Information Systems 47 (2015) 1–14

Contents lists available at ScienceDirect

Information Systems

journal homepage: www.elsevier.com/locate/infosys

Feasibility and effort estimation models for medium and small size information mining projects

Pablo Pytel [a,b,c], Alejandro Hossian [d], Paola Britos [e], Ramón García-Martínez [b,*]

[a] PhD Program on Computer Science, Computer Science School, National University of La Plata, Buenos Aires, Argentina
[b] Information Systems Research Group, National University of Lanus, Buenos Aires, Argentina
[c] Information Systems Engineering Department, Technological National University at Buenos Aires, Argentina
[d] Intelligent Systems Applied to Engineering Research Group, Technological National University at Neuquén, Neuquén, Argentina
[e] Information Mining Research Group, National University of Rio Negro at El Bolson, Río Negro, Argentina

ARTICLE INFO

Article history:
Received 3 April 2014
Received in revised form
20 June 2014
Accepted 21 June 2014
Available online 28 June 2014

Keywords:
Data mining
Effort estimation model
Feasibility model
SMEs

ABSTRACT

Information mining is a sub-discipline of Information Systems which provides the non-trivial knowledge needed for making decisions inside an organization. Although such projects have different features from Software Engineering ones, they share some of their problems. Among these problems two are highlighted: unmanaged risks and inaccurate estimations of necessary resources to complete the project. In this context, this paper presents two ad-hoc models to be applied in Small and Medium-sized Enterprises: one for assessing project feasibility and the other for estimating the resources and time required to carry out the project. Both models should be applied at the beginning of the project.

1. Introduction

The organization management needs a lot of information for their decision-making process and the generation of strategic plans [1]. This valid and useful non-trivial information is normally referred to as knowledge [2]. This knowledge is located implicitly in the available data repositories in the organization and it can be extracted using the synthesis tools provided by Data Mining [3]. Data Mining focuses on the technology to be applied (i.e. tools and algorithms), while information mining focuses on which task and procedure must be developed to accomplish the project goals. In [4] "Information Mining" term refers to the sub-discipline of Information Systems, it studies, proposes and develops: processes, methods, techniques and methodologies to run this kind of project successfully. Consequently, it can be said that Data Mining is close to programming tasks, while Information Mining is close to Software Engineering activities.

The processes, methods, techniques and tools that come from Software Engineering cannot be used in Information Mining projects because of differences in goals and practical aspects between these two kinds of projects [5]. This means that ad-hoc processes, methods, techniques, tools, and methodologies should be developed considering Information Mining project main features. On the other hand, the methodologies most commonly used for Information Mining projects are CRISP-DM [6], SEMMA [7] and P3TQ [8]. These methodologies are considered as proven by the community, but they exhibit problems when trying to define the phases related to project management [4]:

- project management elements are mixed with the knowledge discovery process,

[*] Corresponding author.
E-mail addresses: ppytel@gmail.com (P. Pytel), alejandrohossian@yahoo.com.ar (A. Hossian), paobritos@gmail.com (P. Britos), rgm1960@yahoo.com (R. García-Martínez).

http://dx.doi.org/10.1016/j.is.2014.06.004
0306-4379/© 2014 Elsevier Ltd. All rights reserved.

- they do not indicate the methods to be used for project monitoring, verification and measurement, and
- project characteristics performed within Small and Medium-sized Enterprises (SMEs) are not analyzed.

Moreover, conducted studies about Information Mining projects have detected that not all projects are completed successfully [9] and that there is a significant percentage of projects that fail [10]. In 2000, 85% of the projects failed to achieve its goals [11]. In other words, only 15 out of a 100 developed projects have been completed successfully. After nine years working, the community has been able to reduce this project failure rate to approximately 50% [12]. Therefore, we can say that the community is working in the right way but there are project elements that should be enhanced yet. In this context, this paper presents two ad-hoc models to be applied in Small and Medium-sized Enterprises: one for assessing project feasibility and the other one for estimating resources and time required to carry out the project. Both models should be applied at the beginning of the project. The article structure is as follows: first, we describe the main problem (Section 2), then we present the proposed models (Section 3) and the validation results (Section 4). Finally, the research work presents the main conclusions (Section 5).

2. Project failures' analysis

Most Software Engineering projects can be considered (at least) partial failures because few projects meet all their cost, schedule, quality or required objectives [13]. From challenged or canceled projects, the project final cost average was 189% over budget, the project final time average was 222% on schedule, and contained only 61% (average) of the originally specified features [14]. Based on a survey carried out by the Standish Group [15], the top 10 reasons causing failure of software development projects are

1. incomplete requirements (13.1%),
2. lack of user involvement (12.4%),
3. lack of resources (10.6%),
4. unrealistic expectations (9.9%),
5. lack of executive support (9.3%),
6. changing requirements & specifications (8.7%),
7. lack of planning (8.1%),
8. did not need it any longer (7.5%),
9. lack of IT management (6.2%) and
10. technology illiteracy (4.3%).

Most of these reasons are related to requirements handling (indicated by 41.7% of surveyed people) and can be solved by applying methodologies and good practices [16], considering the Information Mining project characteristics [17]. However, it is also important to note the problems associated to resource planning (indicated by the 18.7%) and unrealistic objectives and expectations (9.9%). These last two problems must be handled from the project initial activities.

Before starting any traditional software project, the organization must decide whether it is appropriate to do it or not. Making such decisions is complex and depends on multiple factors: it is necessary to know both, the software impact on the organization and its developing associated risks [18]. This requires analyzing the project features by assessing the project technical and economic feasibility (commonly known as feasibility study). Once the project is considered feasible, it is necessary to predict the effort required to perform the complete project. With this information, it is possible to estimate the necessary resources and the associated cost [19].

Information Mining project initial tasks are similar to traditional software projects. By early risk detection, its effects could be reduced during the project development. But, given that the project features are different from traditional software projects, the existing models cannot be applied in Information Mining projects and, therefore, it is necessary to specify ad-hoc ones.

3. Proposed models

This section presents two ad-hoc models proposed to be used at the beginning of an Information Mining project performed within Small and Medium-sized Enterprises (SMEs). First, the SMEs project characteristics are summarized (Section 3.1), then the model to assess the project feasibility is presented (Section 3.2) and, finally, the model that allows estimating the resources and time required to perform the project is described (Section 3.3).

These models have been specified based on actual information mining projects collected by researchers from the following research groups: Research Group on Information Systems from Universidad Nacional de Lanús (GISI-UNLa), Information System Methodologies Research Group from the Universidad Tecnológica of Buenos Aires (GEMIS-FRBA-UTN), and Information Mining Research Group from Universidad Nacional of Río Negro (SAEB-UNRN). All these projects have been performed by applying the CRISP-DM methodology [6] and then the proposed models can be considered reliable only for Information Mining projects developed with this methodology.

3.1. SMEs' information mining projects

According to the Organization for Economic Cooperation and Development (OECD) Small and Medium-sized Enterprises (SMEs) and Entrepreneurship Outlook report [20]: "SMEs constitute the dominant form of business organization in all countries world-wide, accounting for over 95% and up to 99% of the business population depending on the country". However, although the importance of SMEs is well-known, there is no universal criterion for characterizing them. Depending on the country and region, there are different quantitative and qualitative parameters used to recognize a company as SMEs. For instance, in Latin America each country has a different definition [21]: Argentina considers as SME all independent companies that have an annual turnover lower than USD 20,000 (maximum amount in U.S. dollars that depends on the company's activities), Brazil includes all companies with 500 employees or less. On the other hand, the European Union defines as SMEs all companies with

250 employees or less, assets lower than USD 60,000 and gross sales lower than USD 70,000 per year. In that respect, the International Organization for Standardization (ISO) has recognized the necessity to specify a standard software engineering for SMEs and thus it is working on the ISO/IEC 29110 standard "Lifecycle profiles for Very Small Entities" [22]. The term 'Very Small Entity' (VSE) was defined by the ISO/IEC JTC1/SC7 Working Group 24 [23] as being "an entity (enterprise, organization, department or project) that has up to 25 people".

We use as definition of an information mining project for SMEs one that complies with [a] a project performed at a company of 250 employees or less (in one or several locations), [b] the managers are usually the company's owners, and [c] the management needs to use non-trivial knowledge implicit in company data repositories to solve a business problem, avoiding (as much as possible) special risks. As the company's employees do not usually have the necessary experience, the project is performed by contracting outsourced consultants. In our experience, the project team can be restricted to up to 25 people (including both the outsourced consultants and the involved company staff) with maximum project duration of one year.

