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

Barbara Kitchenham[1], Emilia Mendes[2], Guilherme H. Travassos[3]
[1] Department of Computer Science, Keele University, Staffordshire ST5 5GB, UK
and
National ICT Australia, Locked Bag 9013 Alexandria, NSW 1435, Australia
barbara@cs.keele.ac.uk; barbara.kitchenham@nicta.com.au

[2] Computer Science Department, Private Bag 92019, The University of Auckland, Auckland, New Zealand
emilia@cs.auckland.ac.nz

[3] UFRJ/COPPE
Systems Engineering and Computer Science Program,
Caixa Postal 68511, 21941-972 Rio de Janeiro – RJ, Brazil
ght@cos.ufrj.br

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-company and within-company estimation models, however, only seven of the papers presented independent results. Of those seven, three found that cross-company models were as good as within-company models, four found 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 single company data sets (i.e. <20 projects) that used leave-one-out cross-validation all found that the within-company model was significantly more accurate than 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). In addition, experimenters need to standardise their experimental procedures to enable formal meta-analysis.

Keywords: Cost estimation models, cross-company data, within-company data, estimation accuracy, systematic review

INTRODUCTION

Early studies of cost estimation models (e.g. [11] [8]) suggested that general-purpose models such as COCOMO [1] and SLIM [20] 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. [13] suggested that cost estimation models should be developed only from within-company data. However, three main problems can occur when relying on within-company [3], [10]:

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 several studies compared the prediction accuracy of cross-company and within-company models. In 1999, Maxwell et al. [15] 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. This result was confirmed the following year by Briand et al.[3], using a different data set. Two years later, Wieczorek and Ruhe [21] confirmed this same trend using the same database employed by [2]. Three years later, Mendes et al. [18] also confirmed the same trend using yet a different database.

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-workers 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 [14] claimed that the within-company model was more accurate than the cross-company model, using the same data set employed by Wieczorek and Ruhe [21] and Briand et al. [2]. Finally, a year later Kitchenham and Mendes undertook two studies of Web-based projects [10] [12]. 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 within-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 variations in study protocol, i.e. experimental procedure. The 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 dataset.

The paper is organised as follows: we first describe the method we used for our systematic review. Next we present the results and then discuss the results and threats to the validity of the results. The final section in the paper presents our conclusions and plans for future work.

METHOD

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 [9] [12]. There are numerous motivations 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 [12][19]:

- 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 (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.

A well-formulated question generally has four parts [19], 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 factor and across populations.

Research Questions, Population, Intervention

Within the context of this paper we have carried out a systematic literature review using the basic approach identified in [9], 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 worse than within-company estimation models for predicting effort for software/Web projects?
- Question 2: Do the characteristics of the study data sets and the data analysis methods used in the study affect the outcome of within-company and cross-company effort estimation accuracy studies?

Since some studies also compared prediction accuracy between prediction techniques, we also had a secondary research question, which will not be discussed in this paper due to lack of space:

- Question 3: Which estimation method(s)/process(es) were best for constructing cross-company effort estimation models?

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 the single company data only. The studies' outcomes that were of interest to our systematic review were the accuracy of the cross- and within-company models. Finally, the experimental design that was of interest to our systematic review was that of observational studies using existing cross-company and within-company data bases, where their estimates for project effort are compared using within-company data hold-out sample(s).

Search Strategy used for Primary Studies

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

- Derive major terms from the questions by identifying the population, intervention and outcome;
- Identify alternative spellings and synonyms for major terms. We also included terms identified via consultations with experts in the field and/or subject librarians;
- Check the keywords in any relevant papers we already have;
- Use the Boolean OR to incorporate alternative spellings and synonyms;
- Use the Boolean AND to link the major terms from population, intervention and outcome.

