Cross- vs. Within-Company Cost Estimation Studies Revisited: An Extended Systematic Review

Emilia Mendes
Blekinge Institute of Technology
Karlskrona 37133, Sweden
+46 0455 - 38 50 00
emilia.mendes@bth.se

Marcos Kalinowski, Daves Martins
University of Juiz de Fora
Juiz de Fora 36036, Brazil
+55 32 2102-3311
{kalinowski, davesmartins}@ice.ufjf.br

Filomena Ferrucci
University of Salerno
Via Ponte don Melillo, 84084, Fisciano, Italy
+39 08996-3374
fferrucci@unisa.it

Federica Sarro
University College London
London, WC1E 6BT, UK
+44 207679037289
f.sarro@ucl.ac.uk

ABSTRACT

[Objective] The objective of this paper is to extend a previously conducted systematic literature review (SLR) that investigated under what circumstances individual organizations would be able to rely on cross-company based estimation models. [Method] We applied the same methodology used in the SLR we are extending herein (covering the period 2006-2013) based on primary studies that compared predictions from cross-company models with predictions from within-company models constructed from analysis of project data. [Results] We identified 11 additional papers; however two of these did not present independent results and one had inconclusive findings. Two of the remaining eight papers presented both, trials where cross-company predictions were not significantly different from within-company predictions and others where they were significantly different. Four found that cross-company models gave prediction accuracy significantly different from within-company models (one of them in favor of cross-company models), while two found no significant difference. The main pattern when examining the study related factors was that studies where cross-company predictions were significantly different from within-company predictions employed larger within-company data sets. [Conclusions] Overall, half of the analyzed evidence indicated that cross-company estimation models are not significantly worse than within-company estimation models. Moreover, there is some evidence that sample size does not imply in higher estimation accuracy, and that samples for building estimation models should be carefully selected/filtered based on quality control and project similarity aspects. The results need to be combined with the findings from the SLR we are extending to allow further investigating this topic.

Categories and Subject Descriptors

D.2.9 [**Software Engineering**]: Cost Estimation

General Terms

Measurement, Experimentation.

Keywords

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

1. INTRODUCTION

Some of the early cost estimation studies (e.g. [15] [9]) suggested that general-purpose models such as COCOMO [1] and SLIM [25] needed to be calibrated to specific companies before they could be used effectively. This view was also supported by Kok et al. [17], which, as per the proposals made by DeMarco [5], suggested that cost estimation models should be developed only from within-company data.

However, problems may arise when relying on within-company [4] [12], such as: (1) 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; and (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, productivity benchmarking and defect prediction, and several studies compared the prediction accuracy of cross-company and within-company models. By the end of 2006, ten studies compared the prediction accuracy between cross- and within-company effort estimation models [13]; however only seven of these presented independent results [13]. Of these seven, three found that cross-company models were not significantly different to within-company models and four found that cross-company models were significantly worse than within-company models [13].

The abovementioned studies and their comparison have been detailed in a Systematic literature review (SLR) [13] [14] that identified and analysed studies published between 1990 and November 2006. This SLR aimed to answer three research questions, as follows:

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

The first two questions were detailed in [13], and the third question was addressed in [14].

No meta-analysis of the results was reported in the SLR because the experimental procedures used by the primary studies differed making it impossible to undertake a formal meta-analysis.

The main trend distinguishing the study results was that all the studies that used small within-company data sets (i.e. <20 projects) in combination with leave-one-out cross-validation found that the within-company model was significantly different (better) to the cross-company model.

The issues that motivated the execution of the abovementioned SLR remain important in practice. In addition, seven years have passed since that SLR was published. These two facts motivated us to extend this SLR covering the period 2006 to 2013, further investigating the factors that influence the outcome of studies comparing within and cross-company models, aiming to assist software companies with small data sets in deciding whether or not to use an estimation model obtained from a benchmarking dataset.

