A Proposal of Effort Estimation Method for Information Mining Projects Oriented to SMEs

Pablo Pytel[1,2,3], Paola Britos[4], and Ramón Garcia-Martinez[2]

[1] PhD Program on Computer Science, Computer Science School,
National University of La Plata, Buenos Aires, Argentina
[2] Information Systems Research Group, National University of Lanus,
Buenos Aires, Argentina
[3] Information System Methodologies Research Group,
Technological National University at Buenos Aires, Argentina
[4] Information Mining Research Group, National University of Rio Negro
at El Bolson, Río Negro, Argentina
{ppytel,paobritos}@gmail.com, rgarcia@unla.edu.ar

Abstract. Software projects need to predict the cost and effort with its associated quantity of resources at the beginning of every project. Information Mining projects are not an exception to this requirement, particularly when they are required by Small and Medium-sized Enterprises (SMEs). An existing Information Mining projects estimation method is not reliable for small-sized projects because it tends to overestimates the estimated efforts. Therefore, considering the characteristics of these projects developed with the CRISP-DM methodology, an estimation method oriented to SMEs is proposed in this paper. First, the main features of SMEs' projects are described and applied as cost drivers of the new method with the corresponding formula. Then this is validated by comparing its results to the existing estimation method using SMEs real projects. As a result, it can be seen that the proposed method produces a more accurate estimation than the existing estimation method for small-sized projects.

Keywords: Effort Estimation method, Information Mining, Small and Medium-sized Enterprises, Project Planning, Software Engineering.

1 Introduction

Information Mining consists in the extraction of non-trivial knowledge which is located (implicitly) in the available data from different information sources [1]. That knowledge is previously unknown and it can be useful for some decision making process [2]. Normally, for an expert, the data itself is not the most relevant but it is the knowledge included in their relations, fluctuations and dependencies. Information Mining Process can be defined as a set of logically related tasks that are executed to achieve [3], from a set of information with a degree of value to the organization, another set of information with a greater degree of value than the initial one [4]. Once the problem and the customer's necessities are identified, the Information Mining Engineer selects the Information Mining Processes to be executed. Each Information

G. Poels (Ed.): CONFENIS 2012, LNBIP 139, pp. 58–74, 2013.

Mining Process has several Data Mining Techniques that may be chosen to carry on the job [5]. Thus, it can be said that, Data Mining is associated to the technology (i.e. algorithms from the Machine Learning's field) while Information Mining is related to the processes and methodologies to complete the project successfully. In other words, while Data Mining is more related to the development tasks, Information Mining is closer to Software Engineering activities [6]. However, not all the models and methodologies available in Software Engineering can be applied to Information Mining projects because they do not handle the same practical aspects [7]. Therefore, specific models, methodologies, techniques and tools need to be created and validated in order to aid the Information Mining practitioners to carry on a project.

As in every Software project, Information Mining projects begin with a set of activities that are referred as project planning. This requires the prediction of the effort with the necessary resources and associated cost. Nevertheless, the normal effort estimation method applied in Conventional Software Development projects cannot be used at Information Mining projects because the considered characteristics are different. For example COCOMO II [8], one of the most used estimation method for Conventional Software projects, uses the quantity of source code lines as a parameter. This is not useful for estimating an Information Mining project because the data mining algorithms are already available in commercial tools and then it is not necessary to develop software. Estimation methods in Information Mining projects should use more representative characteristics, such as, the quantity of data sources, the level of integration within the data and the type of problem to be solved. In that respect, only one specific analytical estimation method for Information Mining projects has been found after a documentary research. This method called Data Mining Cost Model (or DMCoMo) is defined in [9]. However, from a statistical analysis of DMCoMo performed in [10], it has been found that this method tends to overestimate the efforts principally in small-sized projects that are usually required by Small and Medium-sized Enterprises [11].

In this context, the objective of this paper is proposing a new effort estimation method for Information Mining projects considering the features of Small and Medium-sized Enterprises (SMEs). First, the estimation method DMCoMo is described (section 2), and the main characteristics of SMEs' projects are identified (section 3). Then an estimation method oriented to SMEs is proposed (section 4) comparing its results to DMCoMo method using real projects data (section 5). Finally, the main conclusions and future research work are presented (section 6).

