Comparing Cross- vs. Within-Company Effort Estimation Models Using Interval Estimates[*]

Stamatia Bibi[1], Nikolaos Mittas[1], Lefteris Angelis[1], Ioannis Stamelos[1], and Emilia Mendes[2]

[1] Department of Informatics, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece
{sbibi, nmittas, lef, stamelos}@csd.auth.gr
[2] Computer Science Department, The University of Auckland, Private Bag 92019
Auckland, New Zealand
emilia@cs.auckland.ac.nz

Abstract. This paper investigates whether effort predictions for projects from a single company that were obtained using a cross-company (CC) training set can be as accurate as effort predictions obtained using a within-company (WC) training set. We employed five different cost estimation techniques, two providing point estimates (estimation by analogy and stepwise regression) and three providing predefined interval estimates (ordinal regression, classification and regression trees and Bayesian networks). For the development and evaluation of both cross and within company models ISBSG release 9 was utilized. Our results showed no significant differences between CC and WC-based predictions, for all the cost estimation techniques, after comparing the medians of the absolute errors. Other accuracy metrics were also considered, providing in general similar results.

Keywords: Software effort estimation, predefined intervals, cross company estimation models, within company estimation models, regression models, CART, estimation by analogy, Bayesian networks.

1 Introduction

During the past decades, a large number of methods have been proposed for software cost estimation [22], most of them relying on past experience or historical data coming from projects developed by one or more companies. An interesting question is whether one single company should use data from different companies (cross-company data) in order to create an effort estimation model for its projects or to use its own data (within-company data). The use of within-company (WC) data seems a reasonable choice since older finished projects are often similar to the new project for which effort is to be estimated; yet, there is the risk that certain innovative

[*] This work is funded by the Greek Secretariat of Research and Technology, PAVET, action 4.3.1 under the contract of DIERGASIA project.

technologies may not be represented in the historical data [23]. This problem may be resolved by using public domain data sets that represent a variety of new technologies [13]. However cross-company (CC) data are not free of problems related to homogeneity, consistent collection and diversity of implementation practices and methods [15].

Several studies have addressed the problem of comparison of prediction accuracy between WC and CC data. Most of them have pointed out that CC models have similar [4], [5], [23] or even worse [14], [16], [18] estimation accuracy than the WC models. However, in all the previous comparative studies, the estimation methods applied provided only point estimates as output. A systematic review of these studies can be found in [13].

Various studies in software cost estimation [3], [11], [1], [20] suggest that the estimation of intervals is a more realistic approach, accounting for both uncertainty and risk. Usually interval estimates are created during the point estimation process by computing confidence intervals for the prediction [8], [1]. Another alternative is to predefine the intervals and then to use a model that predicts in which of the intervals the cost will fall [19]. This approach can also produce point estimates by computing the mean or the median point of the estimated interval [3].

The purpose of this study is to compare CC and WC interval estimation models. Two kinds of methods are applied, namely methods producing point estimates along with prediction intervals and methods producing interval estimates considering the mean or the median as a point estimate. The main research question addressed here is:

Which one of the models, cross-company or within-company, does estimate more accurately data coming from a single company?

The comparisons for testing this question are performed using statistical tests on the errors of the various models. The main contribution of this paper is the fact that the comparisons are taking into account not only the usual point estimates, but also the prediction intervals. In addition, our cross-validation approach uses an independent hold-out sample, i.e. a sample distinct from both WC and CC training sets.

For our study we used recent projects from the ISBSG data set, release 9 [9]. From these data, CC and WC projects were separated and 60 projects of the WC set were further left out to be used as a hold-out sample, as in [14], [16]. This dataset has been previously used in [17], where CC and WC models gave almost similar accuracy. However models were built using only stepwise regression and a different cross-validation approach to ours. The effort techniques used in this study to obtain point estimates are Estimation by Analogy (EbA) and Stepwise Regression (SR). Ordinal regression (OR), Classification and Regression Trees (CART) and Bayesian Networks (BN) are applied to obtain predefined intervals. The choice of techniques was motivated by several factors, such as: i) these techniques have all been previously used for effort estimation; ii) even the techniques that provide point estimates also provide confidence interval that can be used in our analysis; iii) authors' familiarity with such techniques. Other studies that utilized ISBSG releases ([6], [15]) resulted in lower prediction accuracy for the cross-company models.