At the beginning of the project, the consultants need to elicit both the stakeholders´ necessities and desires, and also the characteristics of the available data sources within the organization (i.e. existing data repositories). Although, the consultants must have knowledge and experience in developing information mining projects, they might not have experience in running projects on the current business type; which could complicate the tasks of understanding the organization and its related data. Additionally, company's data repository staff should be interviewed because data repositories are often not properly documented. However, company experts are normally scarce and reluctant to get involved in the knowledge elicitation sessions. Thus, it is required the willingness and commitment of personnel and supervisors to identify the correct characteristics of the organization and data location needed for the project. As the project duration is quite short and the structure of the organization is centralized, it is considered that the project requirements will not change.

Finally, SMEs infrastructure in Information and Communication Technology (ICT) is analyzed. Currently, most of Latin America's SMEs have ICT infrastructure, but not all of them have automated services and/or proprietary software. Normally, commercial off-the-shelf software (such as spreadsheets, managers and document editors) is used to register the company's management and operational information. The data repositories are not large (less than one million records in our experience) but implemented in different formats and supported by different technologies. Therefore, data formatting, data cleaning and data integration tasks will have a considerable effort due to the lack of available software tools to perform them at SMEs scale.

3.2. Feasibility model for information mining projects

The feasibility model proposal for Information Mining projects [24] requires the identification of the main conditions that should be met to consider a project as feasible. These conditions have been identified based on [25–31] and classified into three groups (or dimensions) based on the same criteria used in Knowledge Engineering (KE) in the project feasibility test [32]:

- *Conditions that determine the project plausibility* include factors that make it possible to perform the project. It can be performed if the following conditions are met: business problem representative data to be solved are contained in the available company's data repositories, the business problem is understood and the team has a minimum knowledge about the information mining process.
- *Conditions that determine the project adequacy* include factors that determine whether Information mining is the appropriate solution for the identified business problem (i.e., it is the best solution for the problem). It is appropriate to apply information mining if the following conditions are met: the available data repositories are in digital format (i.e., they are not only available in paper), the business problem cannot be solved by using traditional statistical techniques, the business problem does not change during the project development and the data quality is good. The following metrics are used for assessing the data quality:
    - ◦ Quantity of attributes and records (measures the availability of enough data to apply the information mining process).
    - ◦ Degree of data credibility (measures how much you can trust on the data accuracy depending on the source and nature).
- *Conditions that determine the project success* include factors that ensure the project accomplishment. An information mining project will be successful if the following conditions are met: data repositories implemented with technologies allowing easy data access and manipulation (i.e., integration, cleaning, and formatting tasks), project stakeholders (senior managers, junior managers, end-users) support the project, it is possible to run the project planning considering best practices with required time, and that the team has experience in similar projects.

Here is a five-step procedure to analyze these conditions and thereby assess project feasibility:

STEP 1: Determining each project feature values

Seeking the information mining project characterization and evaluating its feasibility, the corresponding features should be identified from the interviews conducted in the organization. Such features (specified in Table 1) are based on the identified conditions.

For each feature, the following attributes are defined:

- ◦ *Category*: used only to group the features according to what or who is concerned.
- ◦ *ID*: indicates a code to uniquely identify the property and the dimension to which it belongs (plausibility, adequacy or success).
- ◦ *Condition*: describes the feature to be identified for characterizing the project.

- *Weight*: indicates the relative importance of each feature in the global model. These weights have been obtained after the project behavior simulations with Monte Carlo method [33].
- *Threshold*: indicates the value that the feature must be equal or bigger. If the feature does not exceed the threshold, it can be considered that the project is not feasible and it is not necessary to continue with the next steps.

However, it is not easy to meet these conditions by answering '*yes*'/'*no*' questions (or by giving a numerical value). Then, the model allows utilizing a range of linguistic values to answer each condition: '*nothing*', '*little*', '*regular*', '*much*', and '*all*'.

STEP 2: Converting feature values into fuzzy intervals

Once the linguistic values have been defined for each feature in Table 1, they should be translated into numeric values to calculate the project feasibility. The transformation process is described in the feasibility test of knowledge engineering projects [34]. For each word, the values of a fuzzy interval are defined and expressed by four numbers (ranging from zero to ten) that represents the breakpoints (or corner points) of the corresponding membership function. These intervals with the membership function graphic representation are shown in Fig. 1.

STEP 3: Calculating each dimension value.

We associate the fuzzy intervals with the weights in Table 1 in order to determine each project dimension value. The interval representing each dimension value ($\mathrm{I_d}$) is calculated with the following formula. This formula is formed by combining the harmonic mean and the arithmetic mean of the interval set. We aim to reduce the influence of low values when calculating the dimension value.

$$I_d = \left(\frac{1}{2} \frac{\sum_{i=1}^{n_d} W_{d_i}}{\sum_{i=1}^{n_d} \left(\frac{W_{d_i}}{F_{d_i}} \right)} \right) + \left(\frac{1}{2} \frac{\sum_{i=1}^{n_d} (W_{d_i} F_{d_i})}{\sum_{i=1}^{n_d} W_{d_i}} \right)$$

where I_d represents the fuzzy interval calculated for dimension d (using '*P*' for plausibility, '*A*' for adequacy and '*S*' for success), W_{di} represents the feature weight i for dimension d, F_{di} represents the fuzzy interval that has been assigned to the feature i for dimension d, n_d represents the quantity of features associated to dimension d.

As a result of the previous formula, another fuzzy interval is achieved. To convert this interval into a single numeric value (V_d) the arithmetic average is used as shown:

$$V_d = \frac{\sum_{i=1}^{4} I_{d_i}}{4}$$

where V_d represents the numeric value calculated for dimension d, I_{di} represents position i value of the fuzzy interval calculated for dimension d.

STEP 4: Calculating the overall project feasibility

In this step, the numerical values calculated in the previous step for each dimension (V_d) are combined by using a weighted arithmetic mean obtaining the overall project feasibility value (OV):

$$OV = \frac{8V_P + 8V_A + 6V_S}{22}$$

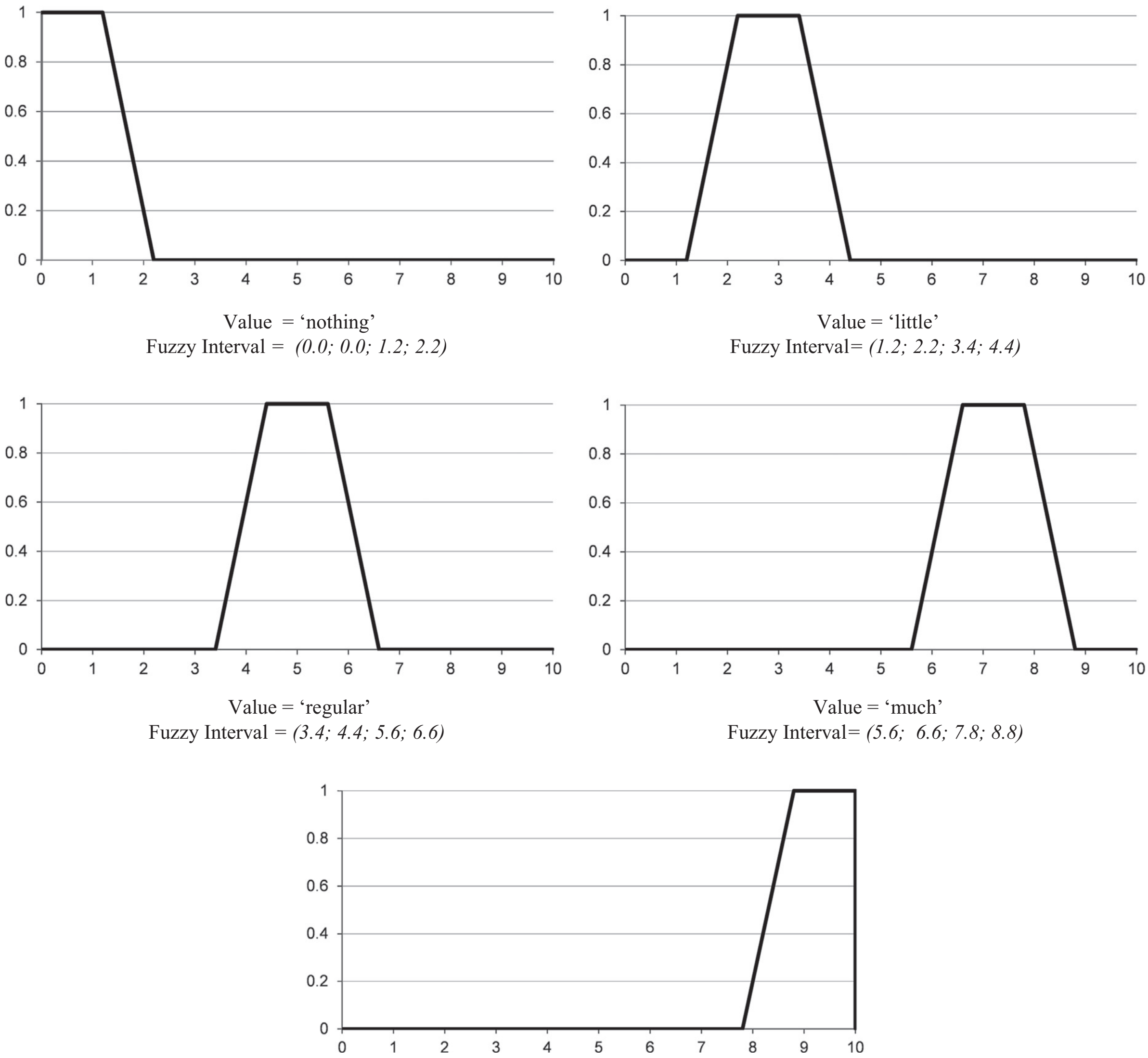
Where OV represents the overall project feasibility value, VP represents the value calculated for dimension plausibility, VA represents the value calculated for dimension adequacy, V represents the value calculated for dimension success.

