Fraunhofer Institut
Experimentelles
Software Engineering

A comparative Study of Cost Modeling Techniques using Public Domain multi-organizational and company-specific Data

Authors:
Ross Jeffery
Melanie Ruhe
Isabella Wieczorek

Accepted for publication in
Proceedings of European Software Control and Metrics, ESCOM2000 and special issue of Information and Software Technology

IESE-Report No. 004.00/E
Version 1.0
January 2000

A publication by Fraunhofer IESE

Fraunhofer IESE is an institute of the Fraunhofer Gesellschaft.
The institute transfers innovative software development techniques, methods and tools into industrial practice, assists companies in building software competencies customized to their needs, and helps them to establish a competetive market position.

Fraunhofer IESE is directed by
Prof. Dr. Dieter Rombach
Sauerwiesen 6
D-67661 Kaiserslautern

Abstract

This research examines the use of the International Software Benchmarking Standards Group (ISBSG) repository, which is a large database of completed software projects from different organizations, for estimating the required effort for new software projects. The accuracy of the estimates based on this repository is compared with the results obtained from using a one-company data set from a company called Megatec. This study investigates two questions: (1) What are the differences in accuracy between a traditional technique such as ordinary least-squares (OLS) regression and Analogy-based estimation? (2) Is there a difference between estimates derived from multi-company data and estimates derived from company-specific data? Regarding the first question, our results show that OLS regression performs as well (when based on one-company data) and significantly better than (when based on multi-organizational data) Analogy-based estimation. This result is in contrast to previous studies that showed promising results applying Analogy on software engineering data. On the other hand, the result confirms the outcomes of investigating Analogy on another large multi-organizational database (called Laturi) from the business applications domain. Addressing the second question, we found two results. When applying Analogy, significantly more accurate models could be built based on company-specific data than based on multi-organizational data. The results reveal that Analogy-based procedures do not seem as robust when using data external to the organization for which the model is built. When applying OLS regression, no significant advantage was found when using local, company-specific data opposed to multi-organizational data. Again, this result is consistent with a previously performed comprehensive comparison on the Laturi database as well as on the ESA database. We plan to further investigate the reasons for consistencies and inconsistencies in the current and previous results to derive generalizable conclusions.

Table of Contents

1 Introduction

Delivering a software product on time, within budget, and to an agreed level of quality is a critical concern for many software organizations. Underestimating software costs can have detrimental effects on software quality and thus on a company's business reputation. On the other hand, overestimation of software cost can result in missed opportunities to fund other projects. In response to industry demand, a myriad of estimation techniques have been proposed during the last three decades. In order to assess the suitability of a cost modeling technique, its performance and relative merits must be compared. Normally, homogenous company-specific data are believed to form a better basis for more accurate estimates. However, those data sets are typically small and cost driver data are tailored and specific such that comparison with other organizations or across the industry is impossible. Moreover, data collection is an expensive and time-consuming process for individual organizations. Industry representative parties have addressed the problem of software data collection in the past few years with the advent of multi-organizational data sets. The collaboration of organizations (such as the International Software Benchmarking Standards Group) to form multi-organizational data sets provides a possibility for reduced data collection costs, faster data accumulation and shared information benefits. Therefore, the pertinent question is whether multi-organizational data are valuable for estimation.

In this study, we used public domain data from the ISBSG. We compared the estimates derived from those data with estimates derived from company-specific data from an Australian company (Megatec). The Megatec data does not form a part of the ISBSG data set. The Megatec projects' main applications are in business areas such as financial, banking, or inventory. The modeling techniques selected for the comparison showed promising results in previous studies. The first technique, OLS regression is one of the most commonly applied techniques. The second is Analogy-based estimation, whose popularity increased in the 90's. We applied different variants of the ACE (Analogical and Algorithmic Cost Estimator) algorithm to our data sets. This algorithm calculates the difference between the target project and each completed project in a database for a set of search metrics. ACE ranks the completed projects in a database according to their similarity. The effort of the most similar project(s) is used to predict the effort for the target project. In addition, size adjustments may be applied to address differences between projects.