Note that due to lack of space, the main focus of this paper will be to detail the findings of our extended SLR, however making a point to compare it to [13] as often as possible. The full integration of our findings with the previous SLR will be the focus of another publication. Finally, despite using the same method as in [13], we will detail it herein for clarity's sake.

The remainder of this paper is organized as follows. Section 2 describes the method employed for our extended systematic review. Section 3 presents the ESLR's partial results. In Section 4 we discuss those results and threats to validity. Throughout the discussion, we also compare the OSLR and ESLR results. The final section in the paper presents our conclusions and plans for future work.

2. METHOD

A consequence of the growing number of empirical studies in software engineering is the need to adopt systematic approaches for aggregating research outcomes in order to provide a balanced and objective summary of evidence on a particular research topic [2]. In this context, SLRs have become a widely used and reliable research method [19].

Guidelines for performing SLRs in the software engineering domain were proposed by Kitchenham and Charters [11], and since then numerous software engineering SLRs have been published. Brereton et al. [2] report lessons learned from applying SLRs in the software engineering domain. The three main Phases of a SLR are [11]: Planning the Review, Conducting the Review, and Reporting the Review. Kitchenham and Charters [11] also suggest that the PICO [24] (population, intervention, comparison, outcome) strategy be used for detailing the research questions in order to support developing the review protocol. These questions should also be sufficiently broad to allow examination of variation in the study factor and across populations.

One of the main advantages of using SLRs in the context of this paper is enabling incremental updates on top of previous SLRs. Another example of such updates is available by Kalinowski et al. [9], where four independent SLR trials were conducted to incrementally produce evidence-based guidelines on defect causal analysis.

Research Questions, Population, Intervention. Within the context of this paper we carried out a SLR that extended a previous SLR [13], using the basic approach identified in [11], 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?
- Question 3: Which estimation method(s)/process(es) were best for constructing cross-company effort estimation models?

Note that, similarly to [13], the results for Question 3 will not be discussed in this paper, and will be the focus of a journal publication.

As in [13], 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 as often as possible the same ones used in [13]; the complete set of search strings is as follows:

(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)

We also employed the same two search phases as in [13]: Initial and Secondary. In addition, whenever a database did not allow the use of complex Boolean search strings we designed different search strings for each of these databases. The search strings were piloted and results documented.

Initial Search Phase. The Initial search phase identified candidate primary sources based on our own knowledge and searches of electronic databases using the derived search string. The searches were performed on the following databases:

- Scopus (only used in [13])
- EI Compendex
- IEEE Xplore
- Science Direct

- Web of Science
- INSPEC
- ACM Digital library

We used the same online databases as in [13], and additionally also carried out the searches using Scopus. No manual searched were needed this time as all the journals/conferences previously searched manually are now indexed by at least one electronic database. Note that the conference 'International Software Metrics Symposium', which was included in [13], was discontinued since 2007.

In relation to the electronic databases, as in [13], we ensured that our search was applied to journals, magazines and conference proceedings published since December 2006, given that [13] had already covered the period from 1990 to November 2006. 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 eight studies, where all had been published when our searches were performed. These eight studies (grey color in Table 1), and another three additional relevant studies were found after searching seven different sources. In total, 1641 papers were retrieved, of which 26 represented the set of 11 primary studies (some papers were retrieved by more than one search engine). Scopus and Web of Science both retrieved the largest number of known papers – 6, followed by the ACM Digital library - 5. None of the search engines missed a known paper of a publication that should be indexed by that search engine.

Secondary Search Phase. The Secondary search phase entailed reviewing 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.

A review of all the references in the known relevant papers found no additional references (see Table 2). As expected, the number of citations received by the other papers within S1-S21 has a negative correlation with the publication year (older papers have more citations). All papers published before 2006 received at least 6 citations. Therefore, the community investigating cross- vs. within-company estimation is aware of these papers and seems to consider them relevant. S2 (Briand et al. [2]) was the most cited paper with 14 citations. Amongst the papers published within the last 10 years, the most cited paper was S8 (Kitchenham and Mendes [10]), which received 10 citations.