STEP 5: Interpreting results

Finally, once the numeric values for each dimension and the overall project feasibility value have been calculated (steps 3 and 4 respectively), the results have to be analyzed. To interpret each dimension feasibility results, it is recommended plotting the corresponding membership function of the obtained fuzzy interval (I_d). The dimension viability can be considered as accepted if it exceeds the range of 'regular' value. Analyzing the dimension numeric value is another way to do it. If the dimension value (V_d) is greater than 5, the dimension can be considered as accepted. For analyzing project feasibility, the following criteria can be used: whether the three dimensions are accepted and the overall project feasibility (OV) is greater than 5,

Table 1
Project features evaluated by feasibility model

Category	ID	Condition	Weight	Threshold
Data	P1	How current is considered the data from the repositories?	8	Little
	P2	How is considered the data representativeness in the repositories in order to solve the business problem?	9	Little
	A1	How many data repositories are in digital format?	4	Little
	A2	How many attributes and records are available in the data repositories?	7	Little
	A3	How much credibility does the available data have?	8	Little
	S1	To what extent does the repository technology support the data manipulation?	6	Nothing
Business problem	P3	Is the business problem understood?	7	Little
	A4	To what extent can't the business problem be solved by traditional statistical techniques?	10	Little
	A5	How stable is considered the business problem during the project?	9	Little
Project	S2	How many stakeholders support the project?	8	Nothing
	S3	To what extent does the project plan consider the required time to perform best practices during the project?	7	Nothing
Project Team	P4	How much knowledge does the team have about information mining?	6	Little
	S4	How much experience does the team have in similar projects?	6	Nothing



Value = ‘nothing’
Fuzzy Interval = *(0.0; 0.0; 1.2; 2.2)*

Value = ‘little’
Fuzzy Interval= *(1.2; 2.2; 3.4; 4.4)*

Value = ‘regular’
Fuzzy Interval = *(3.4; 4.4; 5.6; 6.6)*

Value = ‘much’
Fuzzy Interval= *(5.6; 6.6; 7.8; 8.8)*

Value = ‘all’
Fuzzy Interval = *(7.8; 8.8; 10.0; 10.0)*

Fig. 1. Membership function graphical and fuzzy interval assigned to each word.

then the project is considered feasible. Otherwise, it is not feasible. In both cases, the engineer should also observe the project weaknesses to be strengthened.

3.3. *Effort estimation model for information mining projects*

The existing effort estimation methods applied in traditional software development projects cannot be used in Information Mining projects. Only one specific analytical estimation method for this project type has been found after a state-of-the-art review. The method called Data Mining Cost Model (or DMCoMo), defined in [5]. However, the DMCoMo analysis performed in [35] shows that this method tends to overestimate the efforts mainly in the little-sized projects that are usually required by SMEs. Therefore, it is necessary to specify an ad-hoc effort estimation method for that project type.

Considering the Information Mining project characteristics for SMEs, eight cost drivers are specified. Some cost drivers have been identified in this version because, as explained by [36], when an effort estimation method is created, many of the non-significant data should be ignored. Thus, the model is not too complex (and therefore impractical), the irrelevant and co-dependent variables are removed, and the noise is also reduced. These cost drivers have been selected based on the most critical CRISP-DM methodology tasks. Ref. [37] indicates that building data mining models and finding patterns is quite simple now because 90% of the effort is in the data preparation (i.e., ‘Data Preparation’ tasks performed in the CRISP-DM phase III). From our experience, the other critical tasks are related to

Table 2
OBTY cost driver values.

Value	Description
1	Try to identify the rules that characterize behavior or description of an already known class.
2	Try to identify an available data partition without having a previously known classification.
3	Try to identify the rules that characterize the data partitions without a previously known classification.
4	Try to identify the attributes with a greater incidence frequency on the behavior or description of an already known class.
5	Try to identify the attributes with a greater incidence frequency on a previously unknown class.

Table 3
LECO cost driver values.

Value	Description
1	Both managers and the organization's personnel are willing to collaborate with the project.
2	Only managers are willing to collaborate with the project while the rest of the company's personnel are not concerned with it.
3	Only senior managers are willing to collaborate with the project while junior managers and the rest of the company's personnel are not concerned with it.
4	Only senior managers are willing to collaborate with the project while junior managers are not willing to collaborate.

Table 4
AREP cost driver values.

Value	Description
1	Only 1 data repository available.
2	Between 2 and 5 data repositories of compatible technology.
3	Between 2 and 5 data repositories of non-compatible technology.
4	More than 5 data repositories of compatible technology.
5	More than 5 data repositories of non-compatible technology.

the ‘Business Understanding’ phase (i.e. tasks: ‘understand business background’ and ‘identify project success criteria’).

Proposed cost factors grouped in three as follows:

Cost drivers related to the Project:

- *Information Mining Objective Type* (*OBTY*)
  This cost driver analyses the information mining project objective and therefore the type of process to be applied based on the definition given by [38]. These cost driver values are indicated in Table 2.
- *Organization’s Collaboration Level* (*LECO*)
  The collaboration level from members of the organization is analyzed by checking whether senior management (i.e., usually SME's owners), junior management (supervisors and department heads) and operational personnel are willing to help consultants to understand the business and the related data (especially in the first phases of the project). If the information mining project has been hired, it is assumed that at least the senior management should support it. Possible values for this cost factor are shown in Table 3.

Cost Drivers related to the Available Data:

- *Quantity and type of the available data repositories* (*AREP*)
  The data repositories to be used by the information mining process are analyzed (including database management systems, spreadsheets, documents, etc.). In this case, the data repositories quantity (public or private from the company) and the implementation technology are studied. In this stage, it is not necessary to know the number of tables in each repository because their integration within a repository is relatively simple as it can be performed with a query statement. However, depending on the technology, the complexity of the data integration tasks could vary. The following criteria can be used:
  - If all the data repositories are implemented with the same technology, then the repositories are compatible for integration.
  - If the data can be exported to a common format, then the repositories can be considered compatible for integration because the data integration tasks will be performed by using the exported data.
  - On the other hand, if there are non-digital repositories (i.e., written papers), then the technology should not be considered compatible for integration. But the estimation method is not able to predict the required time to perform the digitalization because it could vary depending on many factors (such as quantity of papers, length, format, diversity, etc.).

  Values for this cost factor are shown in Table 4.
- *Total quantity of available tuples in the main table* (*QTUM*)
  This variable ponders the approximate quantity of tuples (records) available in the main table to be used when applying data mining techniques.