The remainder of the paper is organized as follows: Sections 2 to 7 present the data set used, the accuracy measures, the methods and their results. In Section 8 we discuss our results and conclusions are given in Section 9. We have to note that due to the page limitation the methods are described briefly and only the most important results are presented.

2 Dataset and Accuracy Measures

The prediction methods in this paper were based on projects from Release 9 of the ISBSG database. In order to be consistent with a previous relevant study, we used the same dataset as in [17], which, after applying the analysis that is extensively presented there, contains 4 variables described in Table 1. The estimation methods used, model the relationship between a dependent variable and a set of predictor (independent) variables that can be either categorical (factors) or numerical (covariates). In order to apply the OR, CART and BN techniques on our training data, we had to categorize the effort variable from a ratio to an ordinal scale. For this reason, we divided effort (see Table 1) into four interval categories using as bounds of the intervals the quartiles of its empirical distribution. This in practice means that all categories have almost the same probability to contain the actual effort of a new project.

Table 1. Variables used in this study

Variable	Scale	Description
Effort	Ratio	Project effort in person hours (Categorized into the intervals:) WC: [0, 1029.5], (1029.5, 2353], (2353, 4746.25], > 4746.25 CC: [0, 492.75], (492.75, 1249.5], (1249.5, 3484.75], > 3484.75
Ufp	Ratio	Application size in unadjusted function points
LangType	Nominal	Language type (e.g. 3GL, 4GL)
DevType	Nominal	Describes whether the development was a new development, enhancement or re-development

Regarding the evaluation of the predictive accuracy for each of the estimation methods, the hold-out procedure was adopted. Specifically, 60 projects (33%) from the WC dataset (in total 184 projects) were drawn randomly for the generation of the test set. The remaining 124 (67%) projects constituted the training set for the fitting of the WC models. The entire CC dataset (672 projects) was used in order to fit the CC models.

The validation of cost models was based on the calculation, for each project in the hold-out sample, of the absolute error (AE), i.e. $\left|\text{actual effort} - \text{estimated effort}\right|$ and their median, denoted by MdAE. Other accuracy measures based on the well-known magnitude of relative error (MRE) were also calculated: The Mean Magnitude of Relative Error (MMRE), the Median Magnitude of Relative Error (MdMRE) and the Prediction at 25% level (Pred(25)). For the interval comparisons the hit-rate was calculated. This is simply the percentage of estimated intervals containing the actual effort.

Statistical tests for two related (or paired) samples were used to compare the prediction accuracy of WC and CC models on the hold-out sample. Specifically, the means of MREs (MMRE) were compared using parametric paired-samples t-test while the Wilcoxon Signed Ranks test was used for the comparison of the medians of MREs and absolute errors (MdMRE and MdAE). Regarding the Pred25 and the hit rate, we used the non-parametric McNemar procedure. All tests had level of significance $\alpha = 0.10$. A detailed description of all these tests can be found in numerous statistical textbooks, see for example [21]. The accuracy measures and the tests for their comparisons regarding the estimation of the 60 projects in the hold-out sample for all the WC and the CC models are given in Tables 2 and 3.

3 Estimation by Analogy

Estimation by Analogy (EbA) first finds for the new project the most similar projects (analogies) by evaluating a distance metric computed by the independent variables and then combines their effort to predict the effort of the new project. A detailed description of EbA and of the related methods for calibrating it, can be found in [1]. In order to select the appropriate number of analogies, we applied the jackknife (hold-one-out) technique from one up to ten analogies. We decided to use for the predictions, eight analogies for the WC data and ten analogies for the CC data, i.e. numbers that gave relatively reasonable results for the accuracy measures for each dataset.