All the papers published after 2006 also cite [13]; therefore, results from studies S1-S10 may be indirectly considered. However, note that although some of the recent papers (e.g. S18-S21) cite [13], they only cite some isolated studies conducted after 2006. We believe that such scenario reinforces the importance of our SLR.

Study Selection Criteria and Procedures for Including and Excluding Primary Studies. The criteria for including a primary study was the same one as in [13], and comprised any study that compared predictions of cross-company models with within-company models based on analysis of project data. Studies were excluded if projects were only collected from a small number of different sources (e.g. 2 or 3 companies), and also if 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 third and last authors filtered all the papers by title and abstract; this list was reduced to 100 papers, which were then compared amongst authors aiming at a consensus; 11 studies were selected.

Study Quality Assessment Checklists. The criteria used to determine the overall quality of the primary studies was the same one used in [14] given it was an extended version of the one published in [13]. It was split into two parts (see Table 3). Part I considered the quality of the study itself and Part II the quality of the reporting provided [14]. They were originally attributed

Table 1. Coverage of Search process

			Scopus	El Compendex[1]	IEEE Xplore[1]	Science Direct[1]	Web of Science[2]	INSPEC[1]	ACM Digital Library[3]	Overall
Number of papers retrieved			603	112	7	101	791	11	27	1641
Authors	ID	Year	Did the search identify this paper?							
Lokan, C., Mendes, E.	S11	2006	YES				YES			YES
Mendes, E., Di Martino, S., Ferrucci, F., Gravino, C.	S12	2007	YES				YES		YES	YES
Lokan, C., Mendes, E.	S13	2008							YES	YES
Mendes, E., Lokan, C.	S14	2008	YES				YES			YES
Mendes, E., Di Martino, S., Ferrucci, F., Gravino, C.	S15	2008	YES				YES			YES
Lokan, C., Mendes, E.	S16	2009	YES				YES		YES	YES
Mendes, E., Lokan, C.	S17	2009							YES	YES
Kocaguneli, E., Menzies, T.	S18	2011	YES				YES			YES
Top, O., Ozkan, B., Nabi, M., Demirors, O.	S19	2011	YES	YES	YES		YES			YES
Ferrucci, F., Sarro, F., Mendes, E.	S20	2012	YES				YES		YES	YES
Minku, L.L., Yao, X.	S21	2012	YES				YES		YES	YES
Total relevant papers		11	9	1	1	0	9	0	6	26 (11 papers)
Total irrelevant papers		n/a	594	111	6	101	782	11	21	n/a

[1] Years 2006-2013; Full search string.

[2] Years 2006-2013; 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).

[3] Original search string adapted to search in abstracts returned 386,442 papers, therefore the search was adapted to (cross-company and single-company and software and effort and estimation), which returned 31 papers.

different weights (Part I weight=1.5 and Part II weight =1); however, 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 [14]:

- 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.

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.

Each reviewer assessed each paper assigned to them against each criterion. Scores attributed to our primary studies are presented in Table 3, and indicate that, according to our scoring scheme, the papers that received the highest and lowest quality scores were S11/S13/S16/S17 and S19, respectively.

Data Extraction Strategy. 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.

Data Extraction Process. 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. The first and second author checked each other's extracted data, and both first and second author checked the data extracted by the third author; the fourth author checked the data extracted by the last author.

3. PARTIAL RESULTS

The summary data used to answer research questions 1 and 2 are presented in Tables 4 and 5, respectively, and results are discussed hereafter.

