A Comprehensive Survey of Software Development Cost Estimation Studies

Shatha Al-Qudah
Hashemite University
Software Engineering Department
Zarqa, 13133, Jordan
P.O.Box 33127,00962-0776587544
shatha.a.qudah@itc.hu.edu.jo

Kenza Meridji
University of Petra
Software Engineering Department
Amman, 11196, Jordan
P.O.Box 33127,00962-791996084
kmeridji@uop.edu.jo

Khalid T. Al-Sarayreh
Hashemite University
Software Engineering Department
Zarqa, 13133, Jordan
P.O.Box 33127,00962-798471991
khalidt@hu.edu.jo

ABSTRACT

Context: Cost estimation of software projects is critical activity which requires the use of proper techniques and methods in order to have a good estimation results. This task is facing many obstacles which make it somehow challenging. The software size and how accurately it has been measured greatly affects the accuracy of the estimation. Also project management plays vital role in guiding such estimation processes. A lot of research effort has been accomplished which reflects the increasing demands of high quality software through effective cost estimation.

Purpose: To summarize and differentiate existing literature related to the cost estimation methods and techniques as away to identify estimation model fundamentals based on the analysis of the already available cost estimation models, also trying to identify the possible gaps in this research area.

Scheme: The methodology is based on the survey research method with well defined steps.

Outcomes: The survey search process found as preliminary results 274 papers associated to our topic from strong online databases, around 20 papers have been included in this survey and only 13 studies were selected for in-depth analysis. Data were extracted from these studies and synthesized in respect to the formulated Research questions .The survey results has shown that there are many models for cost estimation and the challenge is which one to use in specific software context, but still there are some commonly used models in both industry and academia.

Conclusion: The result of our work confirm that in order to have an accurate estimation process many software metrics need to be analyzed and used in estimation model or in a combination of models, and the estimation of software needs a lot of data and inputs to be accurate to meets expectation, this can be achieved using benchmarking of datasets.

Categories and Subject Descriptors

Knowledge Engineering, Information and Communication for developing, Technologies as Teaching Strategy.

IPAC '15, November 23-25, 2015, Batna, Algeria

DOI: http://dx.doi.org/10.1145/2816839.2816913

General Terms

Estimation, Design, Standardization.

Keywords

Cost Estimation, Software Metrics, Software Process, Algorithmic and non algorithmic Models.

1. INTRODUCTION

In software cost estimation research area there is a lot of work

Other has been replaced with a new model and in fact no model can be used or fit all types of projects[1]. Using such models assists the estimation process and support project management decision making, planning and feasibility study[2].

There are many aspects affect the estimation process such as unrealistic deadlines imposed over the project, the validity of the model used to measure the software size, software complexity [3]available budget and finally the risk of uncertainty of the estimated software .So the effectiveness of an estimation process depends on the used methods that measure it.

Through our analysis of the included literature we found that the main categorization of estimation models as algorithmic or non-algorithmic. The algorithmic models are sometimes called parametric which are based on mathematical formulas and specific measures of attributes related to the project [4], on the other hand the non-algorithmic models are theoretical and based on heuristics and require a good knowledge to estimate properly[4] .The paper scope is to classify most of the well known studies at software estimation model based on our research methodology. In addition, the main contribution of this paper is the comprehensive survey of the software development estimation studies.This paper is organized as follows. Section 2 presents the research description. Section 3 presents literature review. Section 4 presents preliminary of results. A conclusion is presented in section 5.

2. RESEARCH DESCRIPTION AND PRESENTATION

2.1 Research Methodology

This paper has been conducted as survey which means that it intend to review and study the existing related literature by applying a well defined steps to have strong overview. (Figure 1) below describes the **survey process**:

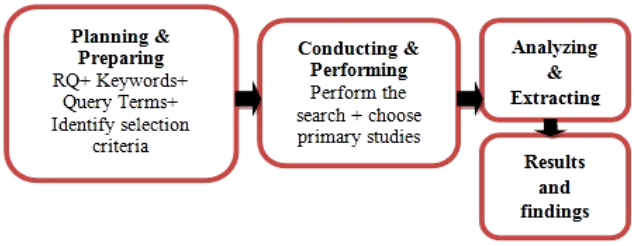

Figure1. Survey Process

Mainly we focus on papers within the (2007-2015) interval and well known publications from online databases which is relevant to our topic and research questions by analyzing and interpreting the studies based on the survey methodology [5].