Our study is motivated by the challenge to assess the feasibility of using multi-organization data to build cost models and the benefits gained from company-specific data collection. The study determines the prediction accuracy of two

different estimation techniques and examines their performance based on both multi-organizational and company-specific contexts. Thus, two important questions are addressed: (1) What are the differences in accuracy between a traditional technique such as ordinary least-squares (OLS) regression and Analogy-based estimation? (2) Is there a difference between estimates derived from multi-company data and estimates derived from company-specific data?

Two other pieces of research have been carried out in this area. The work of Briand et al. [1] was based on the so-called "Laturi-database", which includes 206 business software projects from 26 companies in Finland. In this research the following two questions were investigated. What modeling techniques are likely to yield more accurate results when using typical software development cost data? What are the benefits and drawbacks of using organization-specific data as compared to a multi-organization data set? The second project investigated the same research questions using the ESA data set as the basis for investigation [2]. In both cases the research questions were addressed using organizations within the data set. This research takes an organizational data set from outside the multi-organization set and therefore provides a possibly more stringent test to the use of these types of data sets. It also differs in that the level of knowledge the researchers have of the Megatec data is higher than could be expected of any public data set since the researchers were involved in the collection of the Megatec data in the first place and carried out extensive prior analysis (see [6]). Thus extensive knowledge of the data context, relationships and accuracy was present. There have been two previous studies that utilized the ISBSG data set: The first one is a descriptive study done by the ISBSG itself [9]. Examples of the areas it analyzes are project size, project effort, and other descriptive metrics, e.g. their range, distribution, and their relationship. Lokan [7] investigated the relationship between the five types of elements in function point analysis. Thus this is the first application of this data set to the issue of cost estimation.

2 Research Method

2.1 The Data Sets

Two data sets are used in this work: ISBSG, a publicly available multi-organizational data set consisting of a total of 451 projects, and Megatec, a single-company data set consisting of 19 projects. The Megatec data is used to evaluate estimates made using the ISBSG data and to compare these with estimates made using the Megatec data itself. Not all of the 451 ISBSG projects were used in the study because of the need to match characteristics of the two data sets as closely as possible. The subset used included only those ISBSG projects that fulfill the three criteria. Firstly, resources need to be measured on the same basis as in Megatec. Thus, ISBSG projects where the effort measure reflected the development team and possibly also the people supporting that development team were included. If the effort measure included operations and end user time these projects were excluded, as this time was not included in the Megatec data set. Secondly, projects needed to have no missing values for the metrics function points and team size. These two variables were essential for analysis. These two qualifications resulted in a subset of 225 projects. In order to further match the characteristics of ISBSG compared with Megatec their development type was also used. Megatec projects were completely new developments whereas ISBSG projects can be new developments or re-developments or enhancements. Therefore, the criterion of "new development" was added as a third selection criterion. The result was a subset of 145 ISBSG projects.

Megatec is an Australian software development organization with about 50 employees at the time of data collection that developed and distributed a range of computer products in Australia and the USA. It was one of the first software companies in Australia to gain Australian Standard 3563 (IEEE-Std.-1298), a company that was highly motivated to provide good quality data and that was also interested in research results [6]. The full description of the data set can be obtained from [6].

Software practitioners voluntarily submitted the projects in the ISBSG data set. The ISBSG repository (release 5 March 1998) consisted of projects from fourteen countries around the world; Australia is the largest contributor. There are 38 metrics collected that describe each project. The ISBSG data set was collected using questionnaire. Each project submitted to the ISBSG repository is validated against specific quality criteria. Function points were counted largely using the IFPUG standard. There is a wide range of programming languages; the systems are mainly written using ACCESS, COBOL, NATURAL, PL/1, and

TELON. The range of business area types is also quite wide. The ISBSG software projects were performed in many different areas, whereas the Megatec data is only obtained from projects performed in the software development company Megatec. The application types of ISBSG are mainly management information systems, office information systems or transaction & production systems. Unfortunately, there is neither any data about the experience of the software developers, nor any metric that identifies the company or gives information about the organization type of the company in the repository. Therefore, research as done by Briand et al. [1], which compared estimations for multi-organization and organization-specific data only using data from within the repository, cannot be repeated with the ISBSG data only.

