of the results obtained in each case study using the proposed thresholds and linguistic labels. The best results are obtained for the third case study, and the results obtained in the first and second case studies can be considered acceptable. However, the fourth and sixth case studies have high precision, but very low recall. Finally, the effectiveness results obtained in the fifth case study cannot be considered appropriate, since the precision is low and the recall very low.

TABLE VI. Application of thresholds for each case study

		Recall				
		Very low	Low	Medium	High	Very high
Precision	Very low					
	Low	S5		S1	S2	
	Medium					
	High	S4, S6				S3
	Very high					

Additionally, the obtained thresholds may be individually applied to each business process to discover which processes were retrieved with appropriate levels of recall and precision. For instance, when business processes with a low precision and low recall are obtained, they are not considered relevant. These results also can aid in the improvement of business process mining techniques by observing the features common to the inappropriate business processes, leading to certain modifications.

B. *Threats to the Validity*

This section shows the threats to the validity of this study as well as the list of possible actions to mitigate them. We mainly consider two types of validity: internal and external.

At least two key threats to the *internal validity* exist. The first threat is related to the dependent variable selected to adapt the Bender method. The logical *goodness* variable is selected taking into account the median of the F-measure distribution. In this respect, the concept of *goodness* could be completed by other general characteristics about the retrieval process, for example, the subjective opinion of business experts or their satisfaction with the result, or other internal characteristics of the process such as the structural complexity. However, it is not easy to carry out large empirical studies involving appropriate business experts in each case. In addition, the business expert viewpoint could introduce a bias to those studies.

The second threat to the internal validity is related to the linguistic labels. These labels were established considering five equidistant probabilistic ranges, i.e., 0 to 0.2; 0.2 to 0.4; 0.4 to 0.6; 0.6 to 0.8 and 0.8 to 1.0, by configuring p_0 in the equation (4). However, these ranges could be established considering a different number of ranges as well as considering different sizes for each range of probability. For instance, *Miller* [24] states that seven, plus or minus two, is the most appropriate number to establish different linguistic labels. To mitigate this threat, the study could be replicated by adapting the Bender method using different ranges of probability taking into account expert's opinion according to qualitative research. The results could be then compared to provide strengthened conclusions.

Regarding *external validity*, this study considers the domain of business process mining techniques that use recall and precision to evaluate their effectiveness to be the whole population. Thus, the obtained thresholds could be generalized to this population. However, the experimental data are extracted from the application of a specific technique, MARBLE. This limits the extent to which the thresholds obtained in this study can be generalised to other business process mining techniques based on approaches or principles different to MARBLE. In order to mitigate this threat, the study should be replicated by extending the experimental data population through the consideration of different business process mining techniques. The threshold values extracted through the Bender method are dependent on the experimental data since it uses a logistic regression. For this reason, a larger set of case studies could improve the likelihood of obtaining extrapolated results.

VII. CONCLUSIONS AND FUTURE WORK

A great variety of business process mining techniques and algorithms have been proposed in the literature. The effectiveness of some of these proposals has usually been evaluated using recall and precision, but due to the lack of thresholds for measures, the analysis and interpretation of the obtained results is a difficult and error-prone activity. This paper proposes goodness thresholds for recall and precision values to evaluate the effectiveness of the business process mining technique MARBLE. As a result, five linguistic labels for each measure (very low, low, medium, high, very high) were provided to indicate the goodness of a recall or precision value obtained for a particular case. The results indicate that MARBLE has an appropriate effectiveness when, at least, precision is medium and recall is high or recall is low and precision is high.

Obtaining thresholds for such measures is partially context-specific. However, this work provides a novel approximation which can be applied to other business process mining techniques to obtain their own reference values which can be used to compare their effectiveness. This is an important aspect which can be used to facilitate the selection of the most effective business process mining technique depending on the context of the application.

According to the exposed validity threats, a plan is under way to conduct replications of this study considering other business process mining techniques, as well as different dependent variables in the Bender method.

Acknowledgement

This work was supported by the FPU Program from the Spanish Ministry of Education and the R&D projects ALTAMIRA (JCCM, PII2I09-0106-2463), PEGASO/MAGO (MICIN and FEDER, TIN2009-13718-C02-01), and MEDUSAS (CDTI (MICINN), IDI-20090557).