2.2 Research Framework

The research was oriented on the main research questions that identify and implicitly include concepts and aspects that form a conceptual framework in which the survey try to analyze comprehensively. This conceptual framework is a starting point that help our research to be successful and replicated and define structure of our review, in addition it's important to identify the *estimation process* [6] here as a process that elaborate estimation methods, techniques or models and consider the metrics that affect the software cost in a clear set of steps that estimate the cost of a project[7], finally *software metrics* are attribute [8] of a software system or process attributes that can be measured numerically we will briefly discuss those metrics(See Figure 2).

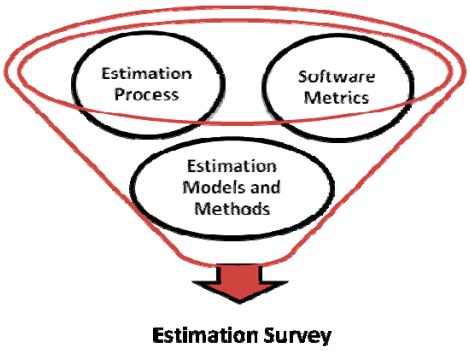

Figure 2.Cost Estimation Research Framework

2.3 Research Objective

Cost Estimation studies: software point of view.

3. LITERATURE REVIEW

Step1: Research planning: In this survey we three research questions, (See Figure 3).

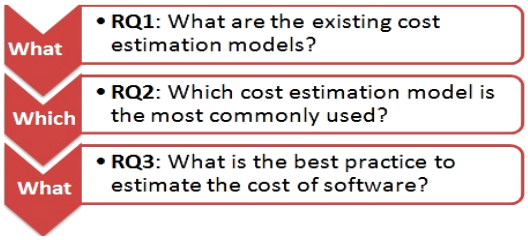

Figure 3. Research Questions

It's important to follow steps that help us to collect the relevant papers from the literature; our search was automatic using the online databases.

- Identify the purpose of this survey.
- Formulate the research questions to specify the search scope.
- Gather potential relevant papers based on the topic and the research question terms.
- Use well-known and leading software database journals and conferences limited by the interval from 2007 until 2015.

Data Sources: the largest and most common scientific databases used to find related cost estimation studies. (Table2) shows the studies along with publisher, type and publication year.

Identify Selection Criteria (See Figure 4):

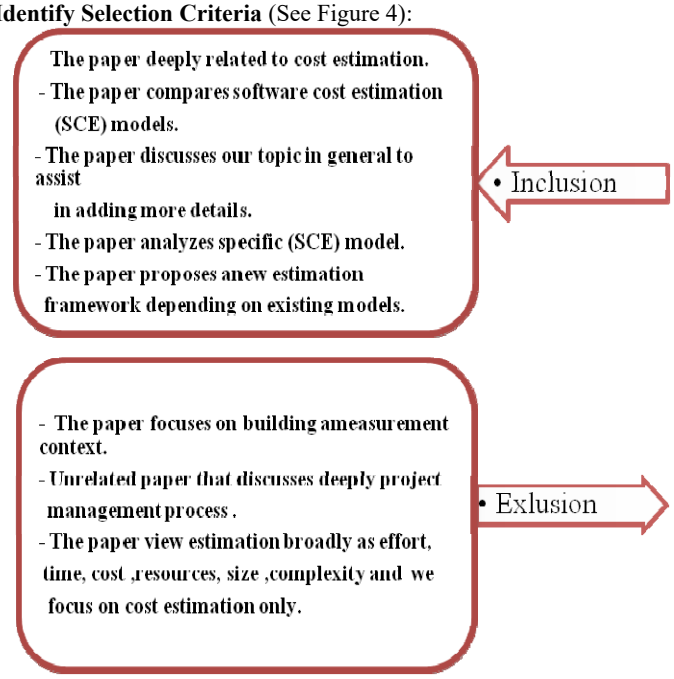

Figure 4. Selection Criteria

Step 2: Conducting the Search.