Table 5
QTUM cost driver values.

Value	Description
1	Up to 100 tuples in main table.
2	Between 101 and 1000 tuples in main table.
3	Between 1001 and 20,000 tuples in main table.
4	Between 20,001 and 80,000 tuples in main table.
5	Between 80,001 and 5,000,000 tuples in main table.
6	More than 5,000,000 tuples in main table.

Table 6
QTUA cost driver values.

Value	Description
1	No auxiliary tables used.
2	Up to 1000 tuples in auxiliary tables.
3	Between 1001 and 50,000 tuples in auxiliary tables.
4	More than 50,000 tuples in auxiliary tables.

Possible values for this cost factor are shown in Table 5.

- *Total quantity of available tuples in auxiliaries tables* (*QTUA*)
This variable ponders the approximate quantity of tuples (records) available in the auxiliary tables (if any) used to add information to the main table (such as a table used for determining the product features associated with the product ID of the sales main table). Normally, these auxiliary tables include fewer records than the main table. Possible values for this cost factor are shown in Table 6.
- *Knowledge level about data sources* (*KLDS*)
Knowledge level about data sources studies whether the data repositories and their tables are properly documented. In other words, if these items are present: a document defining the technology in which it is implemented, a description of the features of the tables' fields and a description of how data is created, modified, and/or deleted. If these documents are not available, it will be necessary to meet the experts (usually in charge of the data administration and maintenance) to create them. As a result, the project effort required will increase depending on the experts' collaboration to help the consultants.
Values for this cost factor are shown in Table 7.

Cost drivers related to Available Resources:

- *Knowledge and experience level of the information mining team* (*KEXT*)
This cost driver studies the outsourced consultants' ability to carry out the project. Team's knowledge and experience in similar previous projects are analyzed, considering similarity of the business type, data to be used and expected goals. It is assumed that when there is greater similarity the effort should be lower. Otherwise, the effort should increase. Possible values for this cost factor are shown in Table 8.

Table 7
KLDS cost driver values.

Value	Description
1	All the data tables and data repositories are properly documented.
2	More than 50% of data tables and data repositories are documented and there are experts available to explain the data sources.
3	Less than 50% of data tables and data repositories are documented, but there are experts available to explain the data sources.
4	Data tables and data repositories are not documented, but there are experts available to explain the data sources.
5	Data tables and data repositories are not documented and the experts available are not willing to explain the data sources.
6	Data tables and data repositories are not documented and there are not experts available to explain the data sources.

- *Functionality and usability of available tools* (*TOOL*)
This cost driver analyzes the features of the information mining tools to be utilized in the project and its implemented functionalities. Data preparation functions and data mining techniques are reviewed. Possible values of this cost factor are shown in Table 9.

Linear Formula:
Once the cost driver values had been specified, they were used to characterize 34 information mining projects with their actual effort collected by co-researchers (mentioned in Section 3).[1] A multivariate linear regression method [39] has been applied to obtain the following linear equation:

$$\begin{aligned}\mathrm{PEM_L} = {} & 0.80\mathrm{OBTY} + 1.10\mathrm{LECO} - 1.20\mathrm{AREP} \\ & - 0.30\mathrm{QTUM} - 0.70\mathrm{QTUA} + 1.80\mathrm{KLDS} \\ & - 0.90\mathrm{KEXT} + 1.86\mathrm{TOOL} - 3.30\end{aligned}$$

where $\mathrm{PEM_L}$ is the linear estimation effort method proposed for SMEs (in man-month), and the following cost drivers: information mining objective type (OBTY), collaboration level from the organization (LECO), data repositories quantity and type available (AREP), tuples total quantity available in the main table (QTUM) and in auxiliaries tables (QTUA), knowledge level about the data sources (KLDS), knowledge and experience level of the information mining team (KEXT), and functionality and usability of available tools (TOOL). The values for each cost driver are defined from Tables 2 to 9 respectively.

Empiric Method:
Although the linear formula is very accurate to estimate the necessary efforts (as shown in the results of Section 4.2), using this kind of formula is not considered completely reliable for all kind of projects. Therefore, a second estimation for this model is also proposed. This new method is similar to the COCOMO family methods [19]. Using the same defined cost drivers, the combination of their values has been analyzed to determine how they

[1] The data of the 34 projects used for the regression is available at http://tinyurl.com/lr5s5gm.

Table 8
KEXT cost driver values.

Value	Description
1	The information mining team has developed projects with similar data in similar business types to obtain the same objectives.
2	The information mining team has developed projects with different data in similar business types to obtain the same objectives.
3	The information mining team has developed projects with similar data in other business types to obtain the same objectives.
4	The information mining team has developed projects with different data in other business types to obtain the same objectives.
5	The information mining team has developed projects with different data in other business types to obtain other objectives.

Table 9
TOOL cost driver values.

Value	Description
1	The tool includes functions for data formatting and integration (allowing the importation of more than one data table) and data mining techniques.
2	The tool includes functions for data formatting and data mining techniques, and it allows importing more than one data table independently.
3	The tool includes functions for data formatting and data mining techniques, and it allows importing only one data table at a time.
4	The tool includes only functions for data mining techniques, and it allows importing more than one data table independently.
5	The tool includes only functions for data mining techniques, and it allows importing only one data table at a time.

Table 10
Decision table to determine CBUS coefficient.

LECO values	OBTY values	KLDS values	
		< 3	**≥ 3**
=**1**	–	*1.00*	*2.00*
≥ **2**	=**1**	2.00	*3.70*
	≥ **2**	3.70	

Table 11
Decision table to determine CDAT coefficient

QTUM values	QTUA values	AREP values	
		=**1**	≥ **2**
=**1**	–	*0.25*	
=**2**	–	*0.50*	
=**3**	=**1**	*1.50*	*2.40*
	≥ **2**	*2.40*	
≥ **4**	–	*2.40*	

Table 12
Decision table to determine CMOD coefficient.

TOOL values	OBTY values	
	=**1**	≥ **2**
< **4**	*0.60*	*0.80*
≥ **4**	*2.70*	*3.80*

may affect the project effort. For this reason, the expert researchers' opinion and Monte Carlo simulation results [33] have been used.

As a result, four coefficients have been identified to characterize a project:

- *Coefficient value of business complexity* (*CBUS*)
  This coefficient is associated with the difficulty of performing the project requirement elicitation and the organization analysis (performed during CRISP-DM 'Business Understanding' phase). To determine the value of this coefficient, the cost factors objective type of information mining (OBTY), collaboration level of the organization (LECO) and knowledge level about the data sources (KLDS) are combined as shown in Table 10.
- *Data complexity coefficient* (*CDAT*)
  This second coefficient is associated with the activities' difficulty connected to 'Data Understanding' and 'Data Preparation' (CRISP-DM phases). In this case, the considered cost factors are the tuples total quantity available in main table (QTUM), tuples total quantity available in auxiliaries tables (QTUA) and data repositories quantity available (AREP). They are combined as shown in Table 11.
- *Modeling complexity coefficient* (*CMOD*)
  This coefficient defines the modeling techniques complexity to be applied in the data prepared to obtain the project results (related to the last three CRISP-DM phases). This means that the cost factors considered are: information mining objective type (OBTY) and functionality of the tools available (TOOL). These are combined in the decision Table 12.
- *Adjustment coefficient of team experience* (*AEXP*)
  In order to set this adjustment value, the team experience and knowledge must be analyzed based on the cost factor KEXT as indicated in Table 13.

  Then, these coefficients are applied in a new formula that has been specified empirically to calculate required effort in man-months to develop the project completely:

$$\mathrm{PEM_E} = (1.80\mathrm{CBUS} + 0.90\mathrm{CDAT} + 1.40\mathrm{CMOD} - 1.50)\mathit{AEXP}$$

where $\mathrm{PEM_E}$ is the linear estimation effort method proposed for SMEs (in man-month); CBUS, CDAT, CMOD and AEXP are the coefficients values defined from Tables 10 to 13 respectively.

4. Proposed models' validation

In this section, the models' validation proposed in Section 3 is performed using 37 information mining projects' collected data. To perform this validation the calculated project values, by the corresponding model, are compared with the real values collected from a researchers' appraisal (considered experts in the domain).

Table 13
Decision table to determine AEXP coefficient.