REFERENCES

[1] Aalst, W.M.P.v.d., et al., Workflow mining: a survey of issues and approaches. Data Knowl. Eng., 2003. 47(2): p. 237-267.

[2] Asencio, A., et al., Relating Expectations to Automatically Recovered Design Patterns, in Proceedings of the Ninth Working Conference on Reverse Engineering (WCRE'02). 2002, IEEE Computer Society. p. 87.

[3] Bender, R., Quantitative Risk Assessment in Epidemiological Studies Investigatin Threshold Effects. Biometrical Journal, 1999. 41(3): p. 305-319.

[4] Benlarbi, S., et al., Thresholds for Object-Oriented Measures. Institute for Information Technology, National Research Council Canada, 2000.

[5] Biermann, A.W. and J.A. Feldman, On the Synthesis of Finite-State Machines from Samples of Their Behavior. IEEE Trans. Comput., 1972. 21(6): p. 592-597.

[6] Brereton, P., et al., Using a protocol template for case study planning, in Evaluation and Assessment in Software Engineering (EASE'08). 2008: Bari, Italia. p. 1-8.

[7] Cardoso, J., Process control-flow complexity metric: An empirical validation. SCC '06: Proceedings of the IEEE International Conference on Services Computing, 2006: p. 167--173.

[8] Castellanos, M., et al., Business Process Intelligence, in Handbook of Research on Business Process Modeling, J. J. Cardoso and W.M.P. van der Aalst, Editors. 2009, Idea Group Inc. p. 456-480.

[9] Cohen, J., Applied multiple regression/correlation analysis for the behavioral sciences. 2003: Lawrence Erlbaum.

[10] Coleman, D., B. Lowther, and P. Oman, The Application of Software Maintainability Models in Industrial Software Systems. Journal of Systems and Software, 1995. 29(1): p. 3-16.

[11] Chidamber, S.R. and C.F. Kemerer, A Metrics Suite for Object Oriented Design. IEEE Transactions on Software Engineering, 1994. 20(6): p. 476-493.

[12] Davis, J. and M. Goadrich, The relationship between Precision-Recall and ROC curves, in Proceedings of the 23rd international conference on Machine learning. 2006, ACM: Pittsburgh, Pennsylvania. p. 233-240.

[13] Frakes, W.B. and T.P. Pole, An Empirical Study of Representation Methods for Reusable Software Components. IEEE Trans. Softw. Eng., 1994. 20(8): p. 617-630.

[14] García, F., et al., Towards a Consistent Terminology for Software Measurement. Information and Software Technology, 2005. 48: p. pg 631-644.

[15] Garcia, V.C., et al., From Specification to Experimentation: A Software Component Search Engine Architecture, in 9th International Symposium on Component-Based Software Engineering (CBSE 2006). 2006, Springer-Verlag: Västerås, Sweden. p. 82-97.

[16] Heuvel, W.-J.v.d., Aligning Modern Business Processes and Legacy Systems: A Component-Based Perspective (Cooperative Information Systems). 2006: The MIT Press.

[17] ISO/IEC, ISO/IEC DIS 19506. Knowledge Discovery Meta-model (KDM), v1.1 (Architecture-Driven Modernization). http://www.iso.org/iso/iso_catalogue/catalogue_ics/catalogue_detail_ics.htm?ics1=35&ics2=080&ics3=&csnumber=32625. 2009, ISO/IEC. p. 302.

[18] Jeston, J., J. Nelis, and T. Davenport, Business Process Management: Practical Guidelines to Successful Implementations. 2nd ed. 2008, NV, USA: Butterworth-Heinemann (Elsevier Ltd.). 469.

[19] Koskinen, J., et al. Software Modernization Decision Criteria: An Empirical Study. in European Conference on Software Maintenance and Reengineering. 2005: IEEE Computer Society.

[20] Lou, J.-G., et al., Mining program workflow from interleaved traces, in Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. 2010, ACM: Washington, DC, USA. p. 613-622.

[21] Lucrédio, D., R.P.M. Fortes, and J. Whittle, MOOGLE: A Model Search Engine, in 11th international conference on Model Driven Engineering Languages and Systems. 2008, Springer-Verlag: Toulouse, France. p. 296-310.

[22] McCabe, T.J., A Complexity Measure. IEEE Transactions on Software Engineering, 1976. SE-2(4): p. 308-320.

