Building Software Cost Estimation Models using Homogenous Data

Rahul Premraj · Thomas Zimmermann
Saarland University
Saarbrücken, Germany
{premraj · zimmerth}@cs.uni-sb.de

Abstract

Several studies have been conducted to determine if company-specific cost models deliver better prediction accuracy than cross-company cost models. However, mixed results have left the question still open for further investigation. We suspect this to be a consequence of heterogenous data used to build cross-company cost models. In this paper, we build cross-company cost models using homogenous data by grouping projects by their business sector. Our results suggest that it is worth to train models using only homogenous data rather than all projects available.

1. Introduction

In the last decade, company-specific cost models have received some attention to determine if they deliver better prediction accuracy than cross-company cost models. Unfortunately, mixed results from studies conducted to date have left the question still open [9]. In contrast to previous work, in this paper we investigate if cost models built using homogenous cross-company project data (i.e., grouped by business sectors) can yield improved prediction accuracy.

Software cost estimation is inherently a challenging task. This is due to numerous factors such as lack of understanding of software processes and their impact on project schedule, constantly changing technologies, and the human factor, which adds substantial variability in productivity. Thus, despite over four decades of on-going research, we continue to witness novel approaches being presented to improve cost estimation accuracy.

One such proposed approach by Maxwell et al. [12] was to build company-specific cost models using data originating from a single company. Such data is likely to be more homogenous and reflect the company's strengths and weaknesses, and in turn, lead to better models. Thereafter, several follow-up studies and replications (discussed in Section 2) were conducted to verify, if indeed, models based on in-house data perform better than those built on data gathered across several companies.

Irrespective of the lack of consensus from different studies, a challenging prerequisite for building company-specific cost models is availability of in-house project data. This might be a constraint for several companies for a variety of reasons such as:

- New companies may be yet to begin implementing projects to extract and record data.
- Recorded data from the past may be irrelevant to estimate costs using current technologies.
- Companies may choose not to collect data for they lack incentives to do so.
- Companies may not know what data to collect and how to do so.

What alternatives do such companies have? Clearly, they have to resort to other sources of data to base their own cost estimates on. For this, it is vital that the data originates from comparable companies undertaking similar projects; else the estimates may be considerably off mark. Natural choices in such a scenario are companies from the same business sector. They likely operate within similar environments that expose them to comparable cost factors, and in turn, comparable development productivity [15] and costs.

Using significant differences in productivity across business as a motivation, we seek to verify if building cost models from homogenous companies delivers better results. Within this large goal, the objectives of our study are:

(a) To develop company-specific cost models for comparisons against other models **[OB1]**.

(b) To develop cross-company cost models to compare their prediction accuracy against company-specific cost models **[OB2]**.

(c) To develop business-specific cost models to compare their prediction accuracy against company-specific and cross-company cost models **[OB3]**.

(d) To develop business-specific cost models to determine if they can be used by companies from other business sectors **[OB4]**.

COMPUTER SOCIETY

The road-map of the paper is as follows: Section 2 discusses related work. Next, the data set used for our study is discussed in Section 3. Our design of experiments is presented in Section 4. Sections 5 and 6 present our results from building company-specific cost models and cross-company cost models respectively using our data (OB1 and OB2). Results from business-specific cost models (OB3) are presented in Section 7, while results from cross-business cost models (OB4) are reported in Section 8. We discuss threats to validity in Section 9, and conclude in Section 10.

2 Related Work

Several studies have been previously conducted to determine if company-specific cost models are superior to cross-company cost models in terms of prediction accuracy. Unfortunately, a systematic review [9] on this topic showed that the results are too mixed to conclude that either of the two models is significantly better than the other.

Table 1 presents ten studies dating from 1999 that compare cost models across companies. Of these, four studies concluded that company-specific cost models are 'not' significantly better than cross-company cost models. All remaining six studies concluded the opposite, i.e. company-specific cost models perform better than cross-company models. However, the last two studies in Table 1 ([12, 10]) conducted no statistical significance tests to arrive at their respective conclusions. Hence, in agreement with Kitchenham and Mendes [9], we report the two studies as inconclusive. Table 1 makes clear the lack of compelling evidence to support claims in favour of either of the two investigated cost models.

