Cross- vs. Within-Company Cost Estimation Studies: A Systematic Review

Barbara A. KITCHENHAM, Member IEEE Computer Society[1], Emilia MENDES, and Guilherme H. TRAVASSOS

Abstract

OBJECTIVE – The objective of this paper is to determine under what circumstances individual organisations would be able to rely on cross-company-based estimation models.

METHOD – We performed a systematic review of studies that compared predictions from cross-company models with predictions from within-company models based on analysis of project data.

RESULTS – Ten papers compared cross- and within-company estimation models, however, only seven presented independent results. Of those seven, three found that cross-company models were not significantly different to within-company models; four found that cross-company models were significantly worse than within-company models. Experimental procedures used by the studies differed making it impossible to undertake formal meta-analysis of the results. The main trend distinguishing study results was that studies with small within-company data sets (i.e. <20 projects) that used leave-one-out cross-validation all found that the within-company model was significantly different (better) to the cross-company model.

CONCLUSIONS – The results of this review are inconclusive. It is clear that some organisations would be ill-served by cross-company models whereas others would benefit. Further studies are needed, but they must be independent (i.e. based on different data bases or at least different single company data sets) and should address specific hypotheses concerning the conditions that would favour cross-company or within-company models. In addition, experimenters need to standardise their experimental procedures to enable formal meta-analysis, and recommendations are made in part III.

Index Terms- ***Cost Estimation, Management, Systematic Review, Software Engineering***

I. Introduction

Early studies of cost estimation models (e.g. [12] [8]) suggested that general-purpose models such as COCOMO [1] and SLIM [24] needed to be calibrated to specific companies before they could be used effectively. Taking this result further and following the proposals made by DeMarco [4], Kok et al. [14] suggested that cost estimation models should be developed only from single-company data. However, three main problems can occur when relying on within-company data sets [3], [2]:

[1] Manuscript received, 2006.

Barbara A. Kitchenham is with Keele University, Keele, Staffordshire, U.K. and National ICT Australia Ltd. Sydney, Australia, (e-mail: Barbara.Kitchenham@nicta.com.au

Emilia Mendes is with Computer Science Department, Private Bag 92019, The University of Auckland. Auckland, New Zealand (e-mail: emilia@cs.auckland.ac.nz)

Guilherme H. Travassos is with UFRJ/COPPE, Systems Engineering and Computer Science Program, Caixa Postal 68511, 21941-972 Rio de Janeiro – RJ, Brazil, (e-mail: ght@cos.ufrj.br)

1. The time required to accumulate enough data on past projects from a single company may be prohibitive.
2. By the time the data set is large enough to be of use, technologies used by the company may have changed, and older projects may no longer be representative of current practices.
3. Care is necessary as data needs to be collected in a consistent manner.

These problems motivated the use of cross-company models (models built using cross-company data sets, which are datasets containing data from several companies) for effort estimation and productivity benchmarking, and subsequently several studies compared the prediction accuracy between cross- and within-company models. In 1999, Maxwell et al. [18] analysed a cross-company benchmarking database by comparing the accuracy of a within-company cost model with the accuracy of a cross-company cost model. They claimed that the within-company model was more accurate than the cross-company model, based on the same hold-out sample. In the same year, Briand et al. [2] found that cross-company models could be as accurate as within-company models. The following year, Briand et al. [3], re-analysed the data set employed by Maxwell et al. [18] and concluded that cross-company models were as good as within-company models. Two years later, Wieczorek and Ruhe [26] confirmed this same trend using the same data set employed by [2]. Three years later, Mendes et al. [20] also confirmed the same trend using yet another data set.

These results seemed to contradict the results of the earlier studies and pave the way for improved estimation methods for companies who did not have their own project data. However, other researchers found less encouraging results. Jeffery and his co-authors undertook two studies, both of which suggested that within-company models were superior to cross-company models ([6], [7]). Two years later, Lefley and Shepperd claimed that the within-company model

was more accurate than the cross-company model, using the same data set employed by Wieczorek and Ruhe [26] and Briand et al. [2]. Finally, a year later Kitchenham and Mendes undertook two studies of Web-based projects ([11], [19]). In both studies, a within-company model was significantly better than a cross-company model.

Given the importance of knowing whether or not it is reasonable to use cross-company estimation models to predict effort for single company projects, we conducted a systematic review in order to determine factors that influence the outcome of studies comparing within- and cross-company models. In addition, we also discuss the different variations in study protocol, i.e. experimental procedure. The main aim of our systematic review is to assist software companies with small data sets in deciding whether or not to use an estimation model obtained from a benchmarking data set. The secondary aim is to provide advice to researchers intending to investigate the potential value of cross-company models.

The results of this systematic review, for our research questions 1 and 2, have been reported previously [13]. In this paper, we also provide the results for research questions 1 and 2, however we detail further our quality evaluation process and we also present the results for our third research question.

The paper is organised as follows: Section 2 describes the systematic review, followed by the presentation of its results in Section 3. Section 4 discusses the results and threats to their validity, followed by conclusions and comments on future work in Section 5.

II. METHOD

A. Introduction

A systematic review is a method that enables the evaluation and interpretation of all accessible research relevant to a research question, subject matter, or event of interest [10], [23]. There are

many reasons for carrying out a systematic literature review, amongst which the most common are:

- To review the existing evidence regarding a treatment of technology, for example, to review existing empirical evidence of the benefits and limitations of a specific Web development method.
- To identify gaps in the existing research that will lead to topics for further investigation.
- To provide a context/framework so as to properly place new research activities.

A systematic review generally comprises the following steps [10], [22]:

- Identification of the need for carrying out a systematic review;
- Formulation of a focused review question;
- A comprehensive, exhaustive search for primary studies;
- Quality assessment of included studies;
- Identification of the data needed to answer the research question;
- Data extraction;
- Summary and synthesis of study results (possibly including formal meta-analysis);
- Interpretation of the results to determine their applicability;
- Report-writing.

Prior to the review, it is desirable to develop a protocol that specifies the plan that the systematic review will follow to identify, assess and collate evidence.

Advice from the medical domain suggests that a well-formulated question generally has four parts [22], identified as PICO (Population, Intervention, Comparison, Outcome):

- The population (e.g. the disease group, or a spectrum of the healthy population);
- The study factor (e.g. the intervention, diagnostic test, or exposure);

- The comparison intervention (if applicable);
- The outcome.

The question should be sufficiently broad to allow examination of variation in the study factors and across populations.

B. Research Questions, Population, Intervention

Within the context of this paper we have carried out a systematic literature review using the basic approach identified in [10], in order to examine studies comparing within- and cross-company models from the point of view of the following research questions:

- Question 1: What evidence is there that cross-company estimation models are not significantly different from within-company estimation models for predicting effort for software/Web projects?
- Question 2: What characteristics of the study data sets and the data analysis methods used in the study affect the outcome of within- and cross-company effort estimation accuracy studies?

Since all the studies used different experimental procedures, we also had one secondary research question:

- Question 3: Which experimental procedure is most appropriate for studies comparing within- and cross-company estimation models?

Some studies also compared prediction accuracy of different prediction techniques, and we intended to investigate this issue. However, Mair and Shepperd [17] have recently undertaken a systematic review of all studies that compared regression and analogy-based techniques of which the studies in this systematic review are a subset.

Our population was that of cross-company benchmarking data bases of software projects, and Web projects, and our intervention included effort estimation models constructed from cross-company data, used to predict single company project effort. The comparison intervention was represented by effort estimation models constructed from single company data only. The studies' outcomes that were of interest to our systematic review were the accuracy of the estimates/predictions made using the within- and cross-company models.

C. Search Strategy used for Primary Studies

The search terms used in our Systematic Review were constructed using the following steps:

1. Derive major terms from the questions by identifying the population, intervention and outcome;
2. Identify alternative spellings and synonyms for major terms;
3. Check the keywords in any relevant papers we already have;
4. Use the Boolean OR to incorporate alternative spellings and synonyms;
5. Use the Boolean AND to link the major terms from population, intervention and outcome.