2 DMCoMo Estimation Method

Analytical estimation methods (such as COCOMO) are constructed based on the application of regression methods in the available historical data to obtain mathematical relationships between the variables (also called cost drivers) that are formalized through mathematical formulas which are used to calculate the estimated effort. DMCoMo [9] defines a set of 23 cost drivers to perform the cost estimation which are associated to the main characteristics of Information Mining projects. These cost drivers are classified in six categories which are included in table 1 as specified in [9]. Once the values of the cost drivers are defined, they are introduced in the mathematical formulas provided by the method. DMCoMo has two formulas which have been defined by linear regression with the information of 40 real projects of different business types (such as marketing, meteorological projects and medical projects). The first formula uses all 23 cost drivers as variables (formula named MM23) and it should be used when the project is well defined; while the second formula only uses 8 cost drivers (MM8) and it should be used when the project is partially defined. As a result of introducing the values in the corresponding formula, the quantity of men x month (MM) is calculated.

Table 1. Cost Drivers used by DMCoMo

Category	Cost Drivers
Source Data	– Number of Tables (NTAB) – Number of Tuples (NTUP) – Number of Table Attributes (NATR) – Data Dispersion (DISP) – Nulls Percentage (PNUL) – Data Model Availability (DMOD) – External Data Level (DEXT)
Data Mining Models	– Number of Data Models (NMOD) – Types of Data Model (TMOD) – Number of Tuples for each Data Models (MTUP) – Number and Type of Attributes for each Data Model (MATR) – Techniques Availability for each Data Model (MTEC)
Development Platform	– Number and Type of Data Sources (NFUN) – Distance and Communication Form (SCOM)
Techniques and Tools	– Tools Availability (TOOL) – Compatibility Level between Tools and Other Software (COMP) – Training Level of Tool Users (NFOR)
Project	– Number of Involved Departments (NDEP) – Documentation (DOCU) – Multisite Development (SITE)
Project Staff	– Problem Type Familiarity (MFAM) – Data Knowledge (KDAT) – Directive Attitude (ADIR)

But, as it has been pointed out by the authors, the behaviour of DMCoMo in projects outside of the 90 and 185 men x month range is unknown. From a statistical analysis of its behaviour performed in [10], DMCoMo always tends to overestimates the estimated efforts (i.e. all project estimations are always bigger than 60 men x month). Therefore, DMCoMo could be used in medium and big-sized projects but it is not useful for small-sized projects. As these are the projects normally required by Small and Medium-sized Enterprises, a new estimation method for Information Mining projects is proposed considering the characteristics of small-sized projects.

3 SMEs' Information Mining Projects

According to the Organization for Economic Cooperation and Development (OECD) Small and Medium-sized Enterprises (SMEs) and Entrepreneurship Outlook report [12]: "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 country". However, although the importance of SMEs is well known, there is no universal criterion to characterise them. Depending on the country and region, there are different quantitative and qualitative parameters used to recognize a company as SMEs. For instance, at Latin America each country has a different definition [13]: while Argentina considers as SME all independent companies that have an annual turnover lower than USD 20,000 (U.S. dollars maximum amount 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, International Organization for Standardization (ISO) has recognized the necessity to specify a software engineering standard for SMEs and thus it is working in the ISO/IEC 29110 standard "Lifecycle profiles for Very Small Entities" [14]. The term 'Very Small Entity' (VSE) was defined by the ISO/IEC JTC1/SC7 Working Group 24 [15] as being "an entity (enterprise, organization, department or project) having up to 25 people".

From these definitions (and our experience), in this paper an Information Mining project for SMEs is demarcated as a project performed at a company of 250 employees or less (at one or several locations) where the high-level managers (usually the company's owners) need non-trivial knowledge extracted from the available databases to solve a specific business problem with no special risks at play. As the company's employees usually do not have the necessary experience, the project is performed by contracted outsourced consultants. From our experience, the project team can be restricted up to 25 people (including both the outsourced consultants and the involved company staff) with maximum project duration of one year.

The Information Mining project's initial tasks are similar to a Conventional Software Development project. The consultants need to elicit both the necessities and desires of the stakeholders, and also the characteristics of the available data sources within the organization (i.e. existing data repositories). Although, the outsourced consultants must have a minimum knowledge and experience in developing Information Mining projects, they might or not have experience in similar projects on the same business type which could facilitate the tasks of understanding the organization and its related data. As the data repositories are not often properly documented, the organization's experts should be interviewed. However, experts are normally scarce and reluctant to get involved in the elicitation sessions. Thus, it is required the willingness of the personnel and the supervisors to identify the correct characteristics of the organization and the data repositories. As the project duration is quite short and the structure of the organization is centralized, it is considered that the elicited requirements will not change.