KEXT values	AEXP
=**1**	*0.60*
=**2**	*0.70*
=**3**	
=**4**	*0.80*
=**5**	*1.00*

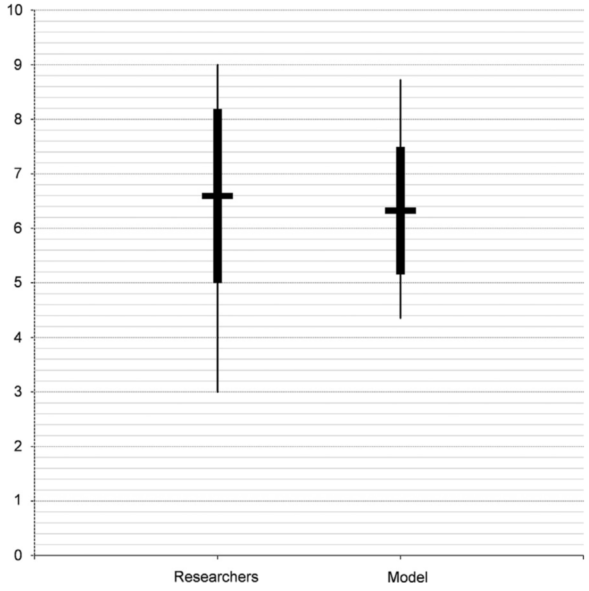


Fig. 2. Boxplot graph for plausibility dimension.

Table 14
Projects data used in the model validation feasibility.

#	Project appraisal provided by researchers				Values calculated by the model			
	Plausibility value	Adequacy value	Success value	Global feasibility value	Plausibility value (V_P)	Adequacy value (V_A)	Success value (V_S)	Overall project feasibility (OV)
P1	8	7	4	6.33	7.20	6.11	5.25	6.27
P2	7	6	5	6.00	6.87	5.07	5.25	5.77
P3	8	5	6	6.33	5.90	5.67	5.31	5.65
P4	6	6	4	5.33	5.12	6.95	4.12	5.51
P5	6	8	7	7.00	5.12	7.82	6.81	6.56
P6	6	5	5	5.33	5.45	5.61	5.25	5.45
P7	5	5	5	5.00	5.45	5.56	5.42	5.48
P8	6	5	6	5.67	6.45	5.80	5.18	5.87
P9	7	6	6	6.33	7.20	5.61	5.57	6.18
P10	6	5	6	5.67	5.85	5.34	5.57	5.59
P11	8	5	6	6.33	6.22	6.56	5.42	6.14
P12	7	8	7	7.33	7.67	7.35	6.45	7.22
P13	7	5	6	6.00	5.93	5.09	7.05	5.93
P14	7	7	6	6.67	6.20	6.59	5.69	6.20
P15	9	7	8	8.00	8.72	6.89	7.66	7.77
P16	7	6	5	6.00	6.45	6.43	5.64	6.22
P17	6	5	5	5.33	6.14	5.83	5.42	5.83
P18	5	5	6	5.33	6.00	5.31	5.42	5.59
P19	8	7	7	7.33	7.01	6.89	5.58	6.58
P20	9	7	5	7.00	8.24	6.75	5.52	6.96
P21	8	6	5	6.33	8.05	6.45	5.25	6.70
P22	7	6	6	6.33	6.45	5.81	6.54	6.24
P23	7	7	8	7.33	6.87	5.20	5.96	6.01
P24	7	8	5	6.67	8.05	6.76	5.81	6.97
P25	5	7	5	5.67	6.00	6.76	5.00	6.00
P26	8	8	8	8.00	6.55	7.00	5.01	6.29
P27	8	6	7	7.00	6.00	6.70	6.54	6.40
P28	8	6	7	7.00	6.39	5.58	5.47	5.85
P29	7	5	7	6.33	7.64	6.27	6.45	6.82
P30	8	8	6	7.33	6.87	5.90	4.97	6.00
P31	7	6	8	7.00	6.52	6.39	6.54	6.48
P32	7	7	8	7.33	6.60	6.39	6.20	6.42
P33	3	4	3	3.33	4.49	4.77	4.99	4.73
P34	4	5	2	3.67	4.36	4.62	2.64	3.99
P35	3	4	3	3.33	4.66	5.34	3.25	4.52
P36	5	3	2	3.33	4.66	3.46	4.21	4.10
P37	4	2	1	2.33	4.63	2.81	3.01	3.52

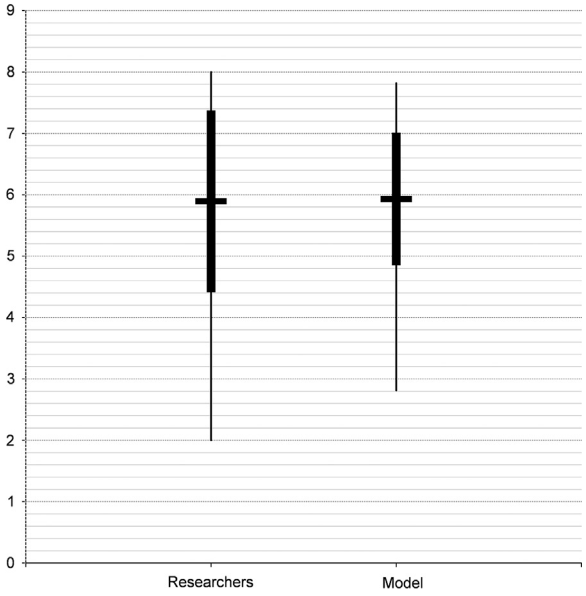


Fig. 3. Boxplot graph for adequacy dimension.

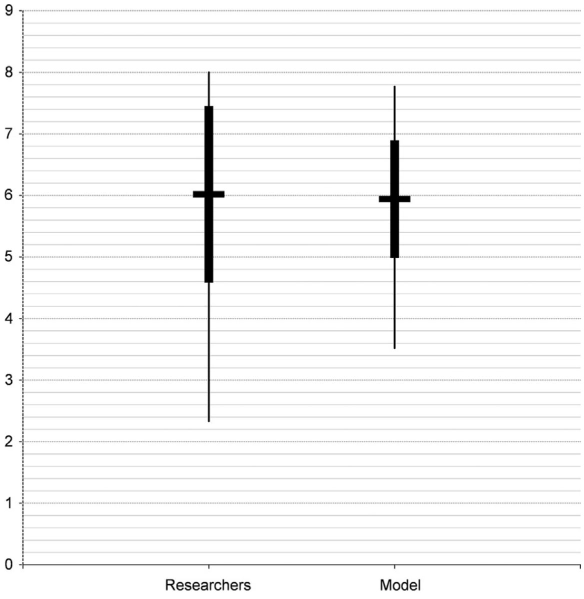


Fig. 5. Boxplot graph for overall project feasibility.

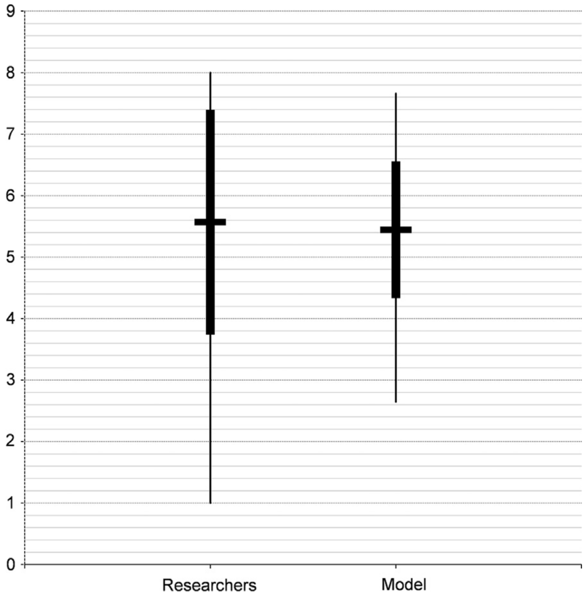


Fig. 4. Boxplot graph for success dimension.

As a result, it is possible to confirm that the feasibility model (Section 4.1) and the effort estimation model (Section 4.2) are reliable to be used within SMEs projects.