Table 2. Citations and New references found

Authors	Study ID	Year	Known references found	New references
Maxwell, K., Wassenhove, L.V., Dutta, S.	S1	1999	-	0
Briand, L.C., El-Emam, K., Maxwell, K., Surmann, D., Wieczorek, I.	S2	1999	-	0
Briand, L.C., Langley, T., Wieczorek, I.	S3	2000	S2	0
Jeffery, R., Ruhe, M., Wieczorek, I.	S4	2000	S1, S2, S3	0
Jeffery, R., Ruhe, M., Wieczorek, I.	S5	2001	S1, S2, S3, S4	0
Wieczorek, I., Ruhe, M.	S6	2002	S2, S3, S4, S5	0
Lefley, M., Shepperd, M. J.	S7	2003	S1, S5	0
Kitchenham, B.A., Mendes, E.	S8	2004	S2, S3, S4, S5, S6	0
Mendes, E., Kitchenham, B.A.	S9	2004	S2, S3, S4, S5, S6, S8	0
Mendes, E., Lokan, C., Harrison, R., Triggs, C.	S10	2005	S1, S2, S3, S4, S5, S6, S7, S8, S9	0
Lokan, C., Mendes, E.	S11	2006	S1, S2, S3, S4, S5, S6, S7, S8, S9, S10	0
Mendes, E., Di Martino, S., Ferrucci, F., Gravino, C.	S12	2007	S1, S2, S3, S4, S5, S6, S7, S8, S9, S10, S11	0
Lokan, C., Mendes, E.	S13	2008	S2, S3, S4, S5, S6, S7, S8, S9, S10, S11, S12	0
Mendes, E., Lokan, C.	S14	2008	S1, S2, S3, S4, S5, S6, S7, S8, S9, S10	0
Mendes, E., Di Martino, S., Ferrucci, F., Gravino, C.	S15	2008	S1, S2, S3, S4, S5, S6, S7, S8, S9, S10, S11, S12	0
Lokan, C., Mendes, E.	S16	2009	S7, S8, S11, S12, S13, S14, S15	0
Mendes, E., Lokan, C.	S17	2009	S2, S3, S4, S5, S6, S7, S8, S9, S10, S11, S13, S14, S15	0
Kocaguneli, E., Menzies, T.	S18	2011	-	0
Top, O., Ozkan, B., Nabi, M., Demirors, O.	S19	2011	S1, S2, S4, S10, S11	0
Ferrucci, F., Sarro, F., Mendes, E.	S20	2012	S8, S9, S12, S15	0
Minku, L.L., Yao, X.	S21	2012	S13, S16, S17	0

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?

Prior to answering this research question it is important to note that there is a clear distinct difference between the evidence presented herein (see Table 4) and the evidence for this same question presented in [13]. This difference relates to an increase in the number of cross-company models to at least two (and also the number of within-company models) in many of the most recent primary studies (4 of the 11 studies). This was not observed previously, when the original SLR was carried out. As a consequence, we have the same study presenting some results where cross-company estimation models are not significantly worse than within-company estimation models for predicting effort for software/Web projects, and other results where cross-company estimation models are significantly different to within-company estimation models for predicting effort for software/Web projects. This applies to studies S14, S15, S16 and S20.

Therefore, the evidence from Table 4 suggests that two (S13, S18) studies show that cross-company estimation models are not significantly different than within-company estimation models for predicting effort for software/Web projects, and four (S14, S15, S16, S20) studies partially show that cross-company estimation models are not significantly different than within-company estimation models for predicting effort for software/Web projects. However, S14 and S15 (both in grey in Tables 4 and 5) cannot be considered independent studies since they used the same data sets employed in S11 and S12 respectively. Thus, they do not add any significant information to the results previously obtained by S11 and S12.

Further, four (S11, S12, S17, S21) studies state that cross-company models were significantly different to within-company models – S11, S12, S17: cross-company models significantly worse than within-company models; S21: cross-company models significantly superior to within-company models; another two (S16, S20) studies partially show that cross-company estimation models are significantly different to (worse than) within-company estimation models for predicting effort for software/Web projects. S19 did not test/report the statistical significance of their results, thus we assume their results as inconclusive.

Thus in summary, four studies provide evidence (or some evidence) that the prediction accuracy of cross-company models is NOT significantly different from the prediction accuracy of within-company models; six studies provide evidence (or some evidence) that the prediction accuracy of cross-company models IS significantly different flenrom the prediction accuracy of within-company models; and one study presented inconclusive results.