The systematic review by Kitchenham and Mendes [9] went further to investigate if quality control for data collection could influence the outcome of the studies. No evidence to suggest the same was found. However, a bias was noticed in studies favouring company-specific cost models since they employed smaller data sets with lower maximum effort values. Also, differences in experimental design and accuracy measures made drawing conclusions in the systematic review challenging.

Clearly, comparable and replicable experiments are required to further investigate the cause for discrepencies in results across different studies. A likely explanation for mixed results is disregarding other factors that influence development costs. One such factor is the business sector that the company operates in. Premraj et al. [15] showed substantial differences in development productivity across business sectors, where the manufacturing sector was nearly three times more productive than banking and insurance sectors. Hence, it is reasonable to assume that differences in productivity (and those from other factors) may distort cross-company cost models and give overall mixed results.

Table 1. Previous Studies and their Model Preferences (adapted from [9])

Model Preference	Study	Data Set
Cross-company	[3]	Laturi
	[4]	ESA
	[17]	Laturi
	[14]	ISBSG
Company-specific	[6]	Megatec and ISBSG
	[7]	ISBSG
	[8]	Tukutuku
	[13]	Tukutuku
Inconclusive	[12]	ESA
	[10]	Laturi

In contrast to the above studies, we take the business sector explicitly into account in the cost model, motivated by a statistically justification [15] for its inclusion, and investigate the influence it casts on prediction accuracy. For this, we use the 'Finnish Data Set' that has previously never been used for studies on this subject.

3 The Finnish Data Set

The 'Experience Pro Data', commonly referred to as the 'Finnish Data', has been used in this study. This section provides a brief background on the data set and furnishes the data cleansing steps undertaken to make it suitable for analysis.

The 'Finnish Data' is the result of a commercial initiative by Software Technology Transfer Finland (STTF) [1] to support the software industry, especially to benchmark software costs, development productivity and software processes. So far, constituent projects originate only from Finland. To have access to the data, companies are subjected to an annual fee, which may be partially waived in proportion to the number of submitted projects by the company. These projects are carefully assessed and graded for quality by experts at STTF, and then committed to the data set.

Since data collection is an on-going process, the data set has been growing over the years. For this study, the version of the data available as of March 2005 is used. We refer to it as *Finnish788* since it comprises 788 projects. The constituent projects span from a wide breadth of business sectors such as banking, insurance, public administration, telecommunications and retail. It includes both, new development projects (88%) and maintenance projects (remaining 12%). Other fields that characterise the projects include effort (in hours), project size (in FiSMA FP), development environment, staff pertaining metrics and many more. In total, over 100 different variables are collected, but a major-

ity of them are difficult to analyse given the ratio of missing values. Maxwell and Forselius [11] provide a fuller description of the data.

It is obvious that a data set of such scale is bound to have noise that may derive misleading results. To counter this, we removed suspect projects from the data set to increase our confidence on the results. Projects that met the following criteria were removed:

- Projects graded as 'X' (i.e. poor or unreliable) for data quality.

- Projects with 0 points for accuracy of size and effort data.

- Projects with size units other than FiSMA FP (for homogeneous comparisons).

- Projects marked as maintenance. These were removed for two reasons. First, they are characterised using a different set of variables than new development projects. Second, previously Premraj et al. [15] showed that these projects exhibit productivity trends different to new development projects and hence, their inclusion may distort results.

- Projects with delivery rates of less than 1 or greater than 30 hours/FP. These implausible values of productivity were determined in discussion with staff from STTF. The rationale here is that projects with pronounced productivity values may be influenced by unreported or misreported factors.

- Projects with business sectors other than Banking, Insurance, Manufacturing, Public administration, Retail, and Telecommunications. Projects from other business sectors were too few in number and could not be meaningfully analysed.

- Projects with any T-variable[1] recorded as −1 or *null*, which symbolises unreported, erroneous or missing values.

The edited data set comprised 395 projects, which we refer to as *Finnish395*. For replication purposes, please note that the same number of projects must be derived from *Finnish788*, irrespective of the order in which projects are removed using the above itemised criteria.

Table 2 reports basic summary statistics for both data sets. Here, effort is recorded in person hours, size in FiSMA FP (a variant of function points) and raw productivity as a ratio of size to effort (i.e. FiSMA FP per person hour). Clearly, the distributions for *Finnish788* indicate the presence of influencing outliers for all three variables. The same is true for *Finnish395*, although to a lesser extent partly due to our data cleaning process.