It is desirable for all methods resulting in a “point estimation” to be accompanied by a confidence interval for this estimation. This is usually computed under specific assumptions for the underlying distributions. However, for EbA there is no such way to compute confidence intervals, but we can use a simulation technique, namely the non-parametric bootstrap [1]. The method is based entirely on the empirical distribution of the data set without any assumption on the population distribution.

4 Stepwise Regression

Stepwise regression is a statistical technique whereby a prediction model (Equation) is built and represents the relationship between independent and dependent variables. This technique builds the model by adding, at each stage, the independent variable with the highest association to the dependent variable [17].

Initially, the variables were checked to make sure the assumptions related to using this model would be satisfied. As a result, both *Effort* and *Ufp* were transformed to a natural logarithmic scale providing the new variables *lneffort* (dependent) and *lnufp* respectively. The stability of each regression model, was checked using residual plots showing residuals vs. fitted values to investigate if the residuals are random and normally distributed, and Cook’s Distance was also calculated to identify influential data points. Both models produce point estimates which are also accompanied by confidence intervals that were also compared.

The best WC model presented an adjusted R^2 of 0.368, thus explaining 36.8% of the variation in effort. The best CC model presented an adjusted R^2 of 0.591, thus explaining 59.1% of the variation in effort. The Equations read from the final model’s output, were transformed back to the raw data scale. The equation for WC and CC models are the following:

$$Effort = 118.51 \times Ufp^{0.620} \times e^{-0.475*Fourthgl} \; (WC\ model)\,. \quad (\mathbf{1})$$

$$Effort = 17.27 \times Ufp^{0.897} \quad (CC\ model)\,. \quad (\mathbf{2})$$

Where by *Fourthgl* we denote the variable resulting from *LangType* after denoting the ‘4gl’ language by 1 and the ‘3gl’ language as 0.

5 Ordinal Regression

The ordinal regression (OR) method is a generalization of the linear regression model. The main difference between the two methods is that the OR is used to model the relationship between an ordinal-dependent variable and a set of predictor variables that can be either categorical (factors) or numerical (covariates), whereas in the linear regression model the dependent variable must be continuous. For a detailed presentation of Ordinal regression (OR) we refer to [19]. Briefly, the ordinal logistic model has the general form:

$$l(c_j) = \theta_j - \sum_{i=1}^{k} \beta_i x_i \quad (\mathbf{3})$$

where c_j is the cumulative probability for the j_{th} ordinal category, θ_j is the threshold for the j_{th} category, $\beta_1,...,\beta_k$ are the regression coefficients, $x_1,...,x_k$ are the predictor variables and k is the number of predictors. The model predicts a transformation of the actual cumulative probabilities and the function $l()$ is called link function. In our analysis we used the logit function that has the form $\log(c/(1-c))$.

The equations of the ordinal regression for the WC and CC model are:

Within-company model

$$l(c_j) = \theta_j + 0.931 * \delta(LangType) + 0.002 * Ufp \qquad \text{for } j = 1,2,3,4$$

$$\delta(x) = \begin{cases} 1 \text{ if } x = 3GL \\ 0 \text{ if } x = 4GL \end{cases} \quad \theta_j = \begin{cases} 0.399 \text{ if } j = 1 \\ 1.656 \text{ if } j = 2 \\ 3.064 \text{ if } j = 3 \\ 0 \text{ if } j = 4 \end{cases} \quad (\mathbf{4})$$

Cross-company model

$$l(c_j) = \theta_j + 0.602 * \delta(LangType) + 0.006 * Ufp \qquad \text{for } j = 1,2,3,4 .$$

$$\delta(x) = \begin{cases} 1 \text{ if } x = 3GL \\ 0 \text{ if x} = 4GL \end{cases} \quad \theta_j = \begin{cases} 0.215 \text{ if } j = 1 \\ 1.631 \text{ if } j = 2 \\ 3.201 \text{ if } j = 3 \\ 0 \text{ if } j = 4 \end{cases} \qquad \textbf{(5)}$$

The dependent ordinal variable in our case has as values the predefined intervals given in Table 1. The predictor variables of the two models were selected based on their significances**.** As the dataset contains only three independent variables, we tried all the combinations in order to fit the WC and the CC model. After removing the variable *DevType* that was not significant for both of the models, we decided to keep the predictors *Ufp* and *LangType*.