However, since the difference in S21 favors cross-company estimation, focusing on the posed research question, we have an overall tie. Five studies indicating that cross-company estimation models are not significantly worse (four showing no significant

Table 3. Quality Scores

Questions	S11	S12	S13	S14	S15	S16	S17	S18	S19	S20	S21
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	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	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	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0	0.5	0
2.2 Was the result of the sensitivity or residual analysis used to remove abnormal data points if necessary?	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0	0.5	0
3. Were accuracy statistics based on the raw data scale?	1	1	1	1	1	1	1	1	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	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	0.5	0.5	0.5	0.5	0.5	0.5	0.17	0.5	0.5	0.5
5. Size WC data set	1	0.33	1	0.5	0.33	1	1	0.42	0	0.5	1
Total Part I	4.5	3.88	4.5	4.0	3.88	4.5	4.5	3.59	2.0	4.0	3.5
Part II											
1. Is it clear what projects were used to construct each model?	1	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	1
3. Is it clear what cross-validation method was used?	1	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	1
Total Part II	4.0	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	10.75	9.82	10.75	10.0	9.82	10.75	10.75	9.39	7.5	10.0	9.25
Total primary study using unweighted scores	8.5	7.88	8.5	8.0	7.88	8.5	8.5	7.59	6.00	8.0	7.5

difference and one showing a difference in favor of cross-company models) and five indicating that they are significantly worse.

The evidence gathered in the original SLR [13] showed slightly different patterns, but representing a similar balanced result, with an equal number of studies (4) where cross-company models were significantly different (in this case all worse), or not different, to within-company models. With regard to inconclusive results, there were two studies in that category. It is noteworthy that in this revisited study , except for S18, the two datasets used – ISBSG and Tukutuku, are databases that contain data on more recent Web or software projects, when compared to databases used in the original SLR (Laturi, ESA, Megatec).

Table 4 also shows that the basis for evaluating how predictive accuracy varied. Some studies used independent hold-out samples; others used different types of cross-validation (e.g. 201-fold, 15-fold, 20-fold, leave-one-out cross-validation). These differences did not make it possible to perform a meta-analysis of the primary study results. These were the same patterns observed in [13].

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?

We revisited the issue raised in the original SLR of whether applying quality controls on data collection would affect the results of cross-company datasets, making their predictions at least as good as those from within-company datasets. Table 5 shows two studies (S18: Nasa93, and S21:Nasa93; Cocomo81) where databases to which quality controls were applied to during data collection were employed, and which provide evidence of cross-company models presenting similar to and significantly superior predictions than within-company models. Our results contradict those reported in the original SLR [13]; however further investigation is clearly needed, where other issues also need to be investigated relating to the estimation techniques and methods used. An in depth discussion on these issues, which are also in line with our third research question, will be reported in a future publication.

In relation to the quality evaluation of the studies, scores show no consistent pattern that the studies' quality influences the results. The score for study S18 is lower than that for studies S11 and S17; however S13, S16, S11 and S17 all share the highest quality score.