4 Experimental Design

This section describes the methodology adopted for our experiments. To recall, the broad objective is to predict project *effort* (dependent variable) using independent variables. We build the cost models using linear regression and then, evaluate and compare their performance using selected accuracy measures.

Although *Finnish395* allowed over 30 variables to be included in the model as independents, we chose to use *size* alone for two reasons.[2] First, during a pilot of this study, we used forward stepwise regression that selected *size* as the most important factor and only two other development environment variables as independents; however, the latter added little to the model. Secondly, the complexity of building a regression model using so many variables of mixed data types would have drifted the focus of the paper away from its intended objectives. Our intention was to keep the models easy to understand and comparable to each other.

One assumption behind building linear regression models is that the included variables must be approximately normally distributed. A Kolmogorov-Smirnov test (a one sample test to check if a variable is normally distributed) on both, *effort* and *size* revealed that neither were normally distributed. Hence, the two variables were transformed into their natural logarithms, which were then normally distributed. The linear regression model took the form $ln(effort) = \alpha + \beta \, ln(size)$, where α is the constant and β is the regression coefficient. When the model is transformed back into its original form, it converts to $effort = \alpha \, size^{\beta}$.

This model has the added advantage of investigating, whether economies of scale exist, i.e., lesser effort is expended to build an additional size unit of software (function points in our case). This is exhibited when $\beta < 1$, since an increase of one unit in size would increase effort less than proportionally. Likewise, for $\beta > 1$ an increase by one unit in size causes effort to increase more than proportionally, thus suggesting diseconomies of scale. Lastly, when $\beta = 1$, effort simply increases as by a multiple of *size* (α), suggesting constant returns to scale.

Another matter of importance are extreme values or outliers in the data set that can cast strong influence on the regression parameters (α and β). These must be identified and removed to increase the model's goodness-of-fit (R^2). To identify such outliers, we computed Cook's distance (D) for each project in the data used to build the model. Then, those projects with $D > 3/4n$, where n is the number of data points, were removed. Thereafter, the model was rebuilt without the outliers. All results reported in the paper are derived from the latter model.

Accuracy of the regression models was measured using

---

[1]These are 21 ordinal variables that characterise the development environment, such as functionality requirements, usability requirements, analysis skills of staff and their team skills.

[2]Our data cleaning was conducted keeping further possible experiments in mind, using the same data.

Table 2. Summary Statistics for Finnish788 and Finnish395

Variable	*Finnish788*				*Finnish395*			
	Mean	Median	Min	Max	Mean	Median	Min	Max
Effort (person hours)	3754	1551	55	67576	4444	2000	55	67576
Project Size (FiSMA FP)	584	293	6	13682	584	321	18	9390
Productivity = Size/Effort (FiSMA FP/hour)	0.22	0.16	0.001	2.67	0.20	0.15	0.033	0.92

- The percentage Pred(x) of predictions that lie within $\pm x\%$ of the actual effort value (for ($x = 25$ and $x = 50$)

- Mean Magnitude of Relative Error *(MMRE):*

$$MMRE = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{|ActualEffort_i - PredictedEffort_i|}{ActualEffort_i} \right)$$

- and Median Magnitude of Relative Error *(MdMRE)* for each experiment.

It is noteworthy that relative error is largely accepted as a biased statistic [5]. However, currently no universally acceptable substitute is known to us and hence, we compute it for our study. Additionally, we wish to make our results comparable to studies in Section 2, since nearly all used the same accuracy measures. In Tables 3, 4, 5, and 6, we present our results using these accuracy measures. However, this has been done visually to ease comparing performance of the different models.

For each experiment, we paired Pred(25) with Pred(50) and *MMRE* with *MdMRE*. Then, each pair was plotted as a bar scaled from 0 to 1. The bars are interpreted as follows — for Pred(x), longer black bars represent higher accuracy and vice versa for *MMRE* and *MdMRE*.

	Pred(25); Pred(50)	*MMRE*; *MdMRE*
Low Accuracy		
Avg. Accuracy		
High Accuracy		

For readers interested in actual values, corresponding results in tables have been provided in the Appendix section.

The objectives of our study require different training and testing sets to be extracted to build the cost models. These are illustrated in Fig. 1 and are described below. Note that the enumerations (i.e., (a), (b), etc.) are mapped with those of the objectives in Section 1 and the subfigures in Fig. 1.