The main search terms are:

Population: software, Web, project.

Intervention: cross-company, project, effort, estimation, model.

Comparison: single-company, project, effort, estimation, model

Outcomes: prediction, estimate, accuracy**.**

The complete set of search strings is presented in the Appendix and all the intermediate steps that lead to it are detailed in [13]. Whenever a database did not allow the use of complex Boolean search strings we designed different search strings for each of these data bases. The search strings were piloted and results documented (see [13]).

D. Search Process

Our search process was organised into two separate phases: *Initial* and *Secondary*. The *Initial* search phase identified candidate primary sources based on our own knowledge and searches of electronic databases using the derived search strings. The electronic searches were based on six electronic databases and seven individual journals and conference proceedings chosen because they had published articles we already knew about.

Compared to our original search process [13] we have extended our search to cover the years 1990-1998 and 2005 to November 2006. This time the full search string could be used for all searches except the ACM. All ten known papers were found after searching the 13 different sources. No new relevant papers were found. 1,344 papers were retrieved, of which 25 represented the set of ten known relevant papers (the same papers were retrieved by several search engines). We examined all 1,344 papers using a manual scan of titles and, if unsure, we also read the abstracts. However, we did not measure inter-rater agreement.

E. The Secondary search phase

The *Secondary* search phase had two sub-phases: i) to review the references of each of the primary sources identified in the *first* phase looking for any other candidate primary sources. This process was to be repeated until no further reports/papers seemed relevant; ii) to contact researchers who authored the primary sources in the *first* phase, or who we believe could be working on the topic. Six researchers were contacted and no one was working on the topic either directly, or via supervision of MSc/PhD students (see [13]).

F. Study Selection Criteria and Procedures for Including and Excluding Primary Studies

The criteria for including a primary study comprised any study that compared predictions of cross-company models with within-company models *based on analysis of single company*

project data. We excluded studies where projects were only collected from a small number of different sources (e.g. 2 or 3 companies), and where models derived from a within-company data set were compared with predictions from a general cost estimation model. The list of selected studies is shown in Table 1.

Table 1 Authors' and sources

Authors	Study ID	Year	Reference (source)
Maxwell, K., L.V. Wassenhove, and S. Dutta	S1	1999	[18]
Briand, L.C., K. El-Emam, K. Maxwell, D. Surmann, I. Wieczorek[2]	S2	1999	[2]
Briand, L.C., T. Langley, I. Wieczorek	S3	2000	[3]
Jeffery, R., .M. Ruhe and I. Wieczorek	S4	2000	[6]
Jeffery, R., M. Ruhe and I. Wieczorek	S5	2001	[7]
Wieczorek, I. and M. Ruhe.	S6	2002	[26]
Lefley, Martin and Shepperd, Martin, J.	S7	2003	[15]
Kitchenham, B.A., and E. Mendes.	S8	2004	[11]
Mendes, E. and B.A. Kitchenham.	S9	2004	[19]
Mendes, E., C. Lokan, R. Harrison, C. Triggs	S10	2005	[20]

G. Study Quality Assessment Checklists

The criteria used to determine the overall quality of the primary studies was split into two parts (see Table 2). Part I considered the quality of the study itself and Part II the quality of the reporting provided [23]. Although we attributed different weights to Part I (weight=1.5) and Part II (weight =1) we also report the final scores considering equal weights. Part I has four top-level questions and an additional quality issue related to the size of the within-company data set:

- Less than 10 projects: Poor quality (score = 0)
- Between 10 and 20 projects: Fair quality (score = 0.33)
- Between 21 and 40 projects: Good quality (score = 0.67)
- More than 40 projects: Excellent quality (score = 1)

Whenever a study used more than one within-company data set, the average score was used.

[2] Briand et al. referenced a technical report on which the conference paper was based.

The size of the within-company data set was considered as part of the study quality criteria because we expected that larger within-company data sets would lead to more reliable comparisons between within- and cross-company models. General statistical principles (and power analysis) favour large data sets over small data sets. However, this principle presupposes that the data set is a sample from a homogenous distribution. If we sample from a heterogeneous population, large and small samples will be equally "messy" (e.g. exhibiting multiple modes, or an unstable mean and variance).

Part II has four top-level questions. For both parts, top-level questions without sub-questions were answered Yes/No, corresponding to scores 1, and 0 respectively. Whenever a top-level question had sub-questions, scores were attributed to each sub-question such that the overall score for the top-level question would range between 1 and 0. For example, question 1 had two sub-questions, thus each “Yes”, and “No” for a sub-question contributed scores of 0.5, and 0 respectively. The overall quality score for a paper for Part I, after applying a weight of 1.5, ranged from 0 to 7.5, representing very poor and excellent quality, respectively. The overall quality score for a paper for Part II ranged from 0 to 4, representing very poor and excellent quality, respectively. Therefore, using weighted scores, the overall quality score for a paper ranged from 0 to 11.5, and with equal weights, from 0 to 9. The quality data extraction was performed as part of the overall data extraction process and used the same process to ensure that data extraction was accurate.

The quality criteria was employed in our investigation in two different ways: First, as an overall score to ensure that results were not largely confounded with quality; Second, as a source of moderator values to investigate systematic differences between studies.

We did not include as part of our study quality assessment any criterion related to the quality of the estimation models because our aim was to assess the study itself, not the accuracy of prediction models it used. We took the view that a model's poor accuracy should not be used to determine a study's quality, even if such models are not appropriate for practical use. Furthermore, even if model accuracy is poor, it may be useful to a company if it is more accurate than their current method.

Table 2 Quality scores

Questions	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
Part I										
1. Is the data analysis process appropriate?										
1.1 Was the data investigated to identify outliers and to assess distributional properties before analysis?	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
1.2 Was the result of the investigation used appropriately to transform the data and select appropriate data points?	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
2. Did studies carry out a sensitivity or residual analysis?										
2.1 Were the resulting estimation models subject to sensitivity or residual analysis?	0.5	0	0	0.5	0	0	0	0.5	0.5	0.5
2.2 Was the result of the sensitivity or residual analysis used to remove abnormal data points if necessary?	0.5	0	0	0.5	0	0	0	0.5	0.5	0.5
3. Were accuracy statistics based on the raw data scale?	1	1	1	1	1	1	0	1	1	1
4. How good was the study comparison method?										
4.1 Was the single company selected at random (not selected for convenience) from several different companies?	0	0	0	0	0	0	0	0	0	0
4.2 Was the comparison based on an independent hold out sample (0.5) or random subsets (0.33), leave-one-out (0.17), no hold out (0)	0.5	0.33	0.33	0.17	0.17	0.17	0.5	0.17	0.17	0.33
5. Size WC data set	0.67	1	0.67	0.33	0.33	0.44	1	0.33	0.33	1
Total Part I	4.17	3.33	3.0	3.5	2.5	2.61	2.5	3.5	3.5	4.33
Part II										
1. Is it clear what projects were used to construct each model?	1	1	1	1	1	1	1	1	1	1
2. Is it clear how accuracy was measured?	1	1	1	1	1	1	1	1	1	1
3. Is it clear what cross-validation method was used?	1	1	1	1	1	1	1	1	1	1
4. Were all model construction methods fully defined (tools and methods used)?	1	1	1	1	1	1	1	1	1	1
Total Part II	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0
Total primary study using weighted scores	10.26	9.0	8.5	9.25	7.75	7.92	7.75	9.25	9.25	10.5
Total primary study using unweighted scores	8.17	7.33	7.0	7.5	6.5	6.61	6.5	7.5	7.5	8.33

In addition to the quality scores given to each primary study, we have also documented, as part of our data extraction, other quality problems with some of these studies, as follows:

- Study 1: Does not penalise prediction models if they are unable to provide any prediction. For example, the single company model was only able to make predictions for 6 out of the 9 projects, so they based their accuracy statistics on the 6 projects for which they had

estimates. However, if we consider the pred(25) statistic, it might be more accurate to say 3 out of 9 projects had estimates within 25% of the actual rather than 3 out of 6. The first approach would give a pred(25) of 33%, the second approach gives 50% as reported by Maxwell. Counting non-predicted projects as not being within 25% of the actual may give an appropriate penalty for failing to estimate when calculating pred(25) but it is not clear how to make an adjustment for the correlation coefficient or MMRE. One approach might be to assign the value 0 to any missing prediction but then when calculating MMRE, projects with a large actual value would be penalised more than projects with a small actual value.