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. The complete set of search strings was:

> (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)

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 string. The electronic searches were based on:

- Databases
  - INSPEC
  - EI Compendex
  - Science Direct
  - Web of Science
  - IEEExplore
  - ACM Digital library

- Individual journals (J) and conference proceedings[1] (C)
  - Empirical Software Engineering (J)
  - Information and Software Technology (J)
  - Software Process Improvement and Practice (J)
  - Management Science (J)
  - International Software Metrics Symposium (C)
  - International Conference on Software Engineering (C)
  - Empirical Assessment in Software Engineering (manual search) (C)

In relation to the electronic databases we ensured that our search was applied to journals, magazines and conference proceedings published since 1999, i.e. since this was the year when the first cross-company vs. within-company study was published. The search process was assessed by comparing the primary studies found by each search engine with the primary studies we already knew about (see Table 1). At the time the searches were conducted we knew about 10 studies, 9 that had been published and one that was not yet published when our searches were performed and could therefore not have been found by any of the search engines.

All nine known and published papers were found after searching 13 different sources. No new relevant papers were found. 772 papers were retrieved, of which 24 represented the set of nine known relevant papers (the same paper was retrieved by different search engines). Overall IEEExplore retrieved the largest number of known papers – 5, followed by INSPEC, Science Direct and Web of Science, each with 4 papers. Science Direct, Metrics Proceedings search engine, and Management Science search engine (JSTOR) each missed a known paper of a publication that should be indexed by that search engine. ACM Digital library missed both known papers published at conference proceedings indexed by this search engine, and did not even retrieve any false positive papers.

The *Secondary* search phase

The *Secondary* search phase had two sub-phases: i) to review the references in 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.

A review of all the references in the known relevant papers found no new references (see Table 2). Although it was one of the first papers on the topic, Maxwell et al [15] is less cited than Briand et al [2], perhaps because it was published in a management science journal rather than a software engineering journal, or because it did not produce the unexpected results that [2] produced. It is however, unusual that Maxwell was an author on both papers and did not cross-reference her own work. Briand et al. [2] and Lefley and Shepperd [14] are the most and least cited papers, respectively. A reason for Lefley and Shepperd [14] being cited only once may be because this paper was published in the conference proceedings of a conference that was not primarily about software engineering.

In general, all known papers identified a relatively high proportion of the preceding papers on the topic. The exceptions are: Lefley and Shepperd [14], which referenced only two out of six possible papers; and Briand et al.[3], which did not reference Maxwell et al [15] despite using the same data set and single company.

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 project data*. We excluded studies where projects were only collected from a small number of different sources (e.g. 2 or 3 companies). We also excluded studies where models derived from a single company dataset were compared with predictions from a general cost estimation model.

As part of our preliminary selection process, the three authors applied the search strategy to identify potential primary studies. Each author used a different set of databases/journals/conference proceedings. No new potential primary studies were identified.

Study Quality Assessment Checklists

The criteria used to determine the overall quality of the primary studies included six top-level questions and an additional quality issue. The overall quality score for a paper ranged from 0 to 7, representing very poor and excellent quality, respectively. Top-level questions without sub-questions were answered Yes/No/Partially,

[1] These conferences were chosen as they had previously published primary studies we knew about.

corresponding to scores 1, 0, and 0.5 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 five sub-questions, thus each “Yes”, “No”, and “Partially” for a sub-question contributed scores of 0.2, 0, and 0.1 respectively.

The six main questions were:

1. Is the analysis process description complete?
    1.1. Was the data investigated to identify outliers and to assess distributional properties before analysis?
    1.2. Was the result of the investigation used appropriately?
    1.3. Were the resulting estimation models subject to sensitivity or residual analysis?
    1.4. Was the result of the sensitivity or residual analysis used appropriately?
    1.5. Were accuracy statistics based on the raw data scale?
2. Is it clear what projects were used to construct each model?
3. Is it clear how accuracy was measured?
4. Is it clear what cross-validation method was used?
5. Were all model construction methods fully defined (tools and methods used)?
6. How good was the study comparison method?
    6.1. Was the single company selected at random (not selected for convenience) from several different companies?
    6.2. Was the comparison based on a completely independent hold out sample or on n-fold cross-validation for the within-company model?

The additional quality issue considered was the size of the within-company data set, measured according to the criteria presented below. Whenever a study used more than one within-company data set, the average score was used:

- 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)

The size of the within-company data set was considered as part of the study quality criteria because it was 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).

Each reviewer assessed each paper assigned to them against each criterion. Scores attributed to our primary studies are presented in Table 4, and indicate that, according to our scoring scheme, the papers that received the highest and lowest quality scores were S10 and S3, respectively.