(a) To achieve **OB1**, we first built company-specific cost models using project data originating from a single company to predict its own projects. The models were built and evaluated using the leave-one-out method **[Fig. 1(a)]**.
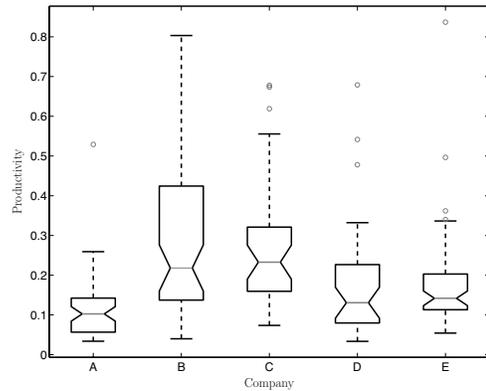
Figure 2. Company Productivity Comparison

(b) To achieve **OB2**, we built cost models using all data points from *Finnish395* except those originating from a company, say Company X. Then the effort of each project from Company X was predicted using the model **[Fig. 1(b)]**.

(c) To achieve **OB3**, we built cost models using data from exclusive business sectors, but leaving out projects from Company X. Then, individual projects from Company X were predicted for using the corresponding cost models of business sectors to which they belong **[Fig. 1(c)]**.

(d) To achieve **OB4**, we built cost models for each business sector exclusively and then predicted for projects from one business sector, say Business X using cost models from every other business sector other than Business X **[Fig. 1(d)]**.

In addition to making comparable models, we wished to do the same for the results. For this, it is vital that the test set remains unchanged (except for **OB4**). We made this possible by selecting five different companies from *Finnish395* with 35 or more projects that exclusively comprised five different test sets. Thus, administering the different test sets to any of the above models would result in the same number of predictions, making like-for-like comparisons possible. Productivity (defined as the ratio of effort to size) comparisons across companies is plotted in Fig. 2 where we can observe statistically significant differences.
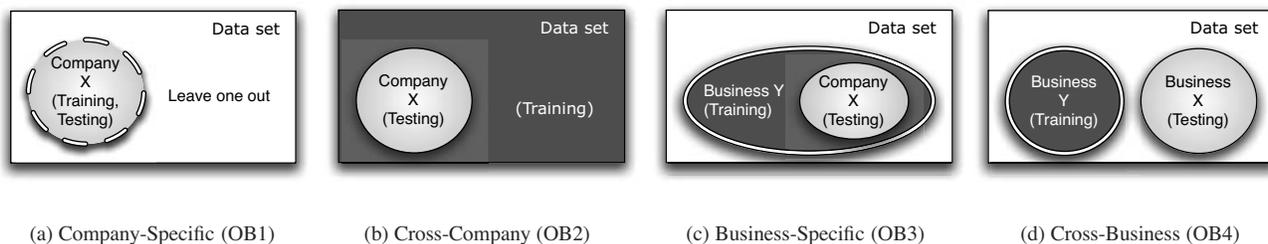
(a) Company-Specific (OB1)   (b) Cross-Company (OB2)   (c) Business-Specific (OB3)   (d) Cross-Business (OB4)

Figure 1. Training and Testing Sets for Different Cost Models

Table 3. Company-Specific Cost Models

Company	β	Pred(25); Pred(50)	*MMRE*; *MdMRE*
A	0.96		
B	0.98		
C	0.87		
D	1.15		
E	1.24		
		▷ better	better ◁

5 Company-Specific Cost Models

To meet our first objective (OB1), we built company-specific cost models to predict their own projects using the leave-out-out approach. Our results from these experiments are presented in Table 3 (and Table 7 in the Appendix).

It appears that company-specific cost models do not fare well in all cases. The Pred(x) values for companies A, B, and C indicate very weak predictive power, which is also reflected by their *MMRE* and *MdMRE*. On the other hand, cost models for companies D and E, especially for the latter, perform very well. Respective *MMRE* and *MdMRE* values for the two companies are also comparably low, suggesting a better model fit.

With regards to economies of scale, companies A and B showed marginal increasing returns to scale, while in the case of company C, the returns to scale were even higher. On the other hand, Companies D and E showed diseconomies of scale. Such results partly shed light upon why in some studies, cross-company models fail to perform as well as company-specific models. Data drawn from companies differing in their returns to scale may build poor models rendering them ineffective for prediction purposes. Banker et al. [2] noted that economies of scale prevails for small software projects, while diseconomies of scale exist when projects get larger in size. Investigating this case for our data set is beyond the scope of this paper since it involves building different models and deviates us from the objectives of our research.