On the other hand, the Information and Communication Technology (ICT) infrastructure of SMEs is analysed. In [16] it is indicated that more than 70% of Latin

America's SMEs have an ICT infrastructure, but only 37% have automated services and/or proprietary software. Normally commercial off-the-shelf software is used (such as spread-sheets managers and document editors) to register the management and operational information. The data repositories are not large (from our experience, less than one million records) but implemented in different formats and technologies. Therefore, data formatting, data cleaning and data integration tasks will have a considerable effort if there is no available software tools to perform them because ad-hoc software should be developed to implement these tasks.

4 Proposed Effort Estimation Method Oriented to SMEs

For specifying the effort estimation method oriented to SMEs, first, the cost drivers used to characterize a SMEs' project are defined (section 4.1) and then the corresponding formula is presented (section 4.2). This formula has been obtained by regression using real projects information. From 44 real information mining projects available, 77% has been used for obtaining the proposed method's formula (section 4.2) and 23% for validation of the proposed method (section 5). This means that 34 real projects have been used for obtaining the formula and 10 projects for validation.

These real Information Mining projects have been collected by researchers from the Information Systems Research Group of the National University of Lanus (GISI-DDPyT-UNLa), the Information System Methodologies Research Group of the Technological National University at Buenos Aires (GEMIS-FRBA-UTN), and the Information Mining Research Group of the National University of Rio Negro at El Bolson (SAEB-UNRN). It should be noted that all these projects had been performed applying the CRISP-DM methodology [17]. Therefore, the proposed estimation method can be considered reliable only for Information Mining projects developed with this methodology.

4.1 Cost Drivers

Considering the characteristics of Information Mining projects for SMEs indicated in section 3, eight cost drivers are specified. Few cost drivers have been identified in this version because, as explained in [18], when an effort estimation method is created, many of the non-significant data should be ignored. As a result the model is prevented from being too complex (and therefore impractical), the irrelevant and co-dependent variables are removed, and the noise is also reduced. The cost drivers have been selected based on the most critical tasks of CRISP-DM methodology [17]: in [19] it is indicated that building the data mining models and finding patterns is quite simple now, but 90% of the effort is included in the data pre-processing (i.e. "Data Preparation" tasks performed at phase III of CRISP-DM). From our experience, the other critical tasks are related to "Business Understanding" phase (i.e. *"understanding of the business' background"* and *"identifying the project success"* tasks). The proposed cost factors are grouped into three groups as follows:

4.1.1 Cost Drivers Related to the Project

- *Information Mining objective type (OBTY)*
  This cost driver analyses the objective of the Information Mining project and therefore the type of process to be applied based on the definition performed in [5]. The allowed values for this cost drivers are indicated in table 2.

Table 2. Values of OBTY cost driver

Value	Description
1	It is desired to identify the rules that characterize the behaviour or the description of an already known class.
2	It is desired to identify a partition of the available data without having a previously known classification.
3	It is desired to identify the rules that characterize the data partitions without a previous known classification.
4	It is desired to identify the attributes that have a greater frequency of incidence over the behaviour or the description of an already known class.
5	It is desired to identify the attributes that have a greater frequency of incidence over a previously unknown class.

- *Level of collaboration from the organization (LECO)*
  The level of collaboration from the members of the organization is analysed by reviewing if the high-level management (i.e. usually the SME's owners), the middle-level management (supervisors and department's heads) and the operational personnel are willing to help the consultants to understand the business and the related data (specially in the first phases of the project). If the Information Mining project has been contracted, it is assumed that at least the high-level management should support it. The possible values for this cost factor are shown in table 3.

Table 3. Values of LECO cost drivers

Value	Description
1	Both managers and the organization's personnel are willing to collaborate on the project.
2	Only the managers are willing to collaborate on the project while the rest of the company's personnel is indifferent to the project.
3	Only the high-level managers are willing to collaborate on the project while the middle-level manager and the rest of the company's personnel is indifferent to the project.
4	Only the high-level managers are willing to collaborate on the project while the middle-level manager is not willing to collaborate.

4.1.2 Cost Drivers Related to the Available Data

- *Quantity and type of the available data repositories (AREP)*
  The data repositories to be used in the Information Mining process are analysed (including data base management systems, spread-sheets and documents among others). In this case, both the quantity of data repositories (public or private from

the company) and the implementation technology are studied. In this stage, it is not necessary to know the quantity 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 into a common format, then the repositories can be considered as compatible for integration because the data integration tasks will be performed using the exported data.
- On the other hand, if there are non-digital repositories (i.e. written paper), then the technology should not be considered compatible for the integration. But the estimation method is not able to predict the required time to perform the digitalization because it could vary on many factors (such as quantity of papers, length, format and diversity among others).

The possible values for this cost factor are shown in table 4.