- Study 2: Does not explicitly say if the data used for the cross- and within-company models was transformed, however the final model was not linear. It does not present the equations, resulting from regression analysis, for either cross- or within-company models.
- Study 3: Does not present the equations, resulting from regression analysis, for either cross- or within-company models.
- Study 5: The p-values for the comparisons of the model construction techniques look internally inconsistent. It was not clear what variables were used to build each model, how R^2 was calculated for non-regression methods, what function points methods were considered, exactly what quality rating was used to select projects, and the detailed criteria used to merge some variables (e.g. organization type);
- Study 7: The paper seems to not have carried out a fair comparison with OLS because the most appropriate transformation (i.e. logarithmic transformation) was not used.

If we rank primary studies, according to their overall quality score, using weighted and unweighted scores we have the same ranking, which is as follows: (highest score) S10, S1, S4, S8, S9, S2, S3, S6, S5, S7 (lowest score). We note that overall the quality of the studies was

good. The worst scoring paper achieved a weighted quality score that was 67% of the maximum score and the best scoring paper achieved a score which was 91% of the maximum. Many factors did not vary between papers. Factors that varied between papers were the size of the within-company data set, the method used to make predictions, and the performance of sensitivity analyses. The first two factors were used in the subsequent analysis of factors affecting the outcome of the primary studies.

H. Data Extraction Strategy

Required Data

In addition to the study quality checklist, the following data were extracted for each primary study:

- General: data Extractor; data checker; study identifier; application domain; name of database; number of projects in database (including within-company projects); number of cross-company projects; number of projects in within-company data set; size metric(s): (FP (Yes/No); version used: , LOC (Yes/No); version used:, others (Yes/No); number:); number of companies; number of countries represented; quality control: (were quality controls applied to data collection?, if quality control, please describe); how was accuracy measured?
- Cross-company & Within-company models: what technique(s) was used to construct the model?; what transformations if any were used?; what variables were included in the model?; what cross-validation method was used?; underlying relationship between predictors and effort (linear, non-linear); (only for cross-company models: was the cross-company model compared to a baseline to check if it was better than chance?; what was/were the measure(s) used as benchmark?);

- Comparison: what was the accuracy obtained using the cross-company model?; what was the accuracy obtained using the within-company model?; what measure was used to check the statistical significance of the prediction accuracy (e.g. absolute residuals)?; what statistical tests were used to compare the results?; what were the results of the tests?
- Data summary: data base summary (all projects) for size and effort metrics; cross-company data summary for size and effort metrics; within-company data summary for size and effort metrics.

Data Extraction Process

For each paper a reviewer was nominated at random as data extractor, checker, or adjudicator. The data extractor reads the paper and completes the form; the checker reads the paper and checks that the form is correct. If there is a disagreement in the extracted data between extractor and checker that cannot be resolved, the adjudicator reads the paper and makes the final decision after discussions with the extractor and checker. Roles were assigned at random with the following restrictions:

1. No one should be data extractor on a paper they authored.
2. All reviewers should have an equal work load (as far as possible).

Extracted data was held in tables, one file per paper. After the extracted data was checked a single file containing the final agreed data was constructed. We did not calculate inter-rater agreement statistics, since our process was intended to achieve 100% agreement[3].

III. RESULTS

The summary data used to answer research questions 1, 2 are presented in Tables 3, and 4 respectively; summary data used to answer research question 3 are presented in Table 5. All results are discussed below.

[3] When we could not understand what was reported in the primary study, we approached the authors for clarification.

Question 1: What evidence is there that cross-company estimation models are not significantly different from within-company estimation models for predicting effort for software/Web projects?

Table 3 lists information about the database, the basis for prediction (cross-validation methods), and statistical tests used by each study reporting MMRE, Pred(25) and MdMRE statistics. When several different estimation models were used we report the best accuracy values and corresponding estimation technique. The studies are organised into three groups: studies that reported cross-company models were not significantly different from within-company models; studies that reported cross-company models were significantly worse than within-company models; and finally studies that did not undertake formal statistical testing. There were no studies that reported cross-company models were significantly better than within-company models.

We included prediction accuracy measures for the best model in each study in order to indicate direction of effect. In Table 3, we have indicated in bold the value that is better when the cross-company and within-company statistics are compared for a specific accuracy statistic. For studies where the within-company model was significantly better than the cross-company model, all the accuracy statistics are better for the within-company models than the cross-company models. For the other studies, the accuracy statistics sometimes disagreed. For studies where the cross-company model was not significantly different from the within-company models (S2, S3, S6, S10), the three studies that reported only one cross-company model all had MdMRE values that favoured the within-company models. For S6 four of the six reported MdMRE values favoured the within-company model. For S2 and S10, one of the two MMRE values favoured the cross-company model, and one of the two Pred(25) values favoured the within-company model.

For the remaining two studies (S1 and S7), in both cases the MMRE favoured the cross-company models while the one case that reported Pred(25) favoured the within-company model.

Thus, we prefer to use statistical significance as the criterion to decide how studies were categorised rather than a simple comparison of accuracy values, because otherwise:

- We have to rely on accuracy statistics, which are known to be biased [5].
- We have to make arbitrary distinctions between model accuracy values, for example for S10 the pred(25) for the cross-company model is 20.6 and for the within-company model is 20.8; however, with such a small difference it does not seem sensible to assert that the cross-company model is better than the within-company model.
- We would need additional criteria to decide how to categorise studies when the different accuracy statistics gave conflicting results.

Also, since none of the studies published the absolute residuals, or MREs, we were unable to look at distributions of the MRE values. Neither could we compute statistics such as Cohen's d in order to measure the size of effect between models.

Of the four studies that reported cross-company models were not significantly different from within-company models, S6 cannot be considered an independent study since it used the same data set employed in S2. Although it compared six within-company estimation models, one of the single companies was the same as that used in S2, the others were companies whose data was used to construct the cross-company model in S2. Thus, S6 does not offer additional *independent* evidence to the evidence provided by S2.

The remaining six studies all claimed that cross-company models were less accurate than within-company models; however, unlike S4, S5, S8, and S9, S1 and S7 did not test the statistical significance of their results, so we regard their results as inconclusive and suggest that

they are not used to provide supporting evidence. Furthermore, S1 used the same data set and single company as S3, and S7 used the same data set and single company as S2. Thus, even if S1 and S7 had performed statistical tests they would not have provided any additional independent evidence. Studies S8 and S9 provide independent evidence because the within-company data set used in S9 was not part of the Tukutuku dataset used in S8. It was collected from a single company some time after the first set of data had been collected and analysed. The datasets that we believe do not offer any independent evidence are greyed out in Table 3. This table also shows that the basis for evaluating predictive accuracy varied. Some studies used independent hold-out samples; others used different types of cross-validation (e.g. 3-fold, 20-fold, leave-one-out cross-validation). In addition, some studies based their statistical tests on the absolute (magnitude) relative error (MRE) while others used the absolute residuals. These differences made it impossible to perform any formal meta-analysis of the primary study results.