The metrics in the ISBSG data set used are shown in Table 1. None of the distributions are normal. The range for Function Points and Work Effort in ISBSG is large compared to the Megatec data. Megatec projects required 1947 hours of Effort on average. Team Size also shows a large difference compared with a range of 1 to 10 people for Megatec projects. Brooks [3] says that a high number of team members usually do not improve the productivity because of time used for the group communication and organization. The relationship between Team Size and PDR or Effort is not linear. The other documented metrics have no counterparts in the Megatec data set, but also show a broad spectrum.

	Megatec		ISBSG	
Metric	Scale	Values/Range	Scale	Values/ Range
Total effort in hours	ratio	194 to 13905	ratio	10 to 59809
Unadjusted function points	ratio	39 to 3290	ratio	11 to 9803
Maximum team size	ratio	1 to 10	ratio	1 to 55
Client/server	nominal	yes, no	nominal	yes, no
Language type	nominal	3GL, 4GL	nominal	ApG, 2GL, 3GL, 4GL
Development platform	ordinal	PC, midrange	ordinal	PC, midrange, mainframe

Table 1. Project Metrics used for the Analysis (from Megatec and ISBSG data set)

The project delivery rate (PDR) as a measure of productivity is defined as hours per function point. A high number indicates more hours per function point were needed and, therefore, shows that the productivity is low and vice versa in a small number. Table 2 compares the PDR of the two data sets.

	Megatec			ISBSG		
PDR	min	max	mean	min	max	mean
	1.50	11.44	4.97	0.19	35.87	8.18

Table 2. Comparison of the data sets by their productivity

2.2 Model Building and Application

In order to determine the accuracy of estimates based on company-specific data, we followed the leave-one-out approach (19-fold cross-validation). For each of the 19 projects of the Megatec data set, we used the remaining 18 projects as a basis for model building. The overall accuracy was aggregated across the 19 projects. Calculating the accuracy in this manner emulates the situation when a company derives a cost estimation model using its own data. In order to determine the accuracy of estimates based on multi-organizational data, we used the 145 ISBSG projects as a basis for predicting the 19 Megatec projects. Thus, the cost of each target project was also estimated using the ISBSG data. This emulates the situation when a company uses external data to build prediction models for its own, internal projects.

Figure 1 depicts the main steps:

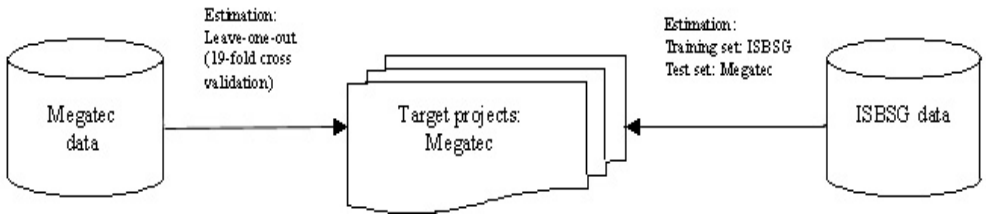

Figure 1. Model Building and Application

2.3 Estimation Techniques Used

Ordinary least-squares regression

We applied multivariate ordinary least-squares regression analysis [8] fitting the data to a specified exponential functional form. The data of the considered variables has been ln-transformed, because the ISBSG and Megatec data are not normally distributed. Moreover, heteroscedasticity was prevalent when using linear model specifications. We used the ratio-scaled variables from Table 1 to form our regression equations.

Analogical and Algorithmic Cost Estimator (ACE)

Analogy-based estimation involves the comparison of a new (target) project with completed (source) projects. The most similar projects are selected from a database as a basis to build an estimate. Major issues are to select an appropriate similarity function, to select relevant project attributes (in our case cost-drivers), and to decide upon the number of similar projects to consider for estimation (analogues).