Data Extraction Strategy

Required Data

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

- Extracted data: data extractor, data checker, study identifier
- Database: name of database, application domain, number of projects in database (including single-company projects), number of companies, number of countries represented, if quality controls were applied to data collection, data summary
- Projects: number of cross-company projects, number of projects in single company, size metric(s),
- Study: how accuracy was measured, cross-company model details, within-company model details, comparison between cross and within-company models

TABLE 1: Coverage of Search process

		INSPEC[2]	EI Compendex[3]	Science Direct[3]	Web of Science[3]	IEEE Xplore 2.1[3]	Metrics Procs[4]	ICSE Procs[5]	EASE Procs[5]	Empirical Sw. Eng.[6]	Inform. Sw. Tech.[7]	Sw. Proc. Impr. & Practice[8]	Mgmt. Science[3]	ACM Digital library[9]	Overall
Number of papers retrieved		224	60	453	19	9	3	2	1	0	1	0	0		772
Authors	Year	Did the search identify this paper?													
Maxwell, K., L.V. Wassenhove, and S. Dutta.	1999				YES								Missed		YES
Briand, L.C., K. El-Emam, K. Maxwell, D. Surmann, I. Wieczorek	1999	YES		YES		YES		YES						Missed	YES
Briand, L.C., T. Langley, I. Wieczorek	2000	YES				YES		YES						Missed	YES
Jeffery, R., .M. Ruhe and I. Wieczorek	2000	YES		YES	YES										YES
Jeffery, R., M. Ruhe and I. Wieczorek	2001		YES	YES		YES	YES								YES
Wieczorek, I. and M. Ruhe	2002					YES	Missed								YES
Lefley, Martin and M.J. Shepperd	2003			Missed	YES										YES
Kitchenham, B.A., and E. Mendes	2004	YES							YES						YES
Mendes, E. and B.A. Kitchenham	2004			YES	YES	YES	YES								YES
Total relevant papers	9	4	1	4	4	5	2	2	1	0	1	0	0		24 (9 papers)
Total irrelevant papers	n/a	220	59	449	15	4	1	0	0	0	0	0	0		n/a

[2] Years 1999-2004; Full search set was used

[3] Years 1999-2004; Search string: (software) 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 company-specific) AND (effort OR cost) AND (estimation)

[4] Years 1999, 2001-2004 (Metrics); 1999-2004 (ICSE); Search string: (cross-company or multi-company) and effort; (cross-organization or multi-organization) and effort; (cross-organisation or multi-organisation) and effort; (multi company or multi organization and effort; (cross company or cross organization and effort; (multi organizational and effort); (multi-organizational and effort); (cross-organizational and effort); (cross organizational and effort); (multiple company and effort).

[5] Years 1999-2004 (hand search)

[6] Years 1999-2004; Simple Search strings failed: (“cross company” or “multi company”) and (cost or effort) caused Server error; (“cross organization” or “multi organization” and (cost or effort) cased “invalid criteria”

[7] Years 1999-2004; Search strings: (cross-company or multi-company) and (cost or effort) / ("cross company" or "multi company") and (cost or effort)

[8] Years 1999-2004; Search strings: ("cross-organization" or "multi-organization") and (cost or effort) / ("cross organization" or "multi organization") and (cost or effort) / ("cross organization" or "multiorganization") and (cost or effort) / ("cross organisation" or "multi organisation") and (cost or effort) / ("cross-company" or "multi-company") and (cost or effort) / ("cross company" or "multi company") and (cost or effort)

[9] Years 1999-2004; Search string: +((software +) +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 +company-specific) +AND +(effort +OR +cost) +AND +(estimation))