The four studies that showed significantly better predictions for within-company models all used a leave-one-out cross-validation. This type of cross-validation uses a single project as the validation set, which, in our view, biases positively towards the within-company data. Their within-company data sets were quite small; however they could have used a leave-two-out cross-validation, using a random selection of pairs. Finally, studies S1 and S7 used an independent hold-out sample, which seems an interesting novel approach given that the validation set is a completely separate data set from the data sets used to build the within- and cross-company models. However, as implemented, the approach introduced several complications. S1 suffered from using models that could only predict a subset of the already small hold-out sample. S7 incorporated the 48 single-company projects not used in the hold-out sample as part of the data set used to construct the cross-company model.

Table 3 Summary of evidence concerning the accuracy of cross-company models

Study	DB	Basis for Predictions[4]	Statistical tests comparing Within (WC) to Cross-company (CC)	Cross-company predictions			Within-company predictions		
				MMRE	Pred(25)	MdMRE	MMRE	Pred(25)	MdMRE
Cross-company model not significantly different to within-company model									
S2	Laturi	6-fold cross-validation (doesn't say what split)	Wilcoxon matched pairs test on MREs, inferred that split used was such that pairing was adequate	**CART+SWR: 52.4%**	CART+SWR: 29%	CART: 46%	CART: 56.9%	**CART: 29%**	**SWR: 41%**
S3	ESA	3-fold cross-validation (doesn't say what split)	Wilcoxon matched pairs test on MREs, inferred that split used was such that pairing was adequate			OLS: 32%			**ANOVA_e: 26%**
S6	Laturi	5 different leave-one-out cross-validations (one for each WC data set), and 1 randomly selected test sets	Wilcoxon matched pairs test. Measure used is unknown			C1: Analogy: 46% C2: **ANOVA: 13%** C3: Analogy: 32% C4: OLS: 30% C5: **Analogy: 31%** C6: ANOVA: 30%			C1: **Analogy: 39%** C2: Analogy: 20% C3: **Analogy: 22%** C4: **Analogy: 25%** C5: Analogy: 32% C6: **OLS: 26%**
S10	ISBSG	20-fold cross-validation (62 projects in validation set)	Mann-Whitney test 2 independent samples on absolute residuals	SWR: 123%	**SWR: 20.6%**	SWR: 61%	**SWR: 102%**	SWR: 20.8%	**SWR: 60%**
Cross-company model significantly different to within-company model									
S4	Megatec and ISBSG	19-fold cross-validation (1 project validation set)	Wilcoxon matched pairs test on MREs (OLS comparison, p<0.05)	OLS: 61%	ACE-2 no SA: 16%	OLS: 38%	**OLS: 37%**	**OLS: 47%**	**OLS: 27%**
S5	ISBSG	12-fold cross-validation (1 project validation set)	Wilcoxon matched pairs test on MREs			ROR: 63.8%			**CART_p: 17.8%**
S8	Tukutuku	13-fold cross-validation (1 project validation set)	Wilcoxon matched pairs test and paired t-test on absolute residuals (p<,0.05)	MSWR: 56.5%	MSWR: 30.8%	MSWR: 44.4%	**MSWR: 24.5%**	**MSWR: 53.8%**	**MSWR: 23.4%**
S9	Tukutuku	14-fold cross-validation (1 project validation set)	Wilcoxon matched pairs test on absolute residuals (p<0.05)	SWR CCM2: 93%	SWR CCM1: 14.3% CCM2: 7.1%	SWR CCM2: 61% A2s CCM1: 93%	**SWR: 38%**	**SWR: 28.6%**	**SWR: 38%**
Inconclusive									
S1	ESA	Independent hold-out (9 projects)	Correlation analysis between actual and estimate, no formal statistical significance test	**GLM: 36% (4 pjs)**	GLM: 25% 11.1% (adjusted for missing predictions		GLM: 65%	**GLM: 50% 33% (adjusted for missing predictions**	
S7	Laturi	Independent hold-out (15 projects)	No formal statistical significance test	**GP: 37.67%**			GP: 37.96%		

Stepwise Regression (**SWR**); Manual Stepwise Regression (**MSWR**); Robust Regression (**ROR**); Ordinary Least-Squares Regression (**OLS**)
Analogy -> Estimation by Analogy, or Case-based reasoning
Projects (**pjs**); Project (**pj**)
Stepwise ANOVA using effort as dependent variable (**ANOVA_e**); Stepwise ANOVA using productivity as dependent variable (**ANOVA_p**)
ACE-2 no SA: uses the average of the two most similar analogues for effort prediction and does not apply any size adjustment.
Genetic Programming (**GP**)
General Linear Model (**GLM**)
Cross-company model fitted without the within-company data (**CCM1**)
Cross-company model fitted with the within-company data (**CCM2**)

[4] Cross-validation for within-company model

Question 2: What characteristics of the study data sets and the data analysis methods used in the study affect the outcome of within- and cross-company effort estimation accuracy studies?

Table 4 reports the values of a variety of study related factors that we believed might have influenced the findings of the primary studies. One of these factors is the quality control of data, where previous studies [2], [11] have hypothesized that studies where the cross-company databases applied quality controls on data collection were those that found cross-company models as good as within-company models. However, study S10 contradicts this view. Furthermore, studies S3 and S1 take a rather different view of the effectiveness of the quality control applied to the projects in the ESA dataset. Maxwell et al. (S1) say "Another limitation, shared by any multi-company database, is that it is extremely difficult to ensure that each company understands each question in the same way. We can attempt to validate answers in a telephone conversation but this will never be as exact as the data that could be obtained in a specific company where one person is in charge of measuring and collecting the data for software development projects." This implies that Maxwell et al. were not convinced that the quality control on data collection was as reliable as Briand et al. (S3) suggest when they say "Once a project questionnaire is filled out, each supplier is contacted to ensure the validity and comparability of the data." Thus, for studies that found cross-company models not significantly different to within-company models we have:

- One database (Laturi) where researches agree that stringent quality control is applied to data collection.
- One database (ESA) where researchers disagree as to the stringency of the quality controls applied to data collection.

- One database (ISBSG) where researchers agree that no quality controls are applied to data collection.

We therefore conclude that quality controls on data collection cannot ensure that cross-company models perform as well as within-company models.

Our quality evaluation of the studies shows no consistent evidence that the quality of the studies influences the results. The scores for two of studies favouring cross-company models (S2, and S3) are lower than that for three studies favouring within-company models (S4, S8, and S9), but S10, which favoured cross-company models, has the highest quality score. This means that the outcome of studies is not confounded with overall quality.

In relation to the number of projects used in the cross-company model (see Table 4) there is a slight difference between studies S2, S3, S10 (median = 131), and studies S4, S5, S8, S9 (median = 99); however there is a more noticeable difference when we compare the number of projects in the within-company models: the median for S2, S3, S10 is 63, whereas the median for S4, S5, S8, S9 is 14. In fact, all the studies where within-company predictions were significantly better than cross-company predictions used small within-company data sets of fair quality.

The number of projects available from a single company may be an indication of the size of the company and the homogeneity of the data set. Certainly the single company in study S4 had about 50 employees [25] and the single companies in studies S8 and S9 had considerably fewer employees (4 and 5 respectively). The single companies in S8 and S9 are specialist Web-application development companies, with S9 specialising in small enhancement projects, so these within-company datasets are relatively homogeneous. In contrast, Stathis [25] reported that the single company in S4 (Megatec) produced a variety of different projects for different clients (9 of the projects were database projects but other projects varied in type). We have no more

detailed information concerning the single company in study S5. Overall, we have no information about company size for the projects reported in the ESA, Laturi or ISBSG data sets.

We can also compare the results for the three studies that used the ISBSG data set as the basis of cross-company models. S4 and S5 both found the within-company model was superior to the cross-company model. In both these cases, the range of project effort values for the within-company data was small relative to the range of values for the cross-company projects, e.g. the maximum effort for S4 single company projects was 23.2% of the maximum effort of the cross-company projects, for S5 the single company maximum was 2.1% of the cross-company maximum. In contrast, S10 found no significant difference between the within- and cross-company models. For S10 the range of effort values for the single company projects was much closer to the range for the cross-company projects e.g. the maximum effort for single company projects was 77.9% of the maximum for the cross-company projects.