A prototype of the Analogical and Algorithmic Cost Estimator (ACE), has been developed as a means to explore the benefits of Analogy-based estimation. ACE estimates effort for a target project by searching through a database of metrics for completed projects and selecting the completed project that it judges most similar to the target project [10]. The tool uses a certain ranking algorithm when searching for similar projects. One part of the ranking algorithm is the calculation of the difference between the target project and each source project for each metric and each source project. According to this difference the completed projects get ranked. This is done for all included variables and the overall ranking determines the most similar project(s). The effort is predicted by the most similar project(s). Additionally, an adjustment can be applied in order to address the differences in size between target and source project. The Adjustment is defined as:

$$Effort_{TARGET} = \frac{Effort_{ANALOGUE}}{FP_{ANALOGUE}} \times FP_{TARGET}$$

We applied four alternative versions of ACE. (1) Using the most similar analogue for effort prediction without any size adjustment, (2) using the average of the two most similar analogues for effort prediction without any size adjustment, (3) using the most similar analogue for effort prediction with size adjustment, and (4) using the two most similar analogues for effort prediction with size adjustment. The abbreviations we use in the following are "ACE-1 no SA", "ACE-2 no SA", "ACE-1 with SA", and "ACE-2 with SA", respectively.

2.4 Evaluation Criteria

The evaluation of the cost estimation models was done by using the following common criteria [4]. Absolute relative error as a percentage of the actual effort for a project, is defined by:

$$ARE = \left| \frac{Effort_{TARGET} - Effort_{ESTIMATED}}{Effort_{TARGET}} \right|$$

The *ARE* is calculated for each observation. Either the mean ARE or the median ARE can achieve the aggregation of multiple observations. The median MRE is less sensitive to extreme values. In addition, we used the prediction level *Pred*. This measure is often used in the literature and is a proportion of a given level of accuracy:

$$Pred(l) = \frac{k}{N}$$

Where, N is the total number of observations, and k the number of observations with an ARE less than or equal to l. A common value for *l* is 0.25, which is used for this study as well.

3 Analysis and Results

3.1 Results based on Megatec Data

Following our cross-validation approach, we built 19 different OLS models. Applying ACE, analogues were identified using 18 remaining projects as source projects for estimating each target project. This procedure was applied for each Megatec project. Table 3 presents the aggregates results obtained when applying OLS regression and ACE. The first column gives the estimation method. For each technique, we provide the mean ARE, the median ARE, as well as the Pred(0.25) values.

Estimation Method	Mean ARE	Median ARE	Pred(0.25)
OLS	0.47	0.30	0.42
ACE-1 no SA	0.63	0.35	0.42
ACE-2 no SA	0.55	0.46	0.16
ACE-1 with SA	0.38	0.27	0.32
ACE-2 with SA	0.37	0.28	0.47

Table 3. Estimates based on Megatec projects applied to Megatec target projects

Looking at the mean ARE values, no large accuracy differences exist among the techniques. The largest difference observed is 0.26 (ACE-1 no SA vs. ACE-2 with SA). This is confirmed when looking at the median ARE values. The largest difference here is 0.19 (ACE-2 no SA vs. ACE-1 with SA). In general, OLS regression and ACE estimates using linear size adjustment perform slightly more accurate than ACE estimates without size adjustment.

3.2 Results based on ISBSG Data

The following OLS regression model was derived based on 145 ISBSG projects to predict the 19 Megatec projects:

*ln (effort) = 2.393 + 0.822*ln (max team size) + 0.684*ln (fp)*

Table 4 summarizes the accuracy results obtained from OLS regression and ACE. Applying ACE, analogues were identified using 145 projects as source projects for estimating each target project from Megatec.

Estimation Method	Mean ARE	Median ARE	Pred(0.25)
OLS	0.55	0.33	0.26
ACE-1 no SA	2.48	0.90	0.05
ACE-2 no SA	1.47	0.66	0.16
ACE-1 with SA	2.39	0.84	0.16
ACE-2 with SA	1.43	0.72	0.05

Table 4. Estimates based on 145 ISBSG projects applied to Megatec target projects

The mean and median ARE is the lowest for estimates derived through OLS regression. This is confirmed looking at the Pred(0.25) value for OLS. The accuracy of ACE when using one analogue is highly affected by producing outlying predictions as indicated through the high mean and median ARE values for ACE-1 no SA and ACE-1 with SA. Applying ACE using two analogues significantly decreases the mean ARE values.