6 Classification and Regression Trees

The CART model consists of an hierarchy of univariate binary decisions and for details we refer to [10]. The CART model for the classification of the WC projects is presented in Fig. 1a, and shows that the model classifies correctly 67.7% of the projects presenting very high functionality to the highest effort interval. The remaining projects are classified to the lowest effort interval with classification ratio 32.3%.

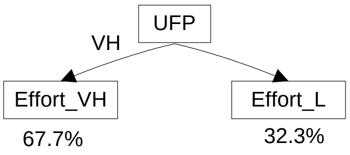

Fig. 1a. WC CART model

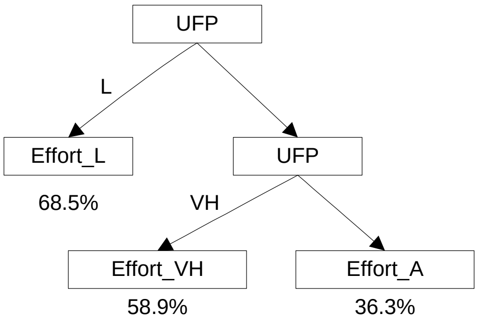

Fig.1b. CC CART model

The best fitting CART model used to classify the cross-company projects into an effort interval is presented in Fig. 1b, and suggests *Ufp* as the sole variable used to estimate the project effort values.

7 Bayesian networks

Bayesian Networks (BNs) are probabilistic models represented as directed acyclic graphs describing the causal relationship between variables. For details we refer to [7]. In this study the BNs built for the estimation of effort intervals use information only from the historical data and are extracted with the application of a BN software that can be found in [2].

As mentioned, BNs can be used both for knowledge representation and classification tasks. The conditional independencies along with the interrelations among the projects attributes are the same for the WC and CC data. The BN model for both data sets is presented in Fig. 2. According to this BN, *Effort* is conditionally independent from *DevType* when the value of *Ufp* and *LangType* is known. *LangType* also affects *Ufp*.

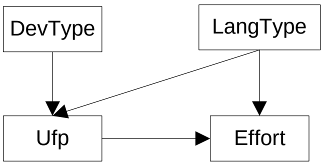

Fig. 2. BN for WC and CC data

In order to provide an estimate for the effort values and evaluate the models we use the General Bayesian Network classifiers [10].

8 Results and Discussion

As described in Section 1, the research question addressed in this study is as follows:

Which one of the models, WC or CC, does estimate more accurately data coming from a single company?

The accuracy metrics for all models and for all metrics are presented in Table 2 while their statistical comparisons are presented in Table 3. Stepwise and ordinal regression were the techniques that presented significant differences between WC and CC predictions, based on a common hold-out sample. In addition, these were the only two techniques for which there were significant differences between hit ratios of WC and CC models.

Except for CART (MdP), all results based on the MMRE showed that WC predictions were significantly superior to CC predictions; however these results are contradicted when using Pred(25) and also the absolute errors. The Hit rate measure also suggests that there were no differences between WC and CC predictions.

The comparison between CC and WC estimation results shows that when the models are evaluated with point-estimate measures based on relative error (MMRE, MdMRE), WC models present a significant difference in almost all methods. However, MdAE which is also a point estimation measure, but based on absolute error, shows no significant differences between WC and CC models. Also, when the models are evaluated with measures that indicate whether the estimation falls within a particular interval (pred(25), hitrate) the results show in most cases that there is no significant difference.

Given that the use of MRE to compare different prediction models is not unanimously agreed upon [12], in this study we prefer to rely upon the results using Pred(25) and MdAE, which all suggest that there were no differences between within- and cross-company predictions.