We can also contrast the single companies in studies S8 and S9 that used the Tukutuku cross-company dataset. The S8 single company project effort ranged from 21 to 1786 hours and the S9 single company project effort ranged from 7 to 148 hours, compared with a range of 6 to 5000 hours for the cross-company projects. Both S8 and S9 found that the within-company model was more accurate than the cross-company model, although the single company data set in S8 was much more similar to the cross-company database than the single company data set in S9. However, the *cross-company* model for S8 was more accurate than the baseline (median) model, whereas this was not the case for S9. The comparison of studies that used ISBSG and Tukutuku suggest that the greater the difference between the within- and cross-company projects the less likely it is that the cross-company model will provide accurate predictions for the single company projects.

No clear patterns were observed for the size metrics used, nor for the procedure used to build the within-company model. S2, S3, S10, S4, S5, and one of the models in study S9 (CMM1) all built models independently; however, studies S8 and S9 (model CMM2) fit a generic cross-company model to select variables applicable to both within- and cross-company models. The use of a generic model was motivated by the existence of several size metrics, and the fact that, if fitted independently, there may be occasions when a cross-company model cannot be applied to the within-company model, as the variables do not match.

Finally, there is no clear indication that the strength of the cross-company relationship is a major factor in determining whether cross-company prediction models are as good as within-company models. The MdMRE for cross-company models were 46%, 32% and 61% for studies S2, S3 and S10; 38%, 63.8%, 44.4% for studies S4, S5, S8, and (61% and 93%) for study S9. Thus, there is no compelling evidence that the cross-company models are more accurate for studies that found cross-company models not to be significantly different from within-company models.

Question 3: Which experimental procedure is most appropriate for studies comparing within- and cross-company estimation models?

We found a large variation in the procedures adopted by different primary studies. In this section, we discuss issues and provide guidelines related to conducting comparative studies between within- and cross-company models.

Tables 5a, 5b and 5c identify a variety of options for performing a comparative study of cross- and within-company estimation models. We consider the pros and cons of each option, and identify which primary study (if any) used that option.

Table 4 Study related factors

Study	Quality control on data collection (Database)	Weighted Quality Score	Number of projects in database (Number used in CC model)	Number of projects in WC	Range of Effort values (converted to person hours)	Size Metric	Was WC model built independently of the CC model	CC data summary (without WC data)	WC data summary	Underlying relationship bt. predictors and effort for CC model	Underlying relationship bt. predictors and effort for WC model
Cross-company models not significantly different to within-company models											
S2	Yes (Laturi)	9.0	206 (56)	63	Min: 480 Max: 63694	Unadjusted Experience Function Points	Yes	ND	Size: (Min: 48; Max: 3634; Mean: 671.4; St.dev.: 777.3) Effort: (Min: 583; Max: 63694; Mean: 8109.54; St.dev.: 10453.9)	Non-linear	ND
S3	Yes (ESA)	8.5	166 (131)	29	Min: 3 Max: 627984	Adjusted KLOC	Yes	ND	Size: (Min: 10.5; Max: 732; Mean: 133.97; St.dev.: 174.15) Effort: (Min: 11.1; Max: 4361; Mean: 558.97; St.dev.: 1063.81)	Non-linear[5]	Non-linear[6]
S6	Yes (Laturi)	7.92	206 (206 – WC size)	63, 13,12, 11,10,10	Min: 250 Max: 63694	Unadjusted Function Points	Yes	ND	C1 (63 pjs) Effort: (Min: 583; Max:63694; Mean:8110; Median:5100) Size: (Min:48; Max:3634; Mean:671; Median:387 C2 (13 pjs) Effort: (Min: 480; Max: 6030; Mean:2426; Median:1979) Size: (Min:219; Max:1613; Mean:593; Median:370) C3 (12 pjs) Effort: (Min: 780; Max: 24788; Mean:14546; Median:15800) Size: (Min:189; Max:2155; Mean:1215; Median:1138) C4 (11 pjs) Effort: (Min: 918; Max: 51100; Mean:1050; Median:6290) Size: (Min:129; Max:2105; Mean:693; Median:546) C5 (10 projects) Effort: (Min: 592; Max: 17745; Mean:5220; Median:4182) Size: (Min:137; Max:1619; Mean:528; Median:422) C6 (10 pjs) Effort: (Min: 1330; Max: 26670; Mean:10922; Median:8649) Size: (Min:176; Max:1364; Mean:804; Median:707)	Non-linear	Non-linear
S10	No (ISBSG)	10.5	872(680)	187	Min: 14 Max: 73920	IFPUG Function Points	Yes	Size: (Min: 3; Max: 809; Mean: 292; Median: 118) Effort: (Min: 14; Max: 73920; Mean: 2710;Median: 1249)	Size: (Min: 16; Max: 6294; Mean: 587.5; Median 293.5) Effort: (Min: 140; Max: 57587; Mean: 4706.5; Median: 2418)	Non-linear	Non-linear
Cross-company models significantly different to within-company models											
S4	No(ISBSG), Yes (Megatec)	9.25	451 (145)	19	Isbsg: Min: 10 Max: 59809 Megatec: Min: 194 Max: 13905	Unadjusted Function Points	Yes	Size: (Min: 11 Max: 9803 Mean: 761) Effort:(Min: 10, Max: 59809 Mean: 761)	Size:(Min: 39; Max: 3290; Mean: 506; St. Dev.: 818) Effort:(Min: 194; Max: 13905; Mean: 1947; St. Dev.: 3115)	Non-linear	Non-linear
S5	No (ISBSG)	7,75	324 (310)	14	Min: 97; Max:59809	Function Points	Yes	ND	Effort: Min: 170; Max: 1238; Mean: 560; Median: 568 Size: Min: 56; Max: 579; Mean: 256; Median: 267	Non-linear	linear
S8	No (Tukutuku)	9.25	53 (40)	13	Min:6	23 different	Not completely	ND (values extracted	ND (values extracted subsequently)	Non-linear	Non-linear

					Max:5000	size measures		subsequently) Size(Web pages) Min: 3 Max: 2500, Mean: 216.07 Effort: Min: 6 Max: 5000 Mean: 434.6	Size (web pages): Min: 7, Max:440, Min: 124.8 Effort: Min 21 Max: 1786, Mean: 354.2		
S9	No (Tukutuku)	9.25	67(53)	14	Min:6 Max:5000	9 different size measures	Yes (CCM1) No (CCM2)	Size: (Min: 3; Max: 2000; Mean: 150.6; Median: 36) Effort:(Min: 6; Max: 5000; Mean: 414.8; Median: 105)	Size (web pages): Min: 1; Max: 86; Mean: 31.5; Median: 12 Effort: Min: 7; Max: 178; Mean: 44.6; Median: 25.5	Non-linear	Non-linear
Inconclusive											
S1	Partially (ESA)	10.26	108 (60)	29	Min: 1123.2 Max: 627984	KLOC	Yes	ND	ND	Non-linear	Non-linear
S7	Yes (Laturi)	7.75	407(149)	48	Not provided	Function points (Laturi variant)	No, CC used 48 WC projects	ND	ND	Linear	Linear
WC–Within-company CC–Cross-company CCM1-Cross-company model fitted without the within-company data CCM2-Cross-company model fitted with the within-company data ND – Not Documented in the paper											

Based on our evaluation of the pros and cons of each option, we suggest that studies aimed at assessing the conditions that would favour (or not) the use of a cross-company model should adopt the following procedure:

- Use new within-company data sets that are independent of existing cross-company data sets *and* that allow specific hypotheses to be tested. For example, use data sets that allow us to test the hypotheses that models derived from heterogeneous and homogenous within-company data sets would produce predictions similar to and better than cross-company predictions, respectively.
- Perform sensitivity analysis using residual analysis for non-regression-based methods and influence analysis for regression-based methods. Use a naïve estimate (e.g. median of data set) for a baseline.
- Use regression analysis as the default model construction method.
- Use a stepwise approach on the cross-company data set based on the variables collected in the within-company data set.
- Apply data transformations appropriate to the specific application (e.g. logarithmic transformation for regression compared with dividing by (max-min) for machine learning methods and analogy).
- Perform statistical tests based on the absolute residuals on the raw data scale.
- Report the residuals for each model or the effort and corresponding prediction for each model, for each project in the single company.