- **Search terms**: Search the primary studies using the appropriate terms that enable us to retrieve as much as possible the needed papers, the main used terms and key words are the following:
  - Software cost estimation, prediction or measurement.
  - Prediction or estimation models.
  - Cost estimation models, techniques, methods or tools.
  - AND/OR Booleans for linking the terms

Table 1.Cost estimation papers

Paper ID	Title	Authors	Citation
P1	Cost Estimation: A Survey of Well-known Historic Cost Estimation Techniques	Syed Ali Abbas, Xiaofeng Liao, Aqeel Ur Rehman, Afshan Azam, M.I. Abdullah	8
P2	An Approach for Software Cost Estimation	Violeta Bozhikova, Mariana Stoeva	4
P3	A Survey on Metric of Software Complexity	Sheng Yu, Shijie Zhou	26
P4	Summarization of Software Cost Estimation	Xiaotie Qin, Miao Fang	1

P5	Software Development Cost Estimation Using Similarity Difference between Software Attributes	Divya Kashyap ,A.K.Misra	2
P6	A Systematic Review of Software Development Cost Estimation Studies	Magne Jørgensen, Martin Shepperd.	563
P7	The Importance of the Use of an Estimation Process	Carlos Eduardo Lima Peixoto, Jorge Luis Nicolas Audy, Rafael Prikladnicki	1
P8	Software Functional Size: For Cost Estimation and More	Baris Ozkan, Oktay Turetken, Onur Demirors	9
P9	Computational Intelligence in Software Cost Estimation: An Emerging Paradigm	Tirimula Rao Benala, Satchidananda Dehuri, Rajib Mall	6
P10	Optimization of Fuzzy Analogy in software cost estimation using linguistic	Malathi, S., and S. Sridhar	4
P11	Software cost estimation using fuzzy logic	Mittal, Anish, Kamal Parkash, and Harish Mittal	35
P12	Overcoming the challenges in cost estimation for distributed software projects.	Ramasubbu, Narayan, and Rajesh Krishna Balan.	7
P13	ISBSG variables most frequently used for software effort estimation: A mapping review	Fernando González-Ladrón-de-Guevara, Marta Fernández-Diego	0

Step 3: Selection of the Primary Studies:

- To select relevant primary studies that help to answer the RQ's an inclusion and exclusion criteria's have been applied with the time range (2007-2015).
- To select primary studies from the 274, first we remove duplicate studies and then the Title and Abstract screened to exclude unrelated papers. Full text reading performed for the 13 included papers and deep analysis has been conducted. (See Table1).

Table2.Cost Estimation paper's publisher

Paper ID	Publisher	Type	Year
P1	JET- CIS	Survey	2012
P2	ACM	Conference Proceedings	2010
P3	IEEE	Survey	2010
P4	Elsevier	Journal	2011
P5	ACM	Conference Proceedings	2013
P6	IEEE	Systematic Review	2007
P7	ACM	Workshop Proceedings	2010
P8	Springer	Journal	2008
P9	ACM	Journal	2012
P10	Elsevier	Journal	2012
P11	ACM	Journal	2010
P12	IEEE	Conference Proceedings	2012
P13	ACM	Conference Proceedings	2014