Table 2. Prediction accuracy of the test set for the WC and CC models for all methods,

Metrics+	MMRE	MdMRE	pred25	MdAE	Hitrate
EbA					
WC model	100%	64%	13%	1612.7	47%
CC model	178%	72%	22%	1919.1	50%
Stepwise Regression					
WC model	137.6%	66.6%	20%	1965.32	93.3%
CC model	95.6%	66%	20	1545.51	43.3%
Ordinal Regression					
WC (MP)	76.0%	59.1%	23%	892.2	53%
WC (MdP)	71.8%	55.6%	25%	901.0	
CC (MP)	133.2%	74.7%	13%	2285.5	37%
CC (MdP)	99.7%	71.5%	17%	1438.5	
CART model					
WC (MP)	88.7%	70%	21.7%	1742.1	45%
WC (MdP)	82.6%	68%	23.3%	1716.5	
CC (MP)	170.3%	66%	23.3%	2580.0	55%
CC (MdP)	118.0%	58%	26.7%	1607.0	
BN model					
WC (MP)	65.3%	45.0%	26.7%	817.66	56.7%
WC (MdP)	63.3%	43.4%	26.7%	781.5	
CC (MP)	179.7%	70.5%	16.7%	2389.92	46.7%
CC (MdP)	125.8%	68.8%	20.0%	1458.0	

PA – Predictive Accuracy MP – Mean Point

Perhaps an explanation for these results is the similarity of application domains between the WC and CC datasets. It seems that when more specific information cannot be used to select a tightly focused CC data set, the performance of a CC model will depend on how broadly similar the CC and WC projects are. Similar results using data sets of similar application domains have also been obtained in [4], [5].

In terms of the validity of these results, the dataset used does not characterize a random sample of projects and therefore the external validity of these results may be compromised. In addition, we also assume that *Ufp* is a reasonable software size measure [22](construct validity). As for the quality of the data used in our analysis, we only employed projects that were rated A or B, as these are considered suitable for sound analysis by the ISBSG (internal validity). This decision is commonly adopted by researchers in this field but leads to the consideration of a few project attributes, a fact that may hinder the generation of accurate results by the applied estimation methods.

Table 3. Significance for theWC and CC models for all methods

Method	MMRE /Paired -t	MdMRE/Wilc.	pred25/McN.	MdAE/Wilc.	Hitrate/Mc.N
EbA	0.002/Sig.	0.041/Sig.	0.267	0.162	0.845
SR	0.003/Sig.	0.049/Sig.	1.00	0.935	0.00/Sig.
OR (MP)	0.002/Sig.	0.008/Sig.	0.180	0.299	0.052/Sig.
OR (MdP)	0.025/Sig.	0.029/Sig.	0.405	0.195	
CART (MP)	0.036/Sig.	0.431	1.00	0.988	0.286
CART (MdP)	0.176	0.763	0.839	0.790	
BN. (MP)	0.01/Sig.	0.02/Sig.	0.263	0.129	0.327
BN. (MdP)	0.03/Sig.	0.08/Sig.	0.556	0.185	

PA: Predictive Accuracy MP: Mean Point McN.: McNemarn MdP: Median Point Wilc.: Wilcoxon

9 Conclusions

This paper investigated the estimation accuracy provided using cross-company and within-company data sets, by considering a new dimension of the problem, namely the generation of both point and interval estimates by a variety of methods. Estimation by analogy and stepwise regression were used to produce point estimates accompanied by confidence intervals, while ordinal regression, classification and regression trees and Bayesian networks were used to estimate effort within predefined intervals. Model generation was performed using cross-company (672 projects) and within-company (124 projects) training data sets, and model accuracy was assessed on the basis of a hold-out sample of 60 projects, selected from the initial within-company dataset.

Overall, our study provides evidence that, even when considering interval estimation, it appears that there is no clear distinction between accuracy obtained using CC and WC data. On the basis of this, practitioners might opt to work with both approaches, depending on the availability of in-house cost estimation data, and combining estimates in some way.

Further research is needed, using multiple within-company data sets to determine the influence of the application domain and other peculiarities of such data sets on estimation accuracy, seeking evidence of when one method might be preferable to the other. In addition, investigation of similarities and differences between the two data sets in an estimation situation may be an interesting topic to investigate, aiming to produce both point and interval estimates of higher accuracy than using each data set separately.