Table 4. Values of AREP cost driver

Value	Description
1	Only 1 available data repository.
2	Between 2 and 5 data repositories compatible technology for integration.
3	Between 2 and 5 data repositories non-compatible technology for integration.
4	More than 5 data repositories compatible technology for integration.
5	More than 5 data repositories no-compatible technology for integration.

- *Total quantity of available tuples in 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. The possible values for this cost factor are shown in table 5.

Table 5. Values of QTUM cost driver

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

- *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 additional information to the main table (such as a table used to determine the product characteristics associated to the product ID of the sales main table). Normally, these auxiliary tables include fewer records than the main table. The possible values for this cost factor are shown in table 6.

Table 6. Values of QTUA cost driver

Value	Description
1	No auxiliary tables used.
2	Up to 1,000 tuples from auxiliary tables.
3	Between 1,001 and 50,000 tuples from auxiliary tables.
4	More than 50,000 tuples from auxiliary tables.

- *Knowledge level about the data sources (KLDS)*

  The knowledge level about the data sources studies if the data repositories and their tables are properly documented. In other words, if a document exits that defining the technology in which it is implemented, the characteristics of the tables' fields, and how the data is created, modified, and/or deleted.

  When this document is not available, it should be necessary to hold meetings with experts (usually in charge of the data administration and maintenance) to explain them. As a result the project required effort should be increased depending on the collaboration of these experts to help the consultants.

  The possible values for this cost factor are shown in table 7.

Table 7. Values of KLDS cost driver

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

4.1.3 Cost Drivers Related to the available Resources

- *Knowledge and experience level of the information mining team (KEXT)*

  This cost driver studies the ability of the outsourced consultants that will carry out the project. Both the knowledge and experience of the team in similar previous projects are analysed by considering the similarity of the business type, the data to be used and the expected goals. It is assumed that when there is greater similarity, the effort should be lower. Otherwise, the effort should be increased.

  The possible values for this cost factor are shown in table 8.

- *Functionality and usability of available tools (TOOL)*

  This cost driver analyses the characteristics of the information mining tools to be utilized in the project and its implemented functionalities. Both the data preparation functions and the data mining techniques are reviewed.

  The possible values for this cost factor are shown in table 9.

Table 8. Values of KEXT cost driver

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

Table 9. Values of TOOL cost driver

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.

4.2 Estimation Formula

Once the values of the cost drivers have been specified, they were used to characterize 34 information mining projects with their real effort[1] collected by co-researchers as indicated before. A multivariate linear regression method [20] has been applied to obtain a linear equation of the form used by COCOMO family methods [8]. As a result, the following formula is obtained:

$$\begin{aligned} PEM = 0.80\,OBTY + 1.10\,LECO - 1.20\,AREP - 0.30\,QTUM - 0.70\,QTUA \\ + 1.80\,KLDS - 0.90\,KEXT + 1.86\,TOOL - 3.30 \end{aligned} \qquad (1)$$

where *PEM* is the effort estimated by the proposed method for SMEs (in men x month), and the following cost drivers: information mining objective type (*OBTY*), level of collaboration from the organization (*LECO*), quantity and type of the available data repositories (*AREP*), total quantity of available tuples 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 in tables 2 to 9 respectively of section 4.1.

[1] The real projects data used for regression is available at: `http://tinyurl.com/bm93wol`

5 Validation of the Proposed Estimation Method

In order to validate the estimation method defined in section 4, the data of other 10 collected information mining projects is used to compare the accuracy of the proposed method with both the real effort with the effort estimated by DMCoMo method. A brief description of these projects with their applied effort (in men x months) are shown in table 10.