Table 3.Cost estimation Models

Algorithmic	Non Algorithmic	Cost Estimation Model	Input or Key parameter
	X	Fuzzy Logic	**Fuzzy set theory, Reasoning [9]**
	X	Neural Networks (NN)	**Functional size**
	X	Expert Judgment (Delphi)(PERT)	**Consultant experience**
	X	Analogy	**Similar project cost**
	X	Nelson Model, SDC	**Man Month**
	X	Parkinson's Law	**Available resources**
	X	Price to Win	**Customer's budget**
	X	Top-down	**Can be applied using heuristics or parametric methods [10]**
	X	Bottom-up	**Can be applied using heuristics or parametric methods [10]**
X		Walston-Felix	**Line of code size**
X		Boeing Model	**Line of code size**
X		Bailey-Basili	**Line of code size**
X		Doty	**Line of code size**
X		Albrecht and Gaffney Model	**Functional size**
X		Kemerer	**Functional size**
X		Matson, Barnett & Mellichamp	**Functional size**
X		TRW (Scep)	**Person month**
X		Aron	**Instructions**
X		COCOMO	**Line of code size**
X		Putnam,SLIM Tool	**Functional Size , Line of code size**
X		GRC	**Line of code size**
X		Scheneider	**Halstead's software science**
X		Kustanowitz	**Man month**
X		Farr & Zagorski	**Man Month**
X		SEER SEM	**Functional size**
X		Mike II	**Functional size**

4. PRELIMINARY RESULTS

4.1 Data Extraction

By reading full text paper and define main items that will form study basics, also we have summarized many cost estimation models briefly and we have listed main advantages and disadvantages for each, (Tables 3 & 4) shows the items that were extracted from papers.

4.2 Evidence Synthesis

Through data synthesis we have answer the research questions through data extracted from studies.

Answers for the *first What* Question:

Algorithmic and Non-Algorithmic Cost Estimation models

There are two categories of cost estimation models; algorithmic and non algorithmic as shown in *Table 4*.The algorithmic models are clear repeatable and uses formulas and equations to calculate the cost, it may be further categorized as linear, non linear or quadratic, while the non algorithmic cost modeling rely on analysis, heuristics and sometimes require knowledge to conduct the cost of software.[4] Some researchers have discussed the existing models within A priori and A posteriori contexts. In A priori the models estimate the software early to help in decision making. On the other hand the A posteriori models rely on data from previous projects which is commonly used[11].In addition there is many other approaches including:
Lexical analysis of requirement, genetic programming, soft computing, and bootstrap-based analogy cost estimation that mostly in research area and researchers tries to combine it with other models [12].

Answers for the *Which* Question:

A comparison has been conducted among the models and we found that the most commonly used model is the algorithmic COCOMO model due to its benefits briefly shown in *Table 4*, other widely used models is the Expert Judgment as a non algorithmic model which the organization uses to reduce the cost of estimation and its accurate to some level as long as there is an expert, knowledge and data to rely on.
As we can see from Table 3, there are different inputs and parameters that are needed to estimate the cost using a model, software metrics are attribute [8] of a software system or process attributes that can be measured numerically. Those are metrics that must be identified and selected such as size metrics:[13] LOC based size, Function point (FP) based size , Object oriented (OO) based size and Use case point (UCP)based size(Pressman 2007).Another metrics are related to complexity such as Halstead Complexity Metric (HCM)(Software Science)[2] and Cyclomatic Complexity Metric (CCM) [3].Now the relation between estimation models, processes and metrics is somehow clear through our results as shown simply in Figure 2.

Answers for the *second What* Question:

It's important to note that no estimation model is appropriate and can handle all software project types [14].
Also there are many challenges that faces managers in estimating the costs as they commonly rely on standard metrics-based estimation models[15].
And as a result of our survey we found that the best practice for cost estimation is a combination of models to gain the best of it, also an important thing to mention that software metrics is a vital part of the estimation process.*"Software metrics let you know when to laugh and when to cry "* [13], as it's the building blocks and inputs to the estimation models. Finally we found that the accurate models are the one that let us calibrate and tune data for the used model.
Based on the results of our survey we recommend the use of benchmarking as away to estimate the software effort, size, cost and also its duration and to enhance the exiting algorithmic models. The ISBGS (The International Software Benchmarking Standards Group) which is repository for a huge number of projects and their related data can be used for benchmarking [16] .
Also we need models that can handle uncertitnity in an appropriate way.
Estimation experimental studies are mainly based on datasets, the most commonly used datasets both simple and complex includes Albrecht dataset, Desharnais dataset, Maxwell dataset and ISBSG (International Software Benchmarking Standards Group)[16], which are repositories for many real world data related to real projects, for the ISBSG it is widely used and encompasses large number of projects but due to the heterogeneous characteristic of it, it's recommended to extract an appropriate subset to use it in cost estimation activities.