We are unable to provide definitive advice on the most appropriate means of cross-validation. However, we do not believe that leave-one-out cross-validation is a sufficiently stringent criterion for assessing the predictive accuracy of the within-company model. Kirsopp and

Shepperd recommend using a large number of training sets (at least 20) to minimise bias when comparing different estimation methods [9]. However, this procedure leads to many different estimates for a specific project using a specific estimation model. It complicates the subsequent analysis and may itself introduce bias since the procedure is only applicable to the within-company model. That is, when we are comparing within- and cross-company models, the prediction accuracy of the models is not based on the same cross-validation process. A similar, but slightly simpler procedure, might to be to use a jack-knife approach in combination with the leave-one-out cross-validation process [21], such that shown in Appendix A2. This process ensures that all predictions are based on a data set with two projects omitted, which is more robust than a leave-one-out cross-validation. The jack-knife procedure reduces the impact of atypical projects and delivers a more reliable estimate and, in addition, the estimate for each project is based on an equal number of predictions. However, it may be somewhat time-consuming if the procedure is done manually.

IV. Discussion

We found that only seven of the ten primary studies provided independent evidence concerning the viability of cross-company prediction models. Of the seven primary studies providing independent evidence, three studies found that the accuracy of the cross-company model was not significantly different from that of the within-company model; four studies found that the prediction accuracy of the within-company model was significantly better than that of the cross-company model. There were no studies where cross-company models were significantly better than within-company models.

Table 5a Study procedure factors – Data preparation, sensitivity analysis and statistical testing

Options for data preparation	**Pros**	**Cons**	**Used in Studies**
Data set transformed in a standard way independent of construction method	Easiest approach.	Risks using an inappropriate transformation.	S7
Data set transformed appropriately for each model construction method	Theoretically the best option.	More time consuming	Regression only (S2, S3, S4, S5, S6, S8, S9, S10)
Options for sensitivity analysis	**Pros**	**Cons**	**Used in studies**
Performed	Good practice because it reduces possibility of results being biased as a result of atypical data values.		S7, S8, S9, S10
Not performed	Simplest option when evaluating many different estimation methods.	Bad practice. Results may be biased by atypical data values.	Studies 1-6
Options for sensitivity analysis methods	**Pros**	**Cons**	**Used in studies**
Module residual analysis	Identifies projects that have a large residual. Re-analyzing the data with those projects omitted tests the resilience of the model. Can be undertaken for any prediction model, statistical or non-statistical.		None
Influence analysis	Identifies projects that have large residuals and have a large influence on the model.	Currently only feasible for regression.	S8, S9, S10
Comparison with naïve model	Provides assurance that the model is better than a simple baseline model.	Researchers may disagree about the baseline model.	S8, S9, S10
Comparison with random model	Provides assurance that the model is better than simple guesswork.	This is a minimal criterion for model validation.	S7
Options for prediction validation	**Pros**	**Cons**	**Used in Study**
Independent hold-out sample	Theoretically the best option particularly if there is a prior justification for the hold-out e.g. using projects started after a certain date as the hold-out.	Not feasible for small data sets	S1, S7
N-fold cross-validation where N<sample size (restricted to ensure one prediction per project)	A reasonable option if there is no obvious hold-out criteria. With a small data set hold-out samples could be at least 2 projects.		S2, S3, Not stated but inferred for S6.
N-fold cross-validation where N<sample size (allowing multiple predictions for each project)	Reduces bias in estimates of mean and variance of absolute residuals when comparing different estimation methods (see [9])	Complicates the analysis because an additional procedure is needed to determine the prediction to be used in any statistical test. If the average is used, this is biased unless each project had an equal number of predictions.	S10
N-fold cross-validation where N=sample size	The easiest cross-validation option practically, usually supported by options in statistical tools.	The worst cross-validation option theoretically since statistics based on a leave-one-out cross-validation are functionally related to statistics based on predictions without cross-validation.	S4, S5, S6(2), S6(3), S6(4), S6(5), S6(6), S8, S9
No hold-out. Simple estimates using the model based on all within company data points	The easiest option	The worst option because the within company model makes no valid predictions.	
Options for basis of statistical significance testing	**Pros**	**Cons**	
MRE		The metric is inherently biased	S2, S3, S4, S5, S6,
Absolute residual	The metric is unbiased.		S8, S9, S10
Options for statistical significance testing	**Pros**	**Cons**	**Used in studies**

Performed	Gives an objective assessment of whether one model is better than another.		S2, S3, S4, S5, S6, S8, S9, S10
Not performed		Does not allow a definitive assessment of whether or not one model is better than the other.	S1, S7

Table 5b Study procedure factors – Model construction options

Option for within-company selection	**Pros**	**Cons**	**Used in Study**
Part of the cross company data set	Will have collected data according to the database standards.		S1, S2, S3, S5, S6, S7, S8, S10
Independent data set	More representative of companies that want to utilize that cross-company data. Easier for experiments since it is easier to vary data set properties to investigate which factors affect the quality of estimates. (There are probably more within-company data sets than cross-company data sets.)	May not have collected appropriate data.	S4, S9
Options for cross-company model construction	**Pros**	**Cons**	**Used in Studies**
Stepwise approach independent of within company model		There is a risk of producing a model that cannot be used on the single company data (because input variables may not have been collected).	S1, S2, S3, S5, S6
Re-calibration of stepwise model obtained from all data (within- and cross-company data)	Ensures that the model can be used on the single company data. Realistic approach for a company that has a reasonable amount of their own data.	The cross-company model is not independent of the within- company model.	S4, S8, S9 (CMM2)
Stepwise approach based on measures collected on the within- company data set that are also collected by the cross-company data set	Ensures that the model can be used on the single company data. Realistic approach for a company that has a reasonable amount of their own data. The cross-company model is only dependent on the within-company model with respect to the choice of metrics not the functional form of the model.		S9 (CMM1)
Cross-company model includes within-company projects	Realistic approach for companies with any data	The cross-company model is not independent of the within-company model.	S7
Options for within-company model construction	**Pros**	**Cons**	**Used in studies**
Stepwise based on data available in benchmarking databases	Suitable if the single company is part of the cross-company data set.		S1, S2, S3, S5, S6, S7, S8, S10
Stepwise based on data collected in the company	Suitable if the single company is not part of the cross-company data set.		S4, S9
Options for model construction method	**Pros**	**Cons**	**Used in studies**
Regression (OLS, Stepwise, Robust)	The most commonly used method. All statistical tools support regression.		All studies
ANOVA (effort or productivity)		Not automated. In most cases equivalent to regression.	S3, S5, S6
CART (effort or productivity)		Requires a specialist tool.	S2, S3, S5
Analogy			S2, S3, S4, S5, S6, S7, S9
Genetic programming		May be difficult for non-experts	S7