4.1. Feasibility model validation

The feasibility model validation utilizes all collected projects. On the one hand, these projects have been characterized by authors using model features and applying the corresponding steps. On the other, a survey has been issued to each researcher to assess one project. The researcher examined the project information (including the plan, meeting notes, status reports, among other things) and indicated a value between 1 and 10 (where 1 is the lowest value and 10 the biggest) to appraise each project dimension (i.e. plausibility, adequacy, success). Then the project feasibility was calculated as the average of them. The obtained values are shown in Table 14. Note that meanwhile the first thirty-two projects (i.e. P1 to P32) finished satisfactory (with some minor problems), the last five projects (i.e. P33 to P37) were canceled before completion.

As soon as the previous values have been collected, Figs. 2, 3, 4, and 5 boxplot graphs have been prepared to show graphically the comparison between the values appraised by the researchers and the values calculated per dimension by the model. These graphs reflect the values behavior assigned by the researchers (on the left of the graph) and the ones calculated by the model (on the right of the graph) indicate the minimum and maximum values (thin line), standard deviation range (thick line) and average value (marker). As seen in the boxplot graphs, the model tends to be more conservative than the appraisal performed by the experts. In general, the model range is shorter than the one assigned (especially for minimum values where the biggest difference is 1.64 for *Success* dimension). But the standard deviation range and average values are almost the same (the biggest difference is lower than 0.30 for *Plausibility* dimension). Thus, from this preliminary analysis it can be said that the model seems to be correct.

To confirm this preliminary analysis results, a more detailed assessment is performed to the model by applying the Wilcoxon signed-rank test [40]. This non-parametric statistical test allows to compare two related samples and define whether their population means differ (i.e. it is a paired difference test). It is an alternative to the paired Student's *t*-test when the population cannot be assumed to be distributed normally but there is a symmetric distribution of the differences around the median. In this test, each project dimension is handled independently. This means that for each dimension, the values provided by the

Table 15
Wilcoxon test results for the feasibility model

Dimension	Sum ranks$^+$ (W^+)	Sum ranks$^-$ (W^-)	Quantity of non-zero pairs (critical value)
Plausibility	*244*	*459*	*37*(*182*)
Adequacy	*393*	*310*	*37*(*182*)
Success	*284*	*382*	*36*(*171*)
Overall Feasibility	*323*	*380*	*37*(*182*)

researchers are tested against the calculated by the proposed model. The used null and alternative hypotheses are:

> *H_0:* the values assigned by the researchers and the values calculated by the model for each dimension have a median difference of zero (in other words, there are no meaningful differences between the researchers and the model values and they can be considered equivalent)
>
> *H_1:* the median difference is not zero (i.e. the researchers and the model values are not equivalent)

The null hypothesis (H0) is accepted or rejected based on the comparison between the minimum sum of ranks (W) and a critical value extracted from the statistical reference table corresponding to quantity of non-zero pairs and a significance level. If W is lower than or equal to the critical value then the null hypotheses can be rejected, meaning that the model is not equivalent to the researchers' assessment. Otherwise, the null hypotheses can be considered as valid (and, in this case, the model can be considered equivalent).

The sums of signed-ranks generated by Wilcoxon test application are shown in Table 15 for each dimension (where W^+ is the sum of all positive ranks and W^- is the sum of all negative ranks). As one dimension (*Success*) has one zero-value pair, the quantity of pairs and the corresponding critical value are also indicated in the table (using in all cases 0.01 significance level).

Based on Table 15 values, the null hypothesis is checked per dimension as follow:

- For *Plausibility*, minimum sum of ranks (*W*) is equal to 244 because W^+ is lower than W^-. As 244 is bigger than 182, the null hypothesis is not rejected. We can conclude that there are no meaningful differences between the researchers and the model plausibility values, they can be considered equivalent.
- For *Adequacy*, minimum value is $W^-=310$, is also higher than 182. This means that H_0 is not rejected and the model adequacy values are also valid.
- For *Success*: $W=W^+=284$ is higher than 171 critical value. This means that success values are also significant.
- Finally, *Project Overall Feasibility* values calculated by the model value can also be considered equivalent because $W=W^+=323>182$.

Therefore, it is confirmed that the proposed model has calculated values equivalent to the experts' appraisal.

4.2. *Effort estimation model validation*

The effort estimation model validation compares the project real effort with the estimation values calculated by the proposed model. Thus, only successful collected projects are used (i.e. P1 to P32). Table 16 shows the real effort for each project with DMCoMo estimations (MM23 formula) and the ones for the proposed model (linear formula and empiric method). Moreover, the absolute and relative errors are presented.

As shown from the obtained results, the estimations calculated by the proposed model methods are more accurate than DMCoMo model values [5]. For this model, only MM23 formula results are presented because they have a smaller error than MM8 formula. Nevertheless, the estimated efforts by DMCoMo are always greater than the real one, producing a large overestimation: the smallest error is almost 24 man-months (i.e. almost 2 man-years) for project P16.

On the other hand, in the proposed method the linear formula generates a smaller error than the empirical method for the analyzed projects. While the average absolute error for the first method is 0.89 man-months (with an error deviation of ±1.53 man-months), the second has an average error of about 1.52 man-months (with a deviation of ±2.21 man-months). Using the collected values, the boxplot graphs show in Figs. 6 and 7 the comparison between the real project effort and the estimated effort calculated by the model. For both methods, the real effort behavior is very similar to the estimated one. We observe that the empiric method tends to underestimate the effort. Nevertheless, the proposed model can be considered correct.

To finalize the estimation method validation, Wilcoxon signed-rank test is also applied. In this case, each method is handled independently. This means that first the linear formula is tested against the real effort provided by the researchers, and then the same operation is performed for the empiric method. Null and alternative hypotheses used in these tests:

> *H_0:* project real efforts and efforts calculated by the method have a median difference of zero (in other words, there are no meaningful differences between the real efforts and the estimated efforts, they can be considered equivalent).
>
> *H_1:* the mean difference is not zero (i.e. the real efforts and the estimation efforts are not equivalent).

The sums of signed-ranks generated by Wilcoxon test application for each method are shown in Table 17. In both cases there are 32 non-zero pairs, then a critical value of 128 is used, which has 0.01 significance level.

Based on Table 17 values, the null hypothesis is checked by the following method:

- For the *Linear Formula*, the minimum sum of ranks (*W*) is equal to 262 because W^+ is lower than W^-. As 262 is bigger than 128, the null hypothesis is not rejected, concluding that there are no meaningful differences

Table 16
Projects data used in the effort estimation model validation.