Table 4. Comparison of Cost Estimation Models

Cost Estimation Model	Advantages	Disadvantages
COCOMO I , COCOMO'81,COCOMO II: Cost Constructive Model (Boehm 1981), Consider three development modes (Organic/Embedded/Semi-detached) [9], Three levels (Basic /Intermediate/Detailed)[2], Uses 15-17Cost Drivers/Factors (Software Attributes)[8].	- Most commonly used model. - Objective & Documented [1] - Can be calibrated and customized for Local project data. - Clear steps to follow. - Practiced at industry.	- Inaccurate size or/and cost drivers measurement will lead to poor estimation - Not always fair in quantifying varying skill Levels.
Putnam's SLIM : Software Life Cycle Management (Lawrence H. Putnam 1978, automated macro model, consider the Rayleigh curve, include 25 factors, have three levels of factors (High, Medium, Low)[2]**.**	- Fast and easy. - Suitable for large scale project - Less attributes and inputs for estimation compared to COCOMO	- Lack of some details required to estimate Parts of the software. - Problems of *constant factor* it use which is Related to the environment.
Walston & Felix :(IBM 1977)	-Based on historical statistical data. -Early proposed model and a foundation of other models	- Haven't been practiced a lot because of Statistical problems.
Nelson : (Nelson 1970), referred as SDC-[2]Software Development Corporation. Include 14 cost drivers.	- Equations to estimate manpower, computer Usage and months elapsed [2].	- Long equations. - Statistical problems.
Boeing:(black 1971)**,**Similar to COCOMO Model	- Simple model of estimation.	- Not possible to calibrate and adjust with Project scope and Not used in practice.
Expert Judgment:(Delphi) expert or group of experts tries to predict the cost and estimate it and they may use tools to assist them, and it can be used in conjunction with parametric models to validate the result.	- Fast estimation. - No need for many resources. - Less cost for estimation process. - Achieves good accuracy results in estimation. - Widely used.	- Subjective [1]as it depends on the person who is judging the decision. - Require experienced consultants with good Knowledge.
Analogy: An estimation approach that requires knowledge, precise, correct and related data based on	- Semi-objective [1]. - provide check for estimation with other	- Relay on human knowledge and Experience.

similar projects from the same domain and then adjusts it with the new project.	Models.	- Require logic and reasoning.
Bottom up: Require the breaking of the project into smaller components, estimate each and then summarize the results to conduct the overall estimation.	- Can estimate each components of the software to lead to a detailed estimation Results.	- Needs a lot of time, resources to Implement it. - Require understanding of the components.
Top Down: based on an overview to the software as a whole and focuses on the high level activities s such as Integration and Configuration Management CM.	- Focuses on system or high level activities.	- May not cover low level components and Activities.
Parkinson (Cyril Parkinson 1995)**:**"*Work expands so as to fill the time available for its completion*."Depends on what are the available resources	- Can be adopted by project Managers.	- May not deliver as scheduled.
Fuzzy Logic: used upon existing formal methods and the appropriate metrics to conduct a range of values that are possible instead of depending on numeric specific value,[9]. Fuzzy Logic take care of values with partially truth [17]**.**	- Used whenever the available information are Uncertain or ambiguous (Fuzzy)[9]. - Flexible and simple to understand as it reflects the logic of humans[17].	- Not clear process so it's not widely used And accepted by estimators.
Neural Networks: consist of neurons and levels, there connections that let attributes to pass through.	- Facilitate and enable the solution of some difficult problems in cost estimation[14].	- Not easy process so it's not widely used And accepted by estimators.