3.3 Comparisons

To address our first stated question, "What are the differences in accuracy between a traditional technique such as ordinary least-squares regression and Analogy-based estimation?", we compared each of the technique's accuracy (1) for estimates derived from Megatec data, and (2) for estimates derived from ISBSG data. We tested statistical significance using the Wilcoxon matched pairs test, a non-parametric analogue to the t-test [5]. Table 5 reports the p-values. The column entitled "Estimates based on Megatec" reports results within the Megatec data set (see also Table 3). No significant differences among the techniques can be observed when models are based on company-specific data. This result is consistent with the study on the Laturi database, where also no significant differences could be found among various applied modeling techniques using company-specific data.[1].

Using multi-organizational data and applying the derived models on Megatec (column "Estimates based on ISBSG"), OLS performs significantly better than any of the ACE variants (see also Table 4). This is also in line with the results from the Laturi, as well as from the ESA study [1][2]. From these results, it seems that using simple OLS regression provides the most accurate results.

Having a closer look at the different ACE variants, we observe significant differences among some of the variants for models based on multi-company data (Table 5, column "Estimates based on ISBSG" and Table 4). In this context, size adjustment does not significantly improve the estimates. Neither for ACE using one analogue, nor for ACE using two analogues the differences of using/not using size adjustment are significant. However, it seems that the use of more than one analogue is a driving factor of significant accuracy improvement. We

would suggest that these results are a reflection of the large distribution with respect to size in the data set and the high likelihood that the selected analogue project will, therefore, be very different in scale from the target. This is much less likely in the Megatec data. Here, the ACE estimates benefit in some instances from the use of the size adjustment algorithm. For the Megatec based comparisons, we see that only ACE using two analogues and size adjustment gets close to a significantly higher accuracy than ACE using two analogues and no size adjustment (p-value= 0.091).

	Estimates based on Megatec				Estimates based on ISBSG			
	ACE-1 no SA	ACE-1 with SA	ACE-2 no SA	ACE-2 with SA	ACE-1 no SA	ACE-1 with SA	ACE-2 no SA	ACE-2 with SA
ACE-1 no SA	-				-			
ACE-1 with SA	0.468	-			0.147	-		
ACE-2 no SA	0.778	0.421	-		0.003	0.064	-	
ACE-2 with SA	0.198	0.573	0.091	-	0.005	0.020	0.717	-
OLS	0.841	0.398	0.355	0.376	0.001	0.003	0.004	0.005

Table 5. Comparison of Techniques (Wilcoxon matched pairs test)

Addressing our second question, "Is there a difference between estimates derived from multi-company data and estimates derived from company-specific data?", Table 6 compares for each technique the model accuracy in different contexts (the median ARE values are provided). When applying Analogy, for almost each ACE variant, significantly different (more accurate) models could be built based on company-specific data than based on multi-company data. The trend for ACE is what one would expect, because of the higher homogeneity of the underlying company-specific data set. Moreover, the average productivity of Megatec projects is higher than for ISBSG projects (Table 2). This may explain a consistent overestimation of Megatec projects, when the predictions are based on ISBSG projects. Selecting ISBSG analogues of similar project size to the targeted Megatec projects lead to overestimated effort values.

When applying OLS regression, no significant advantage was found when using local company-specific data opposed to multi-organizational data. This is consistent with the findings of the Laturi and the ESA studies [1][2].

Comparison	Median ARE for Megatec based estimates (see Table 3)	vs	Median ARE for ISBSG based estimates (see Table 4)	p-value
OLS	0.30	vs	0.31	0.355
ACE-1 no SA	0.35	vs	0.59	0.016
ACE-1 with SA	0.43	vs	0.84	0.000
ACE-2 no SA	0.27	vs	0.74	0.058
ACE-2 with SA	0.28	vs	0.74	0.003

Table 6. Megatec versus ISBSG based estimates (Wilcoxon matched pair test)

4 Discussion and Conclusions

To summarize, we can conclude that for Megatec: (1) Estimates based on the ISBSG data set using OLS as a modeling technique are not significantly different on average to estimates based on their own data. (2) Estimates using their own data are slightly more accurate on average (but not significantly) using Analogy adjusted for the expected size of the target project. (3) Analogy does not seem to be an appropriate method for Megatec if they estimate using the ISBSG data set.