Table 10. Data of the information mining projects used for the validation

#	Business Objectives	Information Mining Objectives	Real Effort (men x month)
P1	The business objective is classifying the different types of cars and reviewing the acceptance of the clients, and detecting the characteristics of the most accepted car.	The process of discovering behaviour rules is used.	2.41
P2	As there is not big increment in the middle segment, the company wants to gain market by attracting new customers. In order to achieve that, it is required to determine the necessities of that niche market.	The process of discovering behaviour rules is used.	7.00
P3	The high management of a company have decided to enhance and expand their market presence by launching a new product. The new concept will be proclaimed as a new production unit which aimed to create more jobs, more sales and therefore more revenue.	The processes of discovering behaviour rules and weighting of attributes are used.	1.64
P4	It is necessary to identify the customer behaviour in order to understand which type of customer is more inclined to buy any package of products. The desired objective is increasing the level of acceptance and sales of product packages.	The process of discovering behaviour rules is used.	3.65
P5	The objectives of the project are performing a personalized marketing campaign to the clients, and locating the ads in the most optimal places (i.e. the places with most CTR).	The process of discovery group-membership rules is used.	9.35
P6	Perform an analysis of the causes why the babies have some deceases when they are born, considering the economic, social and educational level, and also the age of the mother	The processes of discovering behaviour rules and weighting of attributes are used.	11.63
P7	The help desk sector of a governmental organization employs software system to register each received phone call. As a result, it is possible to identify a repairing request, a change or bad function of any computer in order to assign a technical who will solve the problem.	The process of discovering group-membership rules is used.	6.73
P8	The objective is improving the image of the company to the customers by having a better distribution service. This means finding the internal and external factors of the company that affect the delay of the orders to be delivered to customers.	The process of discovering group-membership rules is used.	5.40
P9	The purpose is achieving the best global technologies, the ownership of independent intellectual property rights, and the creation of an internationally famous brand among the world-class global automotive market.	The processes of discovering group-membership rules and weighting of the attributes are used.	8.38
P10	It has been decided to identify the key attributes that produce good quality wines. Once these attributes are detected, they should improve the lesser quality wines.	The processes of discovering behaviour rules and weighting of attributes are used.	1.56

Using the collected project data, the values of the DoCoMo's cost drivers are defined to calculate the corresponding estimation method. Both the formula that uses 8 cost factors (MM8 column) and the formula that uses the 23 cost factors (MM23 column) are applied obtaining the values shown in table 11.

Table 11. Effort calculated by DMCoMo method

#	MFAM	ADIR	KDAT	SITE	DOCU	NDEP	NFOR	COMP	TOOL	SCOM	NFUN	MTEC	MATR	MTUP	TMOD	NMOD	DEXT	DMOD	PNUL	DISP	NATR	NTUP	NTAB	MM8 (men x month)	MM23 (men x month)
P1	3	1	4	3	5	4	3	3	1	1	3	1	1	1	0	1	1	1	0	1	7	1	1	84.23	94.88
P2	5	1	3	1	2	2	3	5	1	1	1	1	2	1	1	2	0	4	1	1	1	1	0	67.16	51.84
P3	5	1	3	1	3	2	3	0	1	1	1	1	2	1	1	2	0	1	1	1	1	1	0	67.16	68.07
P4	3	4	2	0	2	1	3	2	1	0	3	5	3	3	3	3	1	2	2	2	5	5	3	118.99	111.47
P5	1	2	4	2	2	4	1	1	1	1	2	3	3	3	1	3	2	5	1	2	3	3	1	110.92	122.52
P6	5	1	1	1	2	1	1	0	1	1	1	1	2	1	1	2	2	1	2	1	1	1	0	80.27	81.36
P7	3	2	2	3	2	3	3	1	1	1	1	4	2	1	3	2	2	1	2	1	1	1	1	96.02	92.49
P8	0	6	1	0	2	0	2	4	1	1	0	1	3	0	1	1	0	0	1	4	3	0	2	116.87	89.68
P9	4	1	3	3	5	4	3	3	1	1	1	1	2	1	4	2	2	1	1	1	1	1	0	97.63	98.74
P10	4	1	2	3	5	2	1	2	1	1	2	4	2	1	4	2	2	0	1	1	1	1	0	105.32	103.13

Similarly, the same procedure is performed to calculate the effort applying the formula specified in section 3.2 for the proposed estimation method oriented to SMEs (PEM column) as shown in table 12.

Table 12. Effort calculated by the proposed estimation method oriented to SMEs

#	OBTY	LECO	AREP	QTUM	QTUA	KLDS	KEXT	TOOL	PEM (men x month)
P1	1	1	3	3	1	3	2	3	2,58
P2	1	1	1	3	1	3	5	5	6,00
P3	4	1	1	3	3	2	5	3	1,48
P4	1	4	3	5	1	1	2	3	1,68
P5	3	2	2	5	2	3	1	5	9,80
P6	4	1	1	2	1	1	5	5	5,10
P7	3	2	1	4	1	1	2	3	3,78
P8	1	4	1	3	2	1	1	3	4,88
P9	5	1	1	3	3	3	4	5	8,70
P10	4	1	2	2	1	1	4	3	1,08

Finally, in table 13 the estimated efforts are compared with the real effort of each project (REf column) are compared. The efforts calculated by the DMCoMo method (MM8 and MM23 columns) and the proposed method for SMEs (PEM column) are indicating with their corresponding error (i.e. the difference between the real effort and the values calculated by each method). Also, the Relative Error for the estimation of the proposed method is shown (calculated as the error divided by the real effort).