5. CONCLUSION

We survey the state of the art in cost estimation research area in order to gain a clear picture of the estimation models, and we have conducted that there are some dominant models that is used repeatedly but that does not prove its validity in all context of the Measured software .Also there is the uncertitnity risk in software which requires more accurate and suitable models to predict it.

As future work we are studying the ISBSG data set to understand its available data items ,also we plan to enhance existing cost estimation models or contribute in the estimation process as a whole within the software cost estimation field.

6. REFERENCES

[1] V. Bozhikova and M. Stoeva, "An approach for software cost estimation," in *Proceedings of the 11th International Conference on Computer Systems and Technologies and Workshop for PhD Students in Computing on International Conference on Computer Systems and Technologies*, pp. 119-124.

[2] S. A. Abbas, X. Liao, A. U. Rehman, A. Azam, and M. I. Abdullah, "Cost Estimation: A Survey of Well-known Historic Cost Estimation Techniques," *Journal of Emerging Trends in Computing and Information Sciences,* vol. 3, pp. 612-636.

[3] S. Yu and S. Zhou, "A survey on metric of software complexity," in *Information Management and Engineering (ICIME), 2010 The 2nd IEEE International Conference on*, pp. 352-356.

[4] D. Kashyap and A. K. Misra, "Software development cost estimation using similarity difference between software attributes," in *Proceedings of the 2013 International Conference on Information Systems and Design of Communication*, pp. 1-6.

[5] S. Keele, "Guidelines for performing systematic literature reviews in software engineering," in *Technical report, Ver. 2.3 EBSE Technical Report. EBSE*, 2007.

[6] C. E. L. Peixoto, J. L. N. Audy, and R. Prikladnicki, "The importance of the use of an estimation process," in *Proceedings of the 2010 ICSE Workshop on Software Development Governance*, pp. 13-17.

[7] X. Qin and M. Fang, "Summarization of Software Cost Estimation," *Procedia Engineering,* vol. 15, pp. 3027-3031.

[8] I. Sommerville,"Software Engineering ", 9th ed, 2010.

[9] S. Malathi and S. Sridhar, "Optimization of Fuzzy Analogy in Software cost estimation using linguistic Variables," *Procedia Engineering,* vol. 38, pp.177-190.

[10] B. Ozkan, O.Turetken, and O. Demirors, "Software functional size: For cost estimation and more,"in *Software Process Improvement*: Springer, 2008.

[11] K. T. Al-Sarayreh and K. Meridji, "Towards apriori and posteriori estimation models for renewable energy projects in software engineering," in *Applications of Information Technology to Renewable Energy Processes and Systems (IT-DREPS), 2013 1st International Conference & Exhibition*, pp. 101-106.

[12] M. Jorgensen and M. Shepperd, "A systematic review of software development cost estimation studies," *Software Engineering, IEEE Transactions on,* vol. 33, pp. 33-53, 2007.

[13] R. Pressman, "Software Engineering: A Practitioner's Approach", 2007.

[14] T. R. Benala, S. Dehuri, and R. Mall, "Computational intelligence in software cost estimation: an emerging paradigm," *ACM SIGSOFT Software Engineering Notes,* vol. 37, pp. 1-7.

[15] N. Ramasubbu and R. K. Balan, "Overcoming the challenges in cost estimation for distributed software projects," in *Proceedings of the 34th International Conference on Software Engineering*, pp. 91-101.

[16] F. Gonz‹Ílez-Ladr³Ín-de-Guevara and M. Fern‹Índez-Diego, "ISBSG variables most frequently used for software effort estimation: A mapping review," in *Proceedings of the 8th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, p. 42.

[17] A. Mittal, K. Parkash, and H. Mittal, "Software cost estimation using fuzzy logic," *ACM SIGSOFT Software Engineering Notes,* vol. 35, pp. 1-7.