TABLE 2: Citations and New references found

Authors	Study ID	Year	Known references found	New references
Maxwell, K., L.V. Wassenhove, and S. Dutta	S1	1999	0	0
Briand, L.C., K. El-Emam, K. Maxwell, D. Surmann, I. Wieczorek[10]	S2	1999	0	0
Briand, L.C., T. Langley, I. Wieczorek[10]	S3	2000	[2]	0
Jeffery, R., .M. Ruhe and I. Wieczorek	S4	2000	[1],[2],[3]	0
Jeffery, R., M. Ruhe and I. Wieczorek	S5	2001	[1],[2],[3],[4]	0
Wieczorek, I. and M. Ruhe.	S6	2002	[2],[3],[4],[5]	0
Lefley, Martin and Shepperd, Martin, J.	S7	2003	[1],[5]	0
Kitchenham, B.A., and E. Mendes.	S8	2004	[2],[3],[4],[5],[6]	0
Mendes, E. and B.A. Kitchenham.	S9	2004	[2],[3],[4],[5],[6],[8]	0
Mendes, E., C. Lokan, R. Harrison, C. Triggs	S10	2005	[1],[2],[3],[4],[5],[6],[7],[8],[9]	0

Data Extraction Process

For each paper a researcher 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.

RESULTS

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

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

The evidence provided in Table 3 suggests that four (S2, S3, S6, S10) studies show that cross-company estimation models are not significantly worse than within-company estimation models for predicting effort for software/Web projects. However, S6 cannot be considered an independent study since it used the same data set employed in S2. Although it compared six single 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 add any significant information to the results previously obtained by S2.

TABLE 3: Summary of evidence

Study	Database	Basis for Predictions[11]	Statistical tests comparing Within (WC) to Cross-company (CC)
Cross-company model not significantly worse than 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
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
S6	Laturi	6 different leave-one-out cross-validations (one for each WC data set), or randomly selected test sets	Wilcoxon matched pairs test. Measure used is unknown
S10	ISBSG	20-fold cross-validation (62 projects in validation set)	Mann-Whitney test 2 independent samples on absolute residuals
Cross-company model significantly worse than within-company model			
S4	Megatec and ISBSG	19-fold cross-validation (1 project validation set)	Wilcoxon matched pairs test on MREs
S5	ISBSG	12-fold cross-validation (1 project validation set)	Wilcoxon matched pairs test on MREs
S8	Tukutuku	13-fold cross-validation (1 project validation set)	Wilcoxon matched pairs test and paired t-test on absolute residuals
S9	Tukutuku	14-fold cross-validation (1 project validation set)	Wilcoxon matched pairs test on absolute residuals
Inconclusive			
S1	ESA	Independent hold-out (9 projects)	Correlation analysis between actual and estimate, no formal statistical significance test
S7	Laturi	Independent hold-out (15 projects)	No formal statistical significance test

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

[11] Cross-validation for within-company model

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. 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 significant additional evidence.

Table 3 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.

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

Mendes and Kitchenham (S9) noted that studies where the cross-company database 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. (S2) 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 as good as 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 are not a necessary or sufficient cause for cross-company models to perform well.

TABLE 4: Study related factors

Study	Quality control on data collection (Database)	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
Cross-company models not significantly worse than within-company models							
S2	Yes (Laturi)	5.2	206(119-63)	63	Min: 480 Max: 63694	Unadjusted Experience Function Points	Yes
S3	Claims Yes, but not fully (ESA)	4.87	166(131)	29	Min: 3 Max: 627984	Adjusted KLOC	Yes
S6	Yes (Laturi)	5.14	206 (206 – WC size)	63, 13,12, 11,10,10	Min: 250 Max: 63694	Unadjusted Function Points	Yes
S10	No (ISBSG)	6.5	872(680)	187	Min: 14 Max: 73920	IFPUG Function Points	Yes
Cross-company model significantly worse than within-company models							
S4	No(ISBSG), Yes (Megatec)	5.53	451(145)	19	Isbsg: Min: 10 Max: 59809 Megatec: Min: 194 Max: 13905	Unadjusted Function Points	Yes
S5	No (ISBSG)	5.43	324(310)	14	Min: 97; Max:59809	Function Points	Yes
S8	No (Tukutuku)	5.83	53(40)	13	Min:6 Max:5000	23 different size measures	Not completely
S9	No (Tukutuku)	5.83	67(53)	14	Min:6 Max:5000	9 different size measures	Yes (CCM1) No (CCM2)
Inconclusive							
S1	No (ESA)	5.77	108 (60)	29	Min: 1123.2 Max: 627984	KLOC	Yes
S7	Yes (Laturi)	5.9	407(149)	63	Not provided	Function points (Laturi variant)	No, CC used 48 WC projects
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							

Our quality evaluation of the studies shows no consistent evidence that the quality of the studies influences the results. The score for studies S2 and S3 are lower than that for studies S4, S5, S8, and S9, but S10 has the highest quality score.