Table 13. Comparison of the calculated efforts (in men x month)

#	REf	DMCoMo				PROPOSED METHOD		
		MM8	REf - MM8	MM23	REf - MM23	PEM	REf - PEM	Relative Error
P1	2.41	84.23	-81.82	94.88	-92.47	2,58	-0.17	-7.2%
P2	7.00	67.16	-60.16	51.84	-44.84	6,00	1.00	14.3%
P3	1.64	67.16	-65.52	68.07	-66.43	1,48	0.16	9.8%
P4	3.65	118.99	-115.34	111.47	-107.82	1,68	1.97	54.0%
P5	9.35	110.92	-101.57	122.52	-113.17	9,80	-0.45	-4.8%
P6	11.63	80.27	-68.65	81.36	-69.73	5,10	6.53	56.1%
P7	6.73	96.02	-89.29	92.49	-85.76	3,78	2.95	43.8%
P8	5.40	116.87	-111.47	89.68	-84.28	4,88	0.52	9.6%
P9	8.38	97.63	-89.26	98.74	-90.36	8,70	-0.33	-3.9%
P10	1.56	105.32	-103.75	103.13	-101.56	1,08	0.48	30.9%

Table 13. *(Continued)*

#	REf	DMCoMo				PROPOSED METHOD		
		MM8	REf - MM8	MM23	REf - MM23	PEM	REf - PEM	Relative Error
Average Error			88.68		85.64		1.46	
Error Variance			380.28		428.99		3.98	

This comparison is reflected in a boxplot graph (figure 1) where the behaviour of the real and calculated efforts are shown by indicating the minimum and maximum values (thin line), standard deviation range (thick line) and average value (marker).

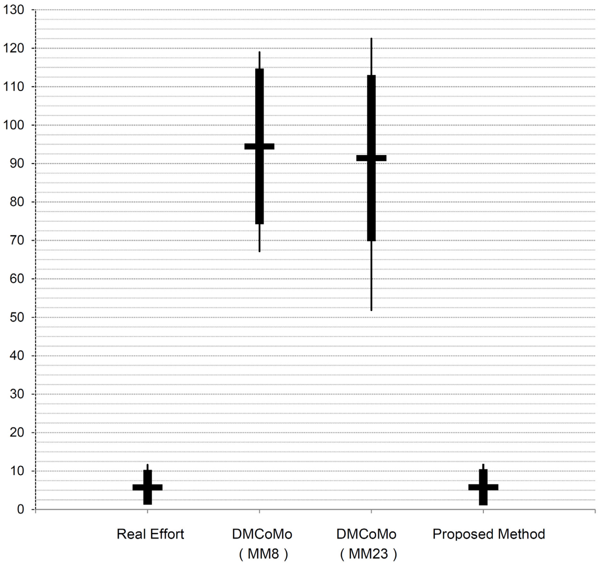

Fig. 1. Boxplot graph comparing the behaviour of the Real Effort with the efforts calculated by DMCoMo and by the proposed estimation method for SMEs

When analysing the results of the DMCoMo method from table 13, it can be seen that the average error is very big (approximately 86 men x months for both formulas) with an error standard deviation of about ± 20 men x months respectively. DMCoMo always tends to overestimate the effort of the project (i.e. the error values are always negative) with a ratio greater than 590% (less difference for the project #6). This behaviour can be seen also graphically in figure 1. In addition, all estimated values are bigger than 60 men x months, which is the maximum threshold value previously identified for SMEs projects. From looking at these results, the conclusions of [9] are confirmed: DMCoMo estimation method is not recommended to predict the effort of small-sized information mining projects.

On the other hand, when the results of the proposed method for SMEs are analysed, it can be seen that the average error is approximately 1.46 men x months with an error standard deviation of approximately ± 2 men x months. In order to study the behaviour of the proposed method with the real effort a new boxplot graph is presented in figure 2. From this second boxplot graph, it seems that the behaviour of the proposed method tends to underestimate the real effort behaviour. There are similar minimum values (i.e. 1.56 men x months for the real effort and 1.08 men x months for the proposed method), maximum values (i.e. 11.63 men x months for REf and 9.80 for PEM), and averages (i.e. 5.77 and 4.51 men x months respectively).