#	Real Effort (RE)	DMCoMo			Proposed model for SMEs					
		MM23 formula			Linear formula (PEM_L)			Empiric method (PEM_E)		
		Calculated effort (MM23)	Error RE – MM23	Relative error RE – MM23/RE (%)	Calculated effort (PEM_L)	Error RE – PEM_L	Relative error RE – PEM_L/RE (%)	Calculated effort (PEM_E)	Error RE – PEM_E	Relative error RE – PEM_E/RE (%)
P1	*2.41*	*94.88*	−92.47	−3837	*2.58*	−0.17	−7.1	*3.57*	−1.16	−48.1
P2	*7.00*	*51.84*	−44.84	−641	*6.00*	1.00	14.3	*7.23*	−0.23	−3.3
P3	*1.64*	*68.07*	−66.43	−4051	*1.48*	0.16	9.8	*3.58*	−1.94	−118.3
P4	*3.65*	*111.47*	−107.82	−2954	*1.68*	1.97	54.0	*3.57*	0.08	2.2
P5	*9.35*	*122.52*	−113.17	−1210	*9.80*	−0.45	−4.8	*7.58*	1.77	18.9
P6	*11.63*	*81.36*	−69.73	−600	*5.10*	6.53	56.1	*6.07*	5.56	47.8
P7	*6.73*	*92.49*	−85.76	−1274	*3.78*	2.95	43.8	*5.91*	0.82	12.2
P8	*5.40*	*89.68*	−84.28	−1561	*4.88*	0.52	9.6	*3.06*	2.34	43.3
P9	*8.38*	*98.74*	−90.36	−1078	*8.70*	−0.32	−3.8	*7.66*	0.72	8.5
P10	*1.56*	*103.13*	−101.57	−6511	*1.08*	0.48	30.8	*1.50*	0.06	4.1
P11	*9.70*	*77.03*	−67.33	−694	*9.60*	0.10	1.0	*12.64*	−2.94	−30.3
P12	*5.24*	*85.74*	−80.50	−1536	*5.80*	−0.56	−10.7	*5.63*	−0.39	−7.4
P13	*5.00*	*93.08*	−88.08	−1762	*4.58*	0.42	8.4	*3.17*	1.84	36.7
P14	*8.97*	*78.20*	−69.23	−772	*9.18*	−0.21	−2.3	*5.91*	3.06	34.1
P15	*2.81*	*93.57*	−90.76	−3230	*3.48*	−0.67	−23.8	*1.11*	1.70	60.4
P16	*11.80*	*35.59*	−23.79	−202	*12.00*	−0.20	−1.7	*7.58*	4.22	35.7
P17	*2.79*	*91.12*	−88.33	−3166	*2.28*	0.51	18.3	*8.44*	−5.65	−202.5
P18	*3.88*	*60.66*	−56.78	−1464	*3.58*	0.30	7.7	*3.57*	0.31	8.0
P19	*5.70*	*69.90*	−64.20	−1126	*6.30*	−0.60	−10.5	*10.11*	−4.41	−77.4
P20	*8.54*	*81.81*	−73.27	−858	*9.18*	−0.64	−7.5	*8.44*	0.10	1.2
P21	*10.61*	*99.45*	−88.84	−837	*11.50*	−0.89	−8.4	*7.33*	3.28	30.9
P22	*6.88*	*130.73*	−123.85	−1800	*6.40*	0.48	7.0	*6.71*	0.17	2.5
P23	*11.20*	*86.93*	−75.73	−676	*9.70*	1.50	13.4	*10.11*	1.09	9.7
P24	*9.70*	*92.03*	−82.33	−849	*12.70*	−3.00	−30.9	*10.93*	−1.23	−12.7
P25	*7.30*	*111.05*	−103.75	−1421	*8.38*	−1.08	−14.8	*6.51*	0.80	10.9
P26	*5.31*	*117.39*	−112.08	−2111	*5.10*	0.21	4.0	*5.63*	−0.32	−6.0
P27	*6.10*	*66.08*	−59.98	−983	*6.70*	−0.60	−9.8	*5.63*	0.47	7.7
P28	*10.00*	*78.27*	−68.27	−683	*9.60*	0.40	4.0	*9.39*	0.61	6.1
P29	*6.43*	*83.52*	−77.09	−1199	*7.12*	−0.69	−10.7	*5.91*	0.52	8.1
P30	*9.80*	*101.39*	−91.59	−935	*10.20*	−0.40	−4.1	*10.11*	−0.31	−3.2
P31	*1.50*	*114.72*	−113.22	−7548	*1.68*	−0.18	−12.0	*1.11*	0.39	25.8
P32	*3.78*	*92.47*	−88.69	−2346	*3.42*	0.36	9.5	*4.04*	−0.26	−6.8

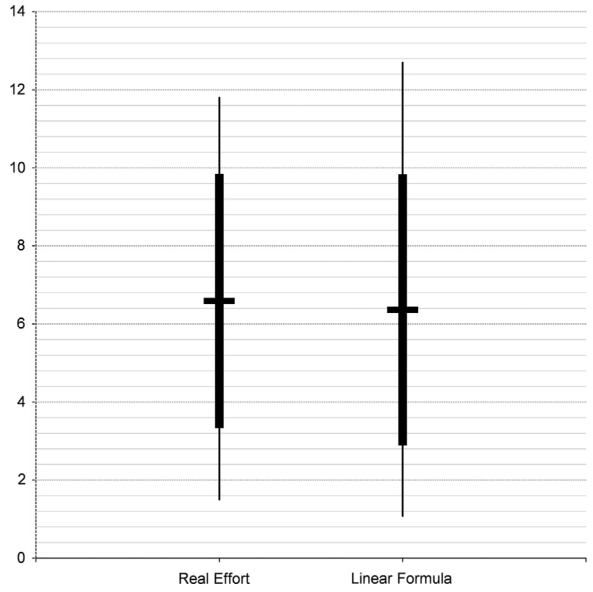


Fig. 6. Boxplot graph for linear formula (PEM_L).

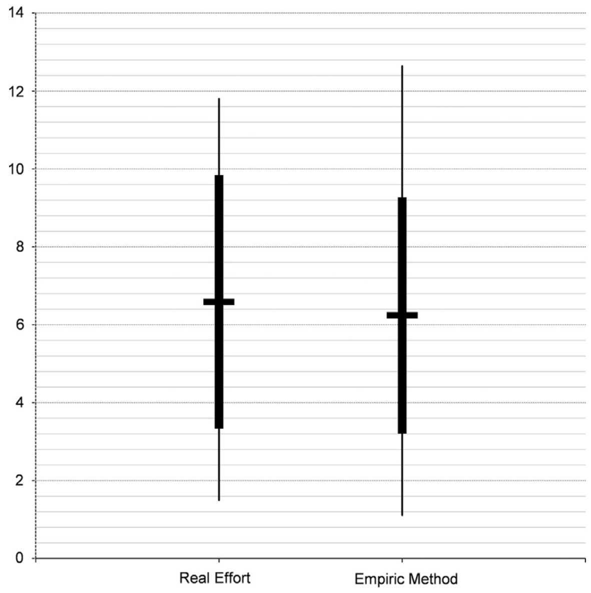


Fig. 7. Boxplot graph for the empiric method (PEM_E).

Table 17
Wilcoxon results for the effort estimation model.

Method	Sum ranks$^+$ (W^+)	Sum ranks $^-$ (W^-)	Quantity of non-zero pairs (critical value)
Linear formula	*266*	*262*	*32*(*128*)
Empiric method	*190*	*338*	*32*(*128*)

between real efforts and the ones calculated by this formula.

- For the *Empiric Method*: $W=W^+=190$ which is bigger than the critical value of 128. This means that the empiric method estimations are also significant.

Therefore, it is confirmed that both proposed model methods can be considered equivalent to estimate the project effort.

5. Conclusions

The term “data mining” is strongly linked to the database concept and goes back to the definition of pattern discovery algorithms on large databases. However, today there are lines of research in fields such as: text mining, image mining, data stream mining, web mining, among others. In this context, authors think that it is more appropriate to use the term “information mining” as a generic one to any of the aforementioned mining types. Then, information mining is a sub-discipline of information systems which provides business intelligence with the non-trivial knowledge needed for making decisions inside an organization. This knowledge is (implicitly) located in the data available from several information sources. Although such projects have different features, they share some of the problems of traditional software engineering and knowledge engineering projects. Most of the projects are not completed successfully, most of them ending in failure.

Among the reasons that produce project failure, two are highlighted: unmanaged risks and needed resources inaccurate estimations. In order to handle these problems, two ad-hoc models have been proposed to be used at early stages of information mining projects.

The early risk detection could reduce the associated effects during the project development. Then the first model's goal is to analyze the project feasibility. This means that, based on the values of 13 features that characterize a project, the model allows to calculate if the project can be performed (i.e., its plausibility), if information mining is an appropriate solution for the identified business problem (i.e., adequacy) and if the project accomplishment can be achieved (i.e., success). Supposing that it is difficult characterizing the project features with answers “yes/no” or “numerical values”, the model uses five linguistic labels for qualifying the features.

We observe that the resulting appraisal by using the proposed model has a similar behavior to the researchers' appraisal for the considered projects. Although the model is usually a bit more conservative to define low values, the general behavior fits. It could be said that researchers have been more critical, having a negative view when assigning each dimension values. This is confirmed by the results obtained by Wilcoxon tests. With 99% confidence level, we conclude that there is no significant difference between the value calculated by the model for all dimensions, and the valuation assigned by researchers. In general, the greater symmetry is detected for the Overall Project Feasibility dimension which is the final result of the model (calculated based on other three dimensions). In some cases, although the dimension valuation result is not equal to the real one, when calculating the overall feasibility by the model, the differences are compensated to obtain a more accurate final result. Therefore, we conclude that the proposed model for the feasibility assessment is valid to be used in information mining projects within SMEs.

The second proposed model allows estimating the resources and time required to perform the project based on the values of 8 project features (also known as cost drivers) and two methods (a linear formula and an empiric method). This model is oriented to estimate small projects which are normally developed by SMEs. From its validation results, it is observed that in general the estimation model has a behavior similar to the real considered projects. The linear estimation formula generates a more accurate result than the empirical method (the relative error for this last method is a little higher). Anyway, when comparing the results of both methods with DMCoMo model, it can be seen that the proposed methods are more accurate for estimating than DMCoMo, which is oriented to large projects. By using the Wilcoxon test, it has been confirmed with 99% confidence level, that there is no significant difference between the calculated effort and the actual effort required to develop the project. Again, the greater symmetry stands for the linear estimation while the empirical method is lower. Therefore, it is concluded that the proposed model for effort estimation is valid to be used in information mining projects within SMEs.