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 this pattern is more noticeable 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. Another difference between studies S2, S3, S10, and studies S4, S5, S8, S9 is the range of effort values for the entire database, which are 73920 and 13905 person hours, respectively.

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.

DISCUSSION

We found that only seven of the ten primary studies provided independent evidence concerning the comparative accuracy of cross-company and within-company prediction models. Overall results were inconclusive. Three studies found that a cross-company model gave prediction accuracy not significantly worse than that of a within-company model; four studies found that a cross-company model gave prediction accuracy significantly worse than that of a within-company model.

Previous studies suggested that data collection following rigorous quality assurance procedures might enable cross-company models to be as accurate as within-company models [2] [10] [21]. However, our results contradict this suggestion. Quality control on data does not appear to be sufficient 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 learnt from the weak points of the 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, and databases where maximum effort was also smaller. We speculate that as within-company data sets grow, they incorporate less similar projects so that differences between within and cross-company data sets cease to be significant.

We conducted a simple systematic review not using formal 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. Therefore, the major validity issues 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 2. 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 sets more fully.

CONCLUSIONS AND FUTURE WORK

This paper presented the results of a systematic review of 10 primary studies that compared within and cross-company effort predictions based on analysis of project data. Of the 10 primary studies, three were not independent studies, leaving seven papers. Of these, three showed that cross-company predictions were as good as within-company predictions, and four showed within-company predictions to be significantly better than cross-company predictions. Studies differed in their experimental procedures and did not publish their residuals, or actual and estimated effort, thus we were unable to carry out a meta-analysis.

Our results showed that strict quality control on data collection did not seem sufficient to ensure that a cross-company model performs as well as a within-company model. In addition, the quality of primary studies did not seem to affect study results.

The main trend in the study related factors was 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, and databases where maximum effort was also smaller.

Given the problems we encountered trying to carry out a meta-analysis, we wish to make a recommendation that residuals, or actual and estimated effort be published in all future papers on the topic so that proper meta-analysis can be performed. We also recommend that future studies in the topic be independent, and that a standard experimental procedure be used to enable formal meta-analyses.

Finally, further details, including the analysis for Question 3 will be the subject of a journal paper.

ACKNOWLEDGMENTS

Barbara Kitchenham is a member of the EBS project team (EP/C51839X/1), which is funded by the UK Engineering and Physical Science Research Council (EPSRC). Guilherme H. Travassos is a CNPq – Brazilian Research Council funded researcher and leads the CNPq eSEE Project (4732135/2005-0).

REFERENCES

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

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

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

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

[5] Foss, T., Stensrud, E., Kitchenham, B., and Myrtveit, I. A Simulation Study of the Model Evaluation Criteria MMRE, *IEEE Transactions on Software Engineering*, 29(11), 2003, pp 985-995.

[6] Jeffery, R., Ruhe, M. and Wieczorek, I. (2000) A Comparative Study of Two Software Development Cost Modeling Techniques using Multi-organizational and Company-specific Data. *Information and Software Technology*, **42**, 1009-1016.

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

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

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

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

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

[12] Kitchenham, B.A., Dyba, T. Jorgensen, M. (2004) Evidence-based software engineering, *Proceedings 26th International Conference on Software Engineering*, (23-28 May 2004), pp. 273 – 281.

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

[14] Lefley, M., and Shepperd, M.J. (2003) Using Genetic Programming to Improve Software Effort Estimation Based on General Data Sets, *Proceedings of GECCO 2003*, LNCS 2724, Springer-Verlag, pp. 2477-2487.

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

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

[17] Mendes, E., Mosley, N. and Counsell, S. (2003) Investigating Early Web Size Measures for Web Cost Estimation, *Proceedings of EASE'2003 Conference*, Keele, April, 1-22.

[18] Mendes, E., Lokan, C., Harrison, R. and Triggs, C. (2005) A Replicated Comparison of Cross-company and Within-company Effort Estimation models using the ISBSG Database, *Proceedings of Metrics'05*, Como.

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

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

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