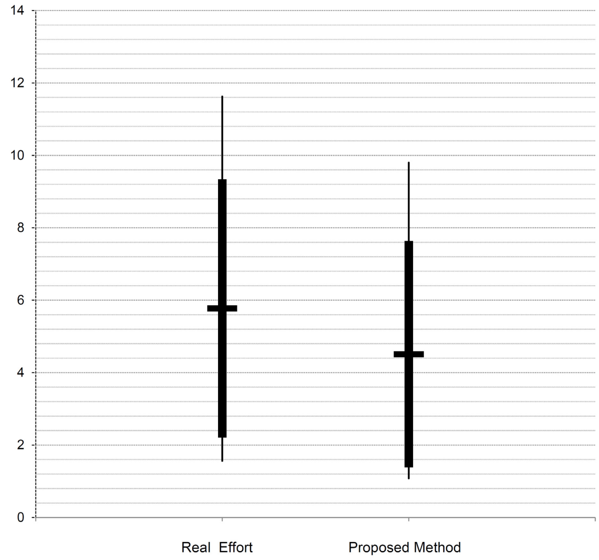

Fig. 2. Boxplot graph comparing the behaviour of the Real Effort with the effort calculated by the proposed estimation method for SMEs

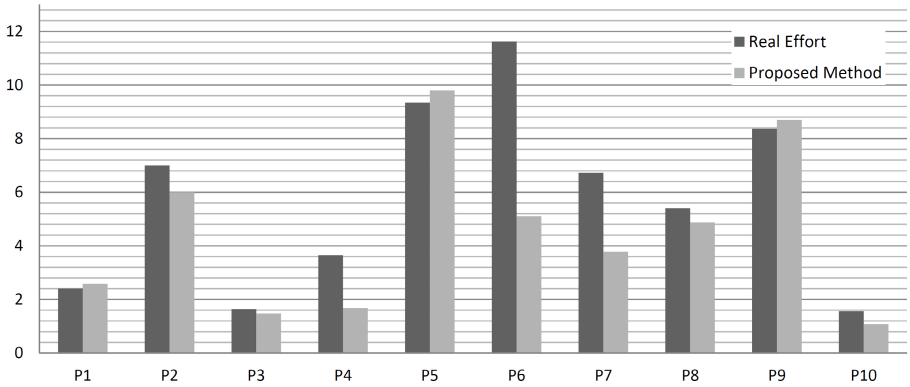

Fig. 3. Bar graph comparing for each project the Real Effort (REf) and the effort calculated by the proposed estimation method for SMEs (PEM)

Finally, if the real and estimated efforts of each project are compared using a chart graph (figure 3), it can be seen that the estimations of the proposed method are not completely accurate:

- Projects #1, #3, #5, #8 and #9 have estimated efforts with an absolute error smaller than one men x month and a relative error lower than 10%.
- Projects #2 and #10 have an estimated effort smaller than the real one with a relative error lower than 35%. In this case, the average error is about 0.74 men x months with a maximum error of one men x months (project *#2*).
- At last, projects #4, #6 and #7 have an estimated with a relative error greater than 35% (but lower than 60%). In this case, the maximum error is nearly 7 men x months (project *#6*) and an average error of 3.81 men x month.

6 Conclusions

Software projects need to predict the cost and effort with its associated quantity of resources at the beginning of every project. The prediction of the required effort to perform an Information Mining project is necessary for Small and Medium-sized Enterprises (SMEs). Considering the characteristics of these projects developed with the CRISP-DM methodology, an estimation method oriented to SMEs has been proposed defining seven cost drivers and formula.

From the validation of the proposed method, it has been seen that that the proposed method produces a more accurate estimation than the DMCoMo method for small-sized projects. But, even though the overall behaviour of the proposed method is similar to real project behaviour, it tends to perform a little underestimation (the average error is smaller than 1.5 men x month). It can be highlighted that 50% of estimations have a relative error smaller than 10%, and the 20% have a relative error between 11% and 35%. For the rest of estimations, the relative error is smaller than 57%. Nevertheless, in all cases the absolute error is smaller than 7 men x months. These errors could be due to the existence of other factors affecting the project effort which have not been considered in this version of the estimation method.

As future research work, the identified issues will be studied in order to provide a more accurate version of the estimation method oriented to SMEs by studying the dependency between the cost drivers and then adding new cost drivers or redefining the existing ones. Another possible approach is modifying the existing equation formula by using an exponential regression with more collected real project data.

Acknowledgements. The research reported in this paper has been partially funded by research project grants 33A105 and 33B102 of National University of Lanus, by research project grants 40B133 and 40B065 of National University of Rio Negro, and by research project grant EIUTIBA11211 of Technological National University at Buenos Aires.

Also, the authors wish to thank to the researchers that provided the examples of real SMEs Information Mining Projects used in this paper.