6 Cross-Company Cost Models

Our second objective (OB2) was to investigate if cross-company cost models perform comparably to company-specific cost models. Table 4 sums up our results, while corresponding values are shown in Table 8 in the Appendix. We see a modest improvement in the Pred(x) values for Company A, while those for companies B and C remain unchanged. The same is true for *MMRE* and *MdMRE*, which changed negligibly too for all three companies. However, a marked difference in performance can be observed in the cases of companies D and E. The cross-company model appears to have considerably poor predictive power for these two companies in comparison to the company-specific model. The Pred(25) values for both fell to 0, while their Pred(50) values declined appreciably as well.

Our results suggest that cross-company cost models perform either comparably or worse than company-specific cost models, except in the case for Company A. To confirm this, we conducted a Kruskal-Wallis test[3] of significance (at $\alpha = 0.05$) on the residuals derived using the two models on each company. The test's results showed no significant differences in the residuals' distribution for all, except for Company E. For the latter, the company-specific model performed significantly better than the cross-company model.

We also performed the statistical test on the *Magnitude of Relative Error MRE* for different companies using the two models. Contrary to the residuals, all comparisons were significantly different with an exception of Company B, thus painting a slightly different picture. From this we learn that although, the residuals using the two models may be comparable to each other, overall prediction accuracy is better when using company-specific cost models.

For this section, it is not meaningful to discuss economies of scale in depth because the training data came from a large number of companies (i.e. the complement of the test data from *Finnish395*). But on the whole, we can observe marginally increasing or constant returns to scale.

---

[3]A non-parametric test comparing the medians of two distributions to determine if they come from the same population.

Table 4. Cross-Company Cost Models

Company	β	Pred(25); Pred(50)	MMRE; MdMRE
A	1.00		
B	0.97		
C	0.96		
D	0.97		
E	0.95		
		▷ better	better ◁

Table 5. Business-Specific Cost Models

Company	β	Pred(25); Pred(50)	MMRE; MdMRE
A	0.89		
B	1.08		
C	1.10		
D	–		
E	–		
		▷ better	better ◁

Perhaps, this is an artefact of data from the different companies negating each other's returns to scale.

7 Business-Specific Cost Models

Our results from building business-specific cost models (OB3) are reported in Table 5 (see also Table 9 in the Appendix). Again, like the previous two models, this one does not exhibit any consistent prediction accuracy pattern. But on the whole, it is more comparable to company-specific cost models.

The Pred(x) values for Company A are 0, while the MMRE and MdMRE are the poorest amongst all three models. But all four accuracy measures indicate that business-specific cost models work best for companies B and C. In their case, while Pred(x) values were the highest, the MMRE and MdMRE were the lowest amongst the three models examined so far. In the case of both, Companies D and E, accuracy measures show an improvement in performance in comparison to cross-company cost models, but not comparable to company-specific cost models.

Again, we conducted the Kruskal-Wallis test of significance to compare residuals from business-specific cost models with company-specific cost models and cross-company cost models on the individual company data. It turned out that none of the comparisons were significantly different from one another.

However, in the case of comparing MRE, all predictions using business-specific cost models were significantly different from company-specific cost models, with an exception of Company D, while the same is true for cross-company cost models with an exception for Company E. Thus, again while residuals may not differ much from one another, there appears to be an overall difference in prediction accuracy.

Interestingly, we see a pattern comparable to the above two models, when examining economies of scale. Company A belonged to the Public Admin. sector which shows increasing returns to scale, while Banking and Insurance sectors (for Companies B and C respectively) exhibit slight diseconomies of scale. For Companies D and E, we report no β values since the projects came from two or three business

sectors. In their case, projects were predicted using models trained with data from the respective business sectors.

8 Cross-Business Cost Models

The last objective (OB4) of this paper was to investigate if cost models built using data from one business sector can be used to predict costs of projects from other sectors. Our results are shown in Table 6, where the business models in the rows are the source of training data to build the cost models, while the columns represent sectors that provide the testing data. The upper part of Table 6 shows the values for Pred(25) and Pred(50) and the lower part the values for MMRE and MdMRE.