Table 5c Study procedure factors – Reporting options

Options for accuracy statistics	**Pros**	**Cons**	**Used in studies**
Pred(25)	Simple measure. Can be adjusted correctly to allow for failure to make a prediction.		S1, S2, S4, S8, S9, S10
MMRE		Ratio-based measures are unstable and can lead to incorrect assessments (see [5]).	S1, S2, S4, S7, S8, S9, S10
MdMRE	Used in other disciplines (e.g. economics).	Ratio-based measures are unstable and can lead to incorrect assessments. (see [5]).	S2, S3, S4, S5, S6, S8, S9, S10
BalancedMRE		Ratio-based measures are unstable and can lead to incorrect assessments. (see [5]).	S7
Mean Absolute residual	Not as unstable or biased as ratio-based accuracy statistics.	Inappropriate for non-Normal distributions. Does not have an obvious baseline value.	S8, S9
Median absolute residual	Not as unstable or biased as ratio-based accuracy statistics.	Does not have an obvious baseline value.	S8, S9
Options for information reported	**Pros**	**Cons**	**Used in studies**
Selected accuracy statistics for within-company and cross-company predictions	Simplest option	This level of information is unsuitable for meta-analysis.	All studies
Mean difference between MRE for within- and cross-company predictions		This level of information is unsuitable for meta-analysis.	None
Mean difference between absolute residuals for within- and cross-company predictions		This level of information is unsuitable for meta-analysis.	None
Mean difference between MRE with standard error	Minimal data sufficient for restricted meta-analysis.	MRE is a biased statistic which would bias any meta-analysis.	None
Mean difference between absolute residuals with standard error	Minimal data required for restricted meta-analysis. MAR is unbiased.		None
Effort actual and predicted for each single company project	Sufficient data for meta-analysis. Makes testing a new model construction method easier (assuming the raw data is available to researchers) – the new method can be easily compared with previous results.	Single-company effort values may be commercially sensitive.	None
Residuals for each method for single company projects	Sufficient data for meta-analysis. Actual effort values remain confidential. Makes testing a new model construction method easier (assuming the raw data is available to researchers).		None

Previous studies suggested that data collection following rigorous quality assurance procedures might ensure that cross-company models were not significantly different from within-company models [2], [19], [26]. However, our results contradict this suggestion. Quality control on data does not appear to ensure that a cross-company model will perform as well as a within-company model.

The quality of the primary studies does not appear to affect the study results. In general, the quality scores for the more recent studies are higher than the quality scores for the earlier studies. This may simply indicate that recent studies have leant from the weak points of earlier studies.

We found that studies where within-company predictions were significantly better than cross-company predictions employed smaller within-company data sets, smaller number of projects in the cross-company models, smaller size projects, and databases where maximum effort was also smaller. In principle we would expect large within-company data sets to lead to more reliable results. This would imply we should put more trust in the results that suggest cross-company models are not significantly different from within-company models. There is, however, one explanation of the results that would favour the conclusion that within-company models based on small data sets are significantly better than cross-company models. Within-company datasets from small companies may be more homogeneous than datasets from larger companies. Furthermore, as within-company data sets grow, they may incorporate less similar projects (i.e. become more heterogeneous), particularly if the company grows and takes on new types of projects; then they would become more similar to the cross-company data set and differences between within- and cross-company models would cease to be significant. This is a hypothesis that would explain the phenomena we have observed (in particular the comparison of studies that used the ISBSG database and the Tukutuku database), and seems consistent with our knowledge

regarding within-company data sets. This is also consistent with the fact that in 3 of the 4 studies that found within-company models better than cross-company models, the single company was small (in the remaining case the size of the company is unknown).

In terms of the quality assessment criteria we employed (see Section II) our results indicated that the outcome of studies was not confounded with their overall quality. However, a more detailed assessment of two quality factors showed that the within-company data sets for studies that found with-company models better than cross-company models were smaller and used a less stringent validation process. It is likely that small within-company data sets are characteristic of small companies.

We found a large variation in the protocols adopted by different primary studies. We strongly recommend that any future researchers adopt a standard protocol such as the one we defined in Section III. This would improve the comparability of different primary studies and hopefully lead to studies that can be combined using meta-analysis. Although MdMRE and MMRE are the most frequently reported statistics they are known to be biased [5], and so they could not be the basis of a reliable meta-analysis. We would suggest basing any meta-analysis on the mean and standard error of the difference between the absolute residual of the within-company estimation model and the absolute residual of the cross-company estimation model for each project (see [16], pp 40-41), assuming that the distribution of the differences is approximately Normal.

The major validity issues facing this systematic review are whether we have failed to find all the relevant primary studies, and whether we have introduced bias because the systematic review authors contributed to three of the primary studies (S8, S9 and S10). To address the first issue we have undertaken a very stringent search strategy as described in Section II. However, we did not undertake an inter-rater reliability study during the first round of the search process, which

means it is possible that we missed some relevant studies. Our inclusion and exclusion criteria were fairly straightforward, and we reviewed the references of all references in identified studies and we contacted other researchers active in the area, so we think it is unlikely that we missed any relevant studies. Nonetheless, it would have been preferable to evaluate inter-rater reliability. With respect to the second issue, there are two concerns: i) we might have biased the quality assessment criteria to reflect our personal preferences with respect to experimental procedures, ii) we might have been less objective in extracting data from papers we ourselves wrote. With respect to the quality criteria, we note that the papers that scored best were written by systematic review authors; however, they were also the most recent studies and were able to avoid weaknesses found in earlier papers. Readers of this review must make their own assessment of the appropriateness of the quality criteria. To address possible data extraction bias, we ensured that no one would be the data extractor on a paper they authored.

Finally, it is important to note that any systematic review is limited to reporting the information provided in the primary studies. Therefore, it is important that future studies attempt to characterise both the within- and cross-company data sets more fully. In particular, it is critical that researchers present more information about the single company, such as the size of the company, the nature of the software applications provided by the company, the quality control procedures, CMMI level etc.

V. Conclusion and Future Work

Based on our review it is clear that some organisations would benefit from using models derived from cross-company benchmarking databases but others would not. The results of our systematic review are unable to provide conclusive explanations of why this occurs, however, we have observed some trends.

In all cases where within-company data sets significantly outperformed cross-company models, the data sets were small and the cross-validation method was not very stringent. It is possible that the cross-validation method biased the results in favour of the within-company models.

In two, out of four cases, where the within-company model presented significantly better predictions than the cross-company models, the single company projects had been collected separately from the cross-company projects. In contrast, in all cases where the within- and cross-company models were not significantly different, the within-company data was a subset of the cross-company data set (i.e. was not collected separately from the cross-company projects). Thus, the studies where the cross-company model was not significantly different from the within-company model, the results might be an artefact of the data set, i.e. the single company projects in cross-company data sets were collected with the cross-company data set in mind and may not have been collected as a homogeneous group of projects.

In three out of four cases, where the within-company model presented significantly better predictions than the cross-company models, the single company projects were volunteered by small companies (Megatec and the two single companies from Tukutuku). In all cases where the within-company model presented significantly better predictions than the cross-company models, the single company projects (in terms of effort) were relatively small.

Therefore the advice we can give currently, based on the results of this study and our own experience, is to consider how similar project data in the cross-company data set are to the projects undertaken in your own company and consider the characteristics of your own company (see Table 6). We would expect most of these factors to be additive, i.e. the more characteristics that agree that a cross-company model is likely to be appropriate, the more likely it is to be appropriate, although some characteristics are related (e.g. C3 and C4). A false positive occurs

when a study exhibits a characteristic that should favour a particular type of model but the study does not favour that type of model. Conversely, a false negative occurs when a study does not exhibit a characteristic that favours a particular type of model but the study does favour that type of model.