Acknowledgment

The research reported in this paper has been partially funded by Research Projects 33A105, 33B102 and 33A167 within Universidad Nacional de Lanús, Research Project 40B133 within Universidad Nacional de Río Negro, and Research Project UTI1867 within UTN in Buenos Aires.

References

[1] E.. Thomsen, BI's Promised Land, Intell. Enterp. 6 (4) (2003) 21–25.
[2] J.. Rowley, The wisdom hierarchy: representations of the DIKW hierarchy, J. Inf. Sci. 33.2 (2007) 163–180.
[3] S. Negash, P. Gray, Business intelligence, in: F. Burstein, C. Holsapple (Eds.), Handbook on Decision Support Systems, vol. 2, Springer, Heidelberg, 2008, pp. 175–193.
[4] R. García-Martínez, P. Britos, P. Pesado, R. Bertone, F. Pollo-Cattaneo, D. Rodríguez, P. Pytel, J. Vanrell, Towards an information mining engineering, in: Software Engineering, Methods, Modeling and Teaching, University of Medellín Press, Medellin (Colombia), ISBN 978-958-8692-32-6, 2011, 83–99.
[5] O. Marbán, E. Menasalvas, C. Fernández-Baizán, A cost model to estimate the effort of data mining projects (DMCoMo), Inf. Syst. 33 (2008) 133–150.

[6] P. Chapman, J. Clinton, R. Keber, T. Khabaza, T. Reinartz, C. Shearer, R. Wirth CRISP-DM 1.0 Step by step BI guide. Edited by SPSS, 2000 ⟨http://tinyurl.com/crispdm⟩.

[7] SAS Enterprise Miner: "SEMMA", 2008 ⟨http://tinyurl.com/semmaSAS⟩.

[8] D. Pyle, Business Modeling and Business Intelligence, Morgan Kaufmann, 2003.

[9] H.A. Edelstein, H.C.. Edelstein, Building, Using, and Managing the Data Warehouse, Data Warehousing Institute. Prentice-Hall PTR, , 1997.

[10] M. Strand, The Business Value of Data Warehouses – Opportunities, Pitfalls and Future Directions (Ph.D. thesis), Department of Computer Science, University of Skovde, 2000.

[11] U.M. Fayyad, "Tutorial report". Summer school of DM, Monash University, Australia, 2000.

[12] O. Marbán, G. Mariscal, J. Segovia, A data mining & knowledge discovery process model. Data Mining and Knowledge Discovery in Real Life Applications, IN-TECH, 2009, p. 8. ⟨http://cdn.intechopen.com/pdfs/5937/InTech-A_data_mining_amp_knowledge_discovery_process_model.pdf⟩.

[13] L.J. May, Major causes of software project failures, CrossTalk: J. Def. Softw. Eng. 11 (6) (2009) 9–121998 11 (2009) 9–12.

[14] R.N. Charette, Why software fails, IEEE Spectr. 42 (9) (2005) 42–49.

[15] Standish Group. "CHAOS Summary Report 2009" ⟨http://kinzz.com/resources/articles/91-project-failures-rise-study-shows⟩.

[16] K. Wiegers, Software Requirements, Microsoft Press, Sebastopol, California, USA, 2003.

[17] P. Britos, O. Dieste, R. García-Martínez, Requirements elicitation in data mining for business intelligence projects, in: David Avison, George M. Kasper, Barbara Pernici, Isabel Ramos, Dewald Roode (Eds.), Advances in Information Systems Research, Education and Practice, IFIP Series, 274, Springer, Boston, 2008, pp. 139–150.

[18] R. Pressman, Software Engineering: A Practitioner's Approach, Mc Graw Hill, New York, USA, 2004.

[19] B. Boehm, C. Abts, A. Brown, S. Chulani, B. Clark, E. Horowitz, R. Madachy, D. Reifer, B. Steece, Software Cost Estimation with COCOMO II, Prentice-Hall, Englewood Cliffs, 2000.

[20] Organization for Economic Cooperation and Development, OECD SME and Entrepreneurship Outlook 2005, OECD Publishinghttp://dx.doi.org/10.1787/9789264009257-en.

[21] Organization for Economic Cooperation and Development/Economic Commission for Latin America and the Caribbean/Inter-American Center of Tax Administrations, Revenue Statistics in Latin America, OECD Publishinghttp://dx.doi.org/10.1787/9789264183889-en-fr.

[22] International Organization for Standardization, ISO/IEC DTR 29110-1 Software Engineering – Lifecycle Profiles for Very Small Entities (VSEs) – Part 1: Overview, International Organization for Standardization (ISO), Geneva, Switzerland, 2011.

[23] C. Laporte, S. Alexandre, A. Renault, Developing international standards for VSEs, IEEE Comput. 41 (3) (2008) 98–101.

[24] P. Pytel, P. Britos, R. García-Martínez Proposal and validation of a feasibility model for information mining projects, in: Proceedings of 25th International Conference on Software Engineering and Knowledge Engineering, 2013, ISBN 978-1-891706-33-2, pp. 83–88.

[25] L.L. Pipino, Y.W. Lee, R.Y. Wang, Data quality assessment, Commun. ACM 45 (4) (2002) 211–218.

[26] H.R. Nemati, C.D. Barko, Key factors for achieving organizational data-mining success, Ind. Manag. Data Syst. 103 (4) (2003) 282–292. (doi: 10.1108/02635570310470692.).

[27] J. Sim, Critical Success Factors in Data Mining Projects (Ph.D. thesis), University of North Texas, 2003.

[28] T.H. Davenport, Make better decisions, Harv. Bus. Rev. 87 (11) (2009) 117–123.

[29] G. Nie, L. Zhang, Y. Liu, X. Zheng, Y. Shi, Decision analysis of data mining project based on Bayesian risk, Expert Syst. Appl. 36 (3) (2009) 4589–4594.

[30] U. Bolea, J. Jakličb, G. Papac, J. Žabkard, Critical Success Factors of Data Mining in Organizations, Ljubljana (Slovenia), 2011.

[31] A. Nothingli, E.N. Kakhky, H.E. Nosratabadi, Evaluating the success level of data mining projects based on CRISP-DM methodology by a Fuzzy expert system, in: Proceedings of the 3rd IEEE International Conference on Electronics Computer Technology (ICECT), vol. 6, Kanyakumari, 2011, pp. 161–165. http://dx.doi.org/10.1109/ICECTECH.2011.5942073.

[32] F. Alonso, N. Juristo, J. Pazos, Trends in life-cycle models for SE and KE: proposal for a spiral-conical life-cycle approach, Int. J. Softw. Eng. Knowl. Eng. 5 (03) (1995) 445–465.

[33] M.H. Kalos, P.A. Whitlock, Monte Carlo Methods. Vol I Basics, John Wiley & Sons, New York, 1986.

[34] J.S. R. Jang, Fuzzy Inference Systems, Prentice-Hall, Upper Saddle River, NJ, 1997.

[35] P. Pytel, P. Britos, R. García-Martínez, A proposal of effort estimation method for information mining projects oriented to SMEs, Lect. Notes Bus. Inf. Process. 139 (2013) 58–74.

[36] Z. Chen, T. Menzies, D. Port, et al., Finding the right data for software cost modeling, IEEE Softw. 22 (6) (2005) 38–46. (Online (04/12).

[37] P. Domingos, C. Elkan, J. Gehrke, J. Han, D. Heckerman, D. Keim, et al., 10 challenging problems in data mining research, Int. J. Inf. Technol. Decis. Mak. 5 (4) (2006) 597–604.

[38] R. Garcia-Martinez, P. Britos, F. Pollo-Cattaneo, D. Rodriguez, P. Pytel, Information mining processes based on intelligent systems, in: Proceedings of II International Congress on Computer Science and Informatics (INFONOR-CHILE), 2011, pp. 87-94. ISBN 978-956-7701-03-2.

[39] S. Weisberg, Applied Linear Regression, John Wiley & Sons, New York, 1985.

[40] F. Wilcoxon, Individual comparisons by ranking methods, Biometrics 1 (1945) 80–83.