[23] Mendling, J., Metrics for Process Models: Empirical Foundations of Verification, Error Prediction, and Guidelines for Correctness. 2008: Springer Publishing Company, Incorporated.

[24] Miller, G.A., The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychological Review, 1956. 63(2): p. 81--97.

[25] Nakache, D., E. Metais, and J.F. Timsit, Evaluation and NLP, in Database and Expert Systems Applications, K.V. Andersen, J. Debenham, and R. Wagner, Editors. 2005, Springer Berlin / Heidelberg. p. 626-632.

[26] Nejmeh, B.A., NPATH: a Measure of Execution Path Complexity and its Applications. ACM, 1988. 31(2): p. 188-200.

[27] OMG, QVT. Meta Object Facility (MOF) 2.0 Query/View/Transformation Specification. http://www.omg.org/spec/QVT/1.0/PDF. 2008, OMG.

[28] OMG, Business Process Model and Notation (BPMN) 2.0. 2009, Object Management Group. p. 496.

[29] Pérez-Castillo, R., I.G.-R. de Guzmán, and M. Piattini, Business Process Archeology using MARBLE. Information and Software Technology, 2011. 53: p. 1023–1044.

[30] Pérez-Castillo, R., et al., Business Process Patterns for Software Archeology, in 25th Annual ACM Symposium on Applied Computing (SAC'10). 2010, ACM: Sierre, Switzerland. p. 165-166.

[31] Pérez-Castillo, R., I. García-Rodríguez de Guzmán, and M. Piattini, Implementing Business Process Recovery Patterns through QVT Transformations, in International Conference on Model Transformation (ICMT'10). 2010, Springer-Verlag: Málaga, Spain. p. 168-183.

[32] Pérez-Castillo, R., et al., A Case Study on Business Process Recovery using an E-Government System. Software Practice & Experience Journal, 2011: p. In Press.

[33] Pérez-Castillo, R., et al., An Empirical Comparison of Static and Dynamic Business Process Mining, in 26th Annual ACM Symposium on Applied Computing (SAC'11). 2011, ACM: TaiChung, Taiwan. p. 269-276.

[34] Philippow, I., et al., An approach for reverse engineering of design patterns. Software and Systems Modeling, 2005. 4(1): p. 55-70.

[35] Povzner, A., et al., Autograph: automatically extracting workflow file signatures. SIGOPS Oper. Syst. Rev., 2009. 43(1): p. 76-83.

[36] Pradel, M., P. Bichsel, and T.R. Gross, A Framework for the Evaluation of Specification Miners Based on Finite State Machines, in 26th IEEE International Conference on Software Maintenance (ICSM'10). 2010: Timişoara, Romania. p. In Press.

[37] Raghavan, V., P. Bollmann, and G.S. Jung, A critical investigation of recall and precision as measures of retrieval system performance. ACM Trans. Inf. Syst., 1989. 7(3): p. 205-229.

[38] Runeson, P. and M. Höst, Guidelines for conducting and reporting case study research in software engineering. Empirical Softw. Eng., 2009. 14(2): p. 131-164.

[39] Sánchez González, L., et al., Measurement in Business Processes: a Systematic Review. Business process Management Journal, 2010. 16(1): p. 114-134.

[40] Shatnawi, R., An Investigation of CK Metrics Thresholds. ISSRE Supplementary Conference Proceedings, 2006.

[41] van der Aalst, W., H. Reijers, and A. Weijters, Business Process Mining: An Industrial Application. Information Systems, 2007. 32(5): p. 713-732.

[42] Van Rijsbergen, C.J., Information Retrieval. 1979: Butterworths.

[43] Weske, M., Business Process Management: Concepts, Languages, Architectures. 2007, Leipzig, Alemania: Springer-Verlag Berlin Heidelberg. 368.

[44] Ye, Y. and G. Fischer, Supporting reuse by delivering task-relevant and personalized information, in Proceedings of the 24th International Conference on Software Engineering. 2002, ACM: Orlando, Florida. p. 513-523.

[45] Yeh, D., et al., An Empirical Study of a Reverse Engineering Method for Aggregation Relationship Based on Operation Propagation, in Proceedings of the 29th Annual International Computer Software and Applications Conference - Volume 01. 2005, IEEE Computer Society. p. 95-100.