Clearly, cross-business cost models perform poorly in general. The models trained from the Banking, Insurance, Public Admin., and Telecommunication sectors give poor Pred(x) values for all business sectors, including themselves. The same deduction can be drawn from the respective MMRE and MdMRE values, which are considerably high. On the contrary, Manufacturing and Retail sector models in comparison perform more favourably, and especially well for themselves giving high Pred(x) values and relatively lower residuals.

It is noteworthy that they also can predict with reasonable accuracy for the Telecommunications sector that fails to predict itself. The Telecommunication sector models consistently performed very poorly across all other sectors having Pred(x) always 0% and MMRE and MdMRE consistently over 95%.

On the whole, our results do not strongly support our hypothesis that some business sectors can be gainfully used for cost modelling for other sectors. However, the Manufacturing and Retail sectors stand out from the rest since their cost models appear to have higher predictive power than the others. We suspect that our results are an artefact of the productivity distribution across different business sectors and number of projects in each.

9 Threats to Validity

As any empirical study, ours has threats to validity, too.

Table 6. Cross-Business Cost Models.

Pred(25); Pred(50)		Business Sector for Testing						
		Banking	Insurance	Manufacturing	Public Admin.	Retail	Telecomm.	Training Sector
Banking	▷							
Insurance	▷							
Manufacturing	▷							
Public Admin.	▷							
Retail	▷							
Telecomm.	▷							

MMRE; *MdMRE*		Business Sector for Testing						
		Banking	Insurance	Manufacturing	Public Admin.	Retail	Telecomm.	Training Sector
Banking	◁							
Insurance	◁							
Manufacturing	◁							
Public Admin.	◁							
Retail	◁							
Telecomm.	◁							

Threats to external validity. These threats are pertaining to the generalisations we can draw from our study. This study was conducted using a sample of only Finnish projects which certainly do not represent global development trends. All the more, these projects may not even represent the population of software projects in Finland.

Threats to internal validity. These threats pertain to our experimental procedure that might affect our results. Firstly, our data cleaning process removed of nearly half the number of projects in *Finnish788*. We believe that this was a necessary evil to have more faith in our results. Secondly, we only used project size as an independent variable. With over 30 other variables to choose from, better modelling techniques can be developed to accommodate more such variables to improve prediction accuracy.

Hence, the findings from our study are local to this very data set and needs further in-depth analysis to draw broader conclusions.

10 Conclusions and Future Work

Many studies have been conducted to determine if company-specific cost models deliver more accurate predictions in comparison to cross-company cost models. Unfortunately, mixed results have constrained the community from drawing any conclusions yet. We suspected the use of heterogenous data to build cross-company cost models, partly to be a cause for their poor performance.

Our hypothesis in this study was that building cost models using more homogenous data would deliver better results. For this, we built cross-company cost models using data from companies belonging to one business sector exclusively. These models were then compared against company-specific cost models and general cross-company cost models.

We found that while none of the models in our study performed consistently well for all test data, company-specific cost models appeared to slightly outperform the others. Business-specific cost models seemed to perform comparably to company-specific cost models, but better than cross-company cost models. Unfortunately, residuals from different models did not differ significantly, however their respective *MRE*s did. Hence, our inference from the results is based purely on the reported accuracy measures.

Thus, while the question of 'which model is better' continues to remain open, there is evidence that researchers should build cross-company cost models using more homogenous data. Homogenous subsets can be extracted from a larger data sets using expert judgement or statistical analysis. A clearer picture of this topic can perhaps be painted by replicating previous studies where cross-company data is handled more systematically, rather than as available.

Small and new software companies can substantially benefit from the lessons learnt from our research. When lacking in-house project data, our results suggest that companies must prefer procuring data from sources that are comparable to themselves. Such data is likely to yield better estimations and in turn, better decisions.

In the future, we aim to build cost models using more

features that could possibly contribute towards an improved prediction accuracy. Alternative modelling techniques could also shed more light on consistency of performance, one of them being Case-Based Reasoning [16] in which predictions are made using a local neighbourhood rather than generalised models. But more importantly, there is a need to devise a protocol for empirical studies to make them comparable to each other.

References

[1] STTF - http://www.sttf.fi.

[2] R. D. Banker, H. Chang, and C. F. Kemerer. Evidence on economies of scale in software development. *Information and Software Technology*, 36(5):275–282, 1994.