References

1. Schiefer, J., Jeng, J., Kapoor, S., Chowdhary, P.: Process Information Factory: A Data Management Approach for Enhancing Business Process Intelligence. In: Proceedings 2004 IEEE International Conference on E-Commerce Technology, pp. 162–169 (2004)
2. Stefanovic, N., Majstorovic, V., Stefanovic, D.: Supply Chain Business Intelligence Model. In: Proceedings 13th International Conference on Life Cycle Engineering, pp. 613–618 (2006)
3. Curtis, B., Kellner, M., Over, J.: Process Modelling. Communications of the ACM 35(9), 75–90 (1992)
4. Ferreira, J., Takai, O., Pu, C.: Integration of Business Processes with Autonomous Information Systems: A Case Study in Government Services. In: Proceedings Seventh IEEE International Conference on E-Commerce Technology, pp. 471–474 (2005)
5. Garcia-Martinez, R., Britos, P., Pollo-Cattaneo, F., Rodriguez, D., Pytel, P.: Information Mining Processes Based on Intelligent Systems. In: Proceedings of II International Congress on Computer Science and Informatics (INFONOR-CHILE 2011), pp. 87–94 (2011) ISBN 978-956-7701-03-2
6. García-Martínez, R., Britos, P., Pesado, P., Bertone, R., Pollo-Cattaneo, F., Rodríguez, D., Pytel, P., Vanrell, J.: Towards an Information Mining Engineering. En Software Engineering, Methods, Modeling and Teaching. Sello Editorial Universidad de Medellín, pp. 83–99 (2011) ISBN 978-958-8692-32-6
7. Rodríguez, D., Pollo-Cattaneo, F., Britos, P., García-Martínez, R.: Estimación Empírica de Carga de Trabajo en Proyectos de Explotación de Información. Anales del XVI Congreso Argentino de Ciencias de la Computación, pp. 664–673 (2010) ISBN 978-950-9474-49-9
8. Boehm, B., Abts, C., Brown, A., Chulani, S., Clark, B., Horowitz, E., Madachy, R., Reifer, D., Steece, B.: Software Cost Estimation with COCOMO II. Prentice-Hall, Englewood Cliffs (2000)
9. Marbán, O., Menasalvas, E., Fernández-Baizán, C.: A cost model to estimate the effort of data mining projects (DMCoMo). Information Systems 33, 133–150 (2008)
10. Pytel, P., Tomasello, M., Rodríguez, D., Pollo-Cattaneo, F., Britos, P., García-Martínez, R.: Estudio del Modelo Paramétrico DMCoMo de Estimación de Proyectos de Explotación de Información. In: Proceedings XVII Congreso Argentino de Ciencias de la Computación, pp. 979–988 (2011) ISBN 978-950-34-0756-1
11. García-Martínez, R., Lelli, R., Merlino, H., Cornachia, L., Rodriguez, D., Pytel, P., Arboleya, H.: Ingeniería de Proyectos de Explotación de Información para PYMES. In: Proceedings XIII Workshop de Investigadores en Ciencias de la Computación, pp. 253–257 (2011) ISBN 978-950-673-892-1
12. Organization for Economic Cooperation and Development: OECD SME and Entrepreneurship Outlook 2005. OECD Publishing (2005), doi: 10.1787/9789264009257-en
13. Álvarez, M., Durán, J.: Manual de la Micro, Pequeña y Mediana Empresa. Una contribución a la mejora de los sistemas de información y el desarrollo de las políticas públicas, CEPAL - Naciones Unidas, San Salvador (2009), `http://tinyurl.com/d5zarna`
14. 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)
15. Laporte, C., Alexandre, S.Y., Renault, A.: Developing International Standards for VSEs. IEEE Computer 41(3), 98–101 (2008)

16. Ríos, M.D.: El Pequeño Empresario en ALC, las TIC y el Comercio Electrónico. Instituto para la Conectividad en las Américas (2006), `http://tinyurl.com/c97qkjd`
17. Chapman, P., Clinton, J., Keber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R.: CRISP-DM 1.0 Step by step BI guide Edited by SPSS (2000), `http://tinyurl.com/crispdm`
18. Chen, Z., Menzies, T., Port, D., Boehm, D.: Finding the right data for software cost modeling. IEEE Software 22(6), 38–46 (2005), doi:10.1109/MS.2005.151
19. Domingos, P., Elkan, C., Gehrke, J., Han, J., Heckerman, D., Keim, D., et al.: 10 challenging problems in data mining research. International Journal of Information Technology & Decision Making 5(4), 597–604 (2006)
20. Weisberg, S.: Applied Linear Regression. John Wiley & Sons, New York (1985)