References

1. L. Angelis, I. Stamelos, A simulation tool for efficient analogy based cost estimation, Empirical Software Engineering, 5, (2000), pp. 35-68.
2. Bayesian Belief Network Software, http://www.cs.ualberta.ca/~jcheng/bnpc.htm
3. S. Bibi, I. Stamelos, L. Angelis: Software Cost Prediction with Predefined Interval Estimates, 1st Software Measurement European Forum, Rome, Italy, January 2004.

4. Briand, L.C., K. El-Emam, K. Maxwell, D. Surmann, I. Wieczorek. An assessment and comparison of common cost estimation models. Proceedings of the 21st International Conference on Software Engineering, ICSE 99, 1999, pp 313-322.
5. Briand, L.C., T. Langley, I. Wieczorek. A replicated assessment of common software cost estimation techniques. Proceedings of the 22nd International Conference on Software Engineering, ICSE 20, 2000, pp 377-386.
6. Jeffery, R., M. Ruhe and I. Wieczorek. Using public domain metrics to estimate software development effort. Proceedings Metrics'01, London, 2001, pp 16-27.
7. Jensen, F., "Bayesian Networks and Decision Graphs", Springer, Denmark, 2002.
8. M. Jorgensen, An effort prediction interval approach based on the empirical distribution of previous estimation accuracy, Information and Software Technology 45 (2003)123-126.
9. International Software Benchmarking Standards Group: www.isbsg.org
10. Hand, D., Mannila, H. and Smyth, P., "Principles of Data Mining", MIT Press, US, 2001.
11. B.A. Kitchenham, S. Linkman , Estimates, uncertainty and risk, IEEE software14(3) ,1997, pp 69-74.
12. B.A. Kitchenham, L.M. Pickard, S.G. MacDonell, and M.J. Shepperd. What accuracy statistics really measure, IEE Proc. - Software Engineering, 2001, 148(3), June.
13. Kitchenham, B., Mendes, E. & Travassos, G. H. (2006), Systematic Review of Cross- vs. Within-Company Cost Estimation Studies, in 'Proceedings of EASE 2006: Evaluation & Assessment in Software Engineering', BCS-eWIC, pp. 89–98.
14. Lefley, M., and M.J. Shepperd, Using Genetic Programming to Improve Software Effort Estimation Based on General Data Sets, Proceedings of GECCO 2003, LNCS 2724, Springer-Verlag, pp 2477-2487, 2003.
15. Chris Lokan, Emilia Mendes: Cross-company and single-company effort models using the ISBSG database: a further replicated study. ISESE 2006: 75-84
16. Maxwell, K., L.V. Wassenhove, and S. Dutta, Performance Evaluation of General and Company Specific Models in Software Development Effort Estimation, Management Science, 45(6), June, pp 787-803, 1999.
17. Mendes, E., C. Lokan, R. Harrison, and C. Triggs, A Replicated Comparison of Cross-company and Within company Effort Estimation models using the ISBSG Database, in Proceedings of Metrics'05, Como, 2005
18. Mendes, E. and B.A. Kitchenham, Further Comparison of Cross-Company and Within Company Effort Estimation Models for Web Applications. Proceedings Metrics'04, Chicago, Illinois September 11-17th 2004, IEEE Computer Society, pp 348-357, 2004.
19. P. Sentas, L. Angelis, I. Stamelos, G. Bleris, Software productivity and effort prediction with ordinal regression, Information and Software Technology 47: (2005) pp. 17-29.
20. Stamelos, L. Angelis, P. Dimou, E. Sakellaris, On the use of Bayesian belief networks for the prediction of software productivity, Information and Software Technology 45 (2003) 51-60
21. Sheskin, D.J., 2004. Handbook of Parametric and Nonparametric Statistical Procedures. Third Edition Chapman & HAL/CRC.
22. Vilet, H., Software Engineering: Priciples and Practice, Wiley and Sons, 1997,UK
23. Wieczorek, I. and M. Ruhe. How valuable is company-specific data compared to multi-company data for software cost estimation? .Proceedings Metrics'02, Ottawa, June 2002, pp 237-246.