Table 6 A summary of advice on factors to consider when considering a cross-company model

Characteristic	Id	Cross-company Model	Evidence in favour	Evidence against
Large company	C1	Yes	Large within-company data sets may indicate large companies. Studies S2, S3 & S10 have large within-company data sets	No contrary examples
Subsets of cross-company data set that contain applications in same business domain.	C2	Yes	Our experience plus all studies using large benchmarking databases made some attempt to match the projects in the single company to the cross-company projects.	S4, S8 (false positives)
Some very large projects.	C3	Yes	S2, S3	S10 (false negative)
Within-company projects broadly similar in size and effort to cross-company projects	C4	Yes (but similar to C3)	S2, S10	S3 (false negative), S8 (false positive partially)
Small company	C5	No	S4, S8, S9	No contrary examples
Specialised products	C6	No	S8, S9	S4 (false negative)
Relatively small projects	C7	No	S4, S5, S8, S9	No contrary examples
Relatively homogeneous within-company data sets	C8	No (but similar to C7 & C6)	S8, S9	S4 (false negative)

Clearly, further research is required to provide definitive advice to organizations deciding whether or not to use cross-company models. We strongly recommend that researchers come to some consensus about the most appropriate experimental procedure for this type of study and use the same procedure for future studies; we recommend one in part III. We also suggest that future studies should aim to test specific hypotheses about the conditions that favour or not the use of cross-company estimation models and should report more information about the characteristics of the single company and its projects. We strongly support studies such as S4 and S9 that obtained data from a single company that was independent of the cross-company data base.

We also believe it is important to try out both the jack-knife approach for effort prediction for small within-company estimates and our proposals for meta-analysis. This we can do for the data sets that we have access to, and we will endeavour to coordinate our work with researchers who have access to the other data sets.

ACKNOWLEDGEMENT

Barbara Kitchenham's research is supported by both the EPSRC EBSE project (EP/C51839X/1) and National ICT Australia**.** National ICT Australia is funded through the Australian Government's Backing Australia's Ability initiative, in part through the Australian Research Council. Guilherme H. Travassos research is supported by CNPq – Brazilian Research Council - eSEE Project (472135/2004-0).

REFERENCES

[1] B.W. Boehm, *Software Engineering Economics*, Prentice-Hall,1981.

[2] L.C. Briand, K. El-Emam, K. Maxwell, D. Surmann, and I. Wieczorek, An assessment and comparison of common cost estimation models, *Proceedings of the 21st International Conference on Software Engineering*, pp. 313-322,1999.

[3] L.C. Briand, T. Langley, and I. Wieczorek, A replicated assessment of common software cost estimation techniques, *Proceedings of the 22nd International Conference on Software Engineering*, pp. 377-386, 2000.

[4] T. DeMarco, *Controlling Software Projects: Management measurement and estimation*, Yourdon Press, New York, 1982.

[5] T. Foss, E. Stensrud, B. Kitchenham, and I. Myrtveit, A simulation study of the model evaluation criterion MMRE. *IEEE Transactions on Software Engineering*, Volume: 29, Issue: 11, Nov. Pages:985 – 995, 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. *Information and Software Technology*, 42, 1009-1016, 2000.

[7] R. Jeffery, M. Ruhe, and I. Wieczorek, Using public domain metrics to estimate software development effort, *Proceedings Metrics'01*, London, pp. 16-27, 2001.

[8] C.F. Kemerer, An empirical validation of software cost estimation models. *Communications of the ACM*, 30(5), 1987.

[9] C. Kirsopp, and M. Shepperd, Making Inferences with Small Numbers of Training Sets. *IEE Proceedings Software*, 149(5), pp123-130, 2002.

[10] B. Kitchenham, Procedures for Performing Systematic Reviews. *Joint Technical Report Software Engineering Group*, Keele University, United Kingdom and Empirical Software Engineering, National ICT Australia Ltd, Australia, 2004.

[11] B.A. Kitchenham, E. Mendes, A Comparison of Cross-company and Single-company Effort Estimation Models for Web Applications, *Proceedings EASE 2004*, pp. 47-55, 2004.

[12] B.A. Kitchenham, and N.R. Taylor, Software cost models. *ICL Technical Journal*, pp. 73-102, 1984.

[13] B.A. Kitchenham, E. Mendes, and G.H. Travassos, A systematic Review of Cross- vs. Within Company Cost Estimation Studies. *Proceedings EASE06*, BCS, 2006. (Available at http://ewic.bcs.org/conferences/2006/ease06/index.htm)

[14] P.A.M. Kok, J. Kirakowski, and B.A. Kitchenham, The MERMAID approach to software cost estimation, *ESPRIT'90*, Kluwer Academic Press, pp 296-314, 1990.

[15] M. Lefley, 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.

[16] M. W. Lipsey, and D.B. Wilson, *Practical Meta-analysis*, Sage Publications Inc., 2001

[17] C. Mair, and M.J. Shepperd, The Consistency of Empirical Comparisons of Regression and Analogy-based Software Project Cost Prediction, Proceedings ISESE'05, Noosa Heads, 17-18 Nov., pp. 509-518, 2005.

[18] K. Maxwell, L.V. Wassenhove, and S. Dutta, Performance Evaluation of General and Company Specific Models in Software Development Effort Estimation, *Management Science*, 45(6), June, 787-803, 1999.

[19] E. Mendes, and B.A. Kitchenham, Further Comparison of Cross-Company and Within Company Effort Estimation Models for Web Applications, *Proceedings Metrics'04*, Chicago, September 11-17th, IEEE Computer Society, pp. 348-357, 2004.

[20] 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, *Proceedings of Metrics'05*, Como, 2005.

[21] F. Mosteller, and J.W. Tukey, Data Analysis and Regression, Addison-Welsey, 1977.

[22] M. Pai, M. McCulloch, J.D. Gorman, N. Pai, W. Enanoria, G. Kennedy, P. Tharyan, and J.M. Jr. Colford, Systematic reviews and meta-analyses: An illustrated step-by-step guide. *The National Medical Journal of India*, 17(2), 86-95, 2004.

[23] M. Petticrew, and H. Roberts, *Systematic Reviews in the Social Sciences: A practical guide*. Blackwell Publishing, 2006.

[24] L.A. Putnam, A general empirical solution to the macro software sizing and estimating problem, *IEEE Transactions on Software Engineering*, 4(4), 1978.

[25] J. Stathis. Early Lifecycle Specification based on Sizing with Function Points. University of New South Wales, BSc Thesis, 1993.

[26] I. Wieczorek, and M. Ruhe, How valuable is company-specific data compared to multi-company data for software cost estimation?, *Proceedings Metrics'02*, Ottawa, June, pp. 237-246, 2002.

APPENDIX

A1. **Search String**

(software OR application OR product OR Web OR WWW OR Internet OR World-Wide Web OR project OR development) AND (method OR process OR system OR technique OR methodology OR procedure) AND (cross company OR cross organisation OR cross organization OR cross organizational OR cross organisational OR cross-company OR cross-organisation OR cross-organization OR cross-organizational OR cross-organisational OR multi company OR multi organisation OR multi organization OR multi organizational OR multi organisational OR multi-company OR multi-organisation OR multi-organization OR multi-organizational OR multi-organisational OR multiple company OR multiple organisation OR multiple organization OR multiple organizational OR multiple organisational OR multiple-company OR multiple-organisation OR multiple-organization OR multiple-organizational OR multiple-organisational OR within company OR within organisation OR within organization OR within organizational OR within organisational OR within-company OR within-organisation OR within-organization OR within-organizational OR within-organisational OR single company OR single organisation OR single organization OR single organizational OR single organisational OR single-company OR single-organisation OR single-organization OR single-organizational OR single-organisational OR company-specific) AND (model OR modeling OR modelling) AND (effort OR cost OR resource) AND (estimation OR prediction OR assessment)

A2 **Jack-knife Algorithm for Estimating Prediction Accuracy**

For a within-company data set of size N the following jack-knife procedure can be followed:

```
For i=1, i<= N, i++ Do
    Omit project i from the data set
    For j = 1, j <= N-1, j++ Do
        Omit project j from data set
        Calculate the best predictive model from the remaining N-2 projects
        Predict the effort of project i
        Return project j to the data set
    End_For
    Calculate the average of the N-1 predictions of project i
    Return project i to the data set
End_For
```