The following observations can be made. It is completely logical that size adjustment needs to be made to the Analogy-based estimate. The regression model-based estimates already include a size adjustment by default since function points is one of the independent variables in the model. It seems that analogue selection in multi-organization data sets is an issue here. It was relatively easy to nominate the possible variables for consideration in the analogue selection process for Megatec as prior work in this organization had been directed at identifying the most pertinent cost drivers for that company [10]. The same variables, however, do not necessarily have the same measured impacts within companies from the ISBSG data set. Thus, linking the cost relationships from Megatec to ISBSG was not terribly successful, resulting in poorer estimates for Analogy than multivariate OLS using the ISBSG data. Whether such cost driver insight is possible in such a data set is an open issue.

The practical implications of these results are that: (1) If an organization were to have a fairly sizeable data set and a very good understanding of their own cost drivers, they would likely be better to use Analogy as their estimation method. This advantage may not derive from the average estimation accuracy when compared with OLS, but experience shows that Analogy is more familiar to practitioners and is the technique preferred in informal estimation procedures. Comparison of estimates derived in this way with OLS model estimates would also be worthwhile. (2) If an organization does not have a sizeable data set of their own, then a regression model derived from public data sets would be more advisable than using Analogy. Cost factors derived from such regression modelling should be checked with conventional wisdom within the organization in order to provide informal model validation.

In this paper, we compared one parametric and one non-parametric cost estimation technique (OLS regression and ACE respectively). In order to derive more generalizable conclusions, future work remains to be done evaluating a variety of other cost estimation techniques on those two data sets. In addition, an investigation of consistencies and differences in the results of previous

studies [1][2] is intended. This also implies an in-depth investigation of the characteristics of the data sets.

Acknowledgements

This work was supported by grants from Megatec, the Fraunhofer Institute for Experimental Software Engineering, the Centre for Advanced Empirical Software Research (CAESAR) at UNSW, and the CSIRO. Thanks also go to the International Software Benchmarking Standards Group (ISBSG) for the repository used in the study.

References

[1] Briand, L.C. El Emam K., Maxwell, K., Surmann, D., Wieczorek, I. An Assessment and Comparison of Common Software Cost Estimation Models. In: Proceedings of the 21st International Conference on Software Engineering, ICSE 99, Los Angeles, USA 1998, pp. 313-322.

[2] Briand, L.C., Langley, T., Wieczorek, I. Using the European Space Agency Database: A replicated Assessment of Common Software Cost Estimation Techniques. Technical Report, ISERN TR-99-15, International Software Engineering Research Network, 1999

[3] Brooks, F. The Mythical Man Month. Addison-Wesley, Inc., 1995

[4] Conte, S.D., Dunsmore, H.E., Shen, V. Y. Software engineering metrics and models. The Benjamin/Cummings Publishing Company, Inc., 1986.

[5] Gibbons, J.D.S. Nonparametric Statistics. In: Quantitative Applications in the Social Sciences 90, Sage Publications, 1993.

[6] Jeffery R., Stathis J. Function Point Sizing: Structure, Validity and Applicability. Empirical Software Engineering, 1996, pp.11-30.

[7] Lokan C.J. An Empirical Study of the Correlations between Function Point Elements. Proceedings of the 6th International Software Metrics Symposium, 1999

[8] Schroeder, L., Sjoquist, D., Stephan, P. Understanding Regression Analysis: An Introductory Guide. No. 57, In: Quantitative Applications in the Social Sciences, Sage Publications, Newbury Park CA, USA, 1986.

[9] Software Project Estimation: A Workbook for Macro-Estimation of Software Development Effort and Duration, ISBSG, 1999

[10] Walkerden F., Jeffery R. An Empirical Study of Analogy-based Software Effort Estimation. Empirical Software Engineering, 4, 2, June 1999, pp. 135-158.

Document Information

Title:	A comparative Study of Cost Modeling Techniques using Public Domain multi-organizational and company-specific Data
Date:	January, 2000
Report:	IESE-004.00/E
Status:	Final
Distribution:	Public