[3] L. C. Briand, K. E. Emam, D. Surmann, I. Wieczorek, and K. Maxwell. An assessment and comparison of common software cost estimation modeling techniques. In *ICSE 1999*, pages 313–322, Los Angeles, May 1999. ACM.

[4] L. C. Briand, T. Langley, and I. Wieczorek. A replicated assessment of common software cost estimation techniques. In *ICSE 2000*, pages 377–386, Limerick, June 2000. ACM.

[5] T. Foss, E. Stensrud, B. Kitchenham, and I. Myrveit. A simulation study of the model evaluation criterion mmre. *IEEE Trans. on Softw. Engg.*, 29(11):985–995, Nov. 2003.

[6] R. Jeffery, M. Ruhe, and I. Wieczorek. A comparative study of two software development cost modeling techniques using multi-organizational and company-specific data. *Info. and Softw. Technology*, 42(14):1009–1016, Nov. 2000.

[7] R. Jeffery, M. Ruhe, and I. Wieczorek. Using public domain metrics to estimate software development effort. In *METRICS 2001*, page 16, London, Apr. 2001. IEEE.

[8] B. Kitchenham and E. Mendes. A comparison of cross-company and within-company effort estimation models for web applications. In *EASE 2004*, pages 47–55, Edinburgh, May 2004.

[9] B. Kitchenham, E. Mendes, and G. H. Travassos. A systematic review of cross- vs. within-company cost estimation studies. In *EASE 2006*, Keele, Apr. 2006.

[10] M. Lefley and M. Shepperd. Using genetic programming to improve software estimation based on general datasets. In *Procs. of Genetic and Evolutionary Computation Conference (GECCO)*, pages 2477–2487. Springer-Verlag, 2003.

[11] K. Maxwell and P. Forselius. Benchmarking software development productivity. *IEEE Software*, 17(1):80–88, Jan./Feb. 2000.

[12] K. Maxwell, L. V. Wassenhove, and S. Dutta. Performance evaluation of general and company specific models in software development effort estimation. *Mgmt. Science*, 45(6):787–803, June 1999.

[13] E. Mendes and B. Kitchenham. Further comparison of cross-company and within-company effort estimation models for web applications. In *METRICS 2004*, pages 348–357, Chicago, Sept. 2004. IEEE.

[14] E. Mendes, C. Lokan, R. Harrison, and C. Triggs. A replicated comparison of cross-company and within-company effort estimation models using the isbsg database. In *METRICS 2005*, page 36, Como, Sept. 2005. IEEE.

[15] R. Premraj, M. Shepperd, B. Kitchenham, and P. Forselius. An empirical analysis of software productivity over time. In *METRICS 2005*, page 37, Como, Sept. 2005. IEEE.

[16] M. Shepperd and C. Schofield. Estimating software project effort using analogies. *IEEE Trans. on Softw. Engg.*, 23(11):736–743, 1997.

[17] I. Wieczorek and M. Ruhe. How valuable is company-specific data compared to multi-company data for software cost estimation? In *METRICS 2002*, page 237, Ottawa, June 2002. IEEE.

Appendix

Table 7. Company-Specific Cost Models

Company	Pred(25)	Pred(50)	*MMRE*	*MdMRE*
A	0.00%	8.33%	77.48%	83.04%
B	0.00%	1.72%	84.35%	87.38%
C	0.00%	0.00%	84.35%	87.38%
D	5.71%	22.86%	65.32%	71.04%
E	45.71%	74.29%	37.85%	34.50%

Table 8. Cross-Company Cost Models

Company	Pred(25)	Pred(50)	*MMRE*	*MdMRE*
A	3.33%	20.00%	70.63%	77.91%
B	0.00%	1.72%	85.02%	88.02%
C	0.00%	0.00%	90.89%	91.67%
D	0.00%	2.86%	85.55%	88.77%
E	0.00%	8.57%	78.42%	82.77%

Table 9. Business-Specific Cost Models

Company	Pred(25)	Pred(50)	*MMRE*	*MdMRE*
A	0.00%	0.00%	85.26%	89.33%
B	0.00%	8.62%	72.87%	78.03%
C	1.85%	5.56%	78.39%	81.53%
D	0.00%	5.71%	82.98%	86.96%
E	28.57%	40.00%	51.89%	100.00%