Table 4. Summary of evidence

Study	DB	Basis for Predictions	Statistical tests comparing Within (WC) to Cross-company (CC)
Cross-company model NOT significantly different to within-company model			
S13	ISBSG	Project-by-project chronological split; 206-fold-cross-validation (1 project validation set); 206-fold-cross-validation (2 projects validation set)	Wilcoxon matched pairs test on absolute residuals ($p<0.05$)
S14 (S4)	ISBSG	29-fold cross-validation (1 project validation set)	Paired Samples T Test
S15 (S4CCM2)	Tukutuku	15-fold cross-validation (1 project validation set)	Wilcoxon matched pairs test on absolute residuals ($p<0.05$)
S16 (CCR2)	ISBSG	Hold-out sample SCR2 (68 random projects)	Wilcoxon matched pairs test on absolute residuals ($p<0.05$)
S18	Nasa93, Maxwell	Nasa93: 12, 37 and 39-fold cross-validation (1 project validation set). Maxwell: 8-fold cross-validation (1 project validation set).	Wilcoxon matched pairs test on absolute residuals ($p<0.05$).
S20 (CC2)	Tukutuku	31, 18-fold cross-validation (1 project validation set).	Wilcoxon matched pairs test on absolute residuals ($p<0.05$).
Cross-company model significantly different to within-company model			
S11	ISBSG	20-fold cross-validation (4 projects validation set)	Mann-Whitney U test on absolute residuals
S12	Tukutuku	15-fold cross-validation (1 project validation set)	Wilcoxon matched pairs test on absolute residuals ($p<0.05$)
S14 (S3)	ISBSG	20-fold cross-validation (4 projects validation set)	Mann-Whitney U test on absolute residuals
S15 (S4CCM1)	Tukutuku	15-fold cross-validation (1 project validation set)	Wilcoxon matched pairs test on absolute residuals ($p<0.05$)
S16 (CCR1, CCSD1 CCSD2)	ISBSG	Hold-out samples: SCSD1(161 projects): SCSD2 (75 projects); SCR1(139 random projects)	Wilcoxon matched pairs test on absolute residuals ($p<0.05$)
S17	Finnish	Project-by-project chronological split; 201-fold-cross-validation (1 project validation set); 201-fold-cross-validation (2 projects validation set)	Paired T-test on absolute residuals ($p<0.05$). WC Superior to CC.
S20 (CC1)	Tukutuku	31, 18-fold cross-validation (1 project validation set).	Wilcoxon matched pairs test on absolute residuals ($p<0.05$). WC superior to CC.
S21	ISBSG Nasa93 Cocomo81 CocomoNasa	Hold-out samples: ISBSG2000 = 119, ISBSG2001 = 69, ISBSG = 187, CocomoNasa = 60	Wilcoxon matched pairs test on mean absolute residuals (and using Holm-Bonferroni corrections) ($p<0.05$).
Inconclusive			
S19	ISBSG, organisations' own data	Hold-out sample: 11 projects, 1 organisation's own data.	WC superior to CC, but no formal statistical significance test.

Further, with regard to the number of projects used in the cross-company model (see Table 4) there is a slight difference between studies S13, S16, S18, S20 (median = 177), and studies S11, S12, S16, S17, S20, S21 (median = 200.5); however this pattern is more noticeable when we compare the number of projects in the within-company models: the median for S13, S16, S18, S20 is 39, whereas the median for S11, S12, S16, S17, S20, S21 is 68.5. Note that the median within-company size for S21 (median = 69) is practically the same median for studies S11, S12, S16, S17, S20.

Our results show a drastic change in the pattern observed in the original SLR. While herein smaller within-company datasets are related to studies where cross-company models did not show statistically different predictions to within-company models, in Kitchenham et al. [13] the opposite was identified: 'all the studies where within-company predictions were significantly better than cross-company predictions used small within-company data sets of fair quality.' We believe that such change may be related with the techniques and methods that have been employed in more recent studies, such as the use of concept drift (S21), nearest neighbour filtering (S20), and chronological split (S13, S16, S17). An in depth discussion on this issue is also in line with our third research question, to be reported in a future publication.

Similar to [13], no clear patterns were observed regarding the size metrics used, or the procedure used to build the within company model that could explain the different results.

Table 5. 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
Cross-company models not significantly different to within-company models							
S13	No (ISBSG)	10.75	909(678)	228(206)	Min: 26 Max: 134211	IFPUG Unadjusted Function Points (FPs)	Yes
S14 (S4)	No (ISBSG)	10.0	119(90)	29	Min: 23 Max: 17688	IFPUG Adjusted FPs	Yes
S15 (S4CCM2)	No (Tukutuku)	9.82	83(68)	15	Min:1.10 Max: 3712	11 different size measures	Yes
S16 (CCR2)	No (ISBSG)	10.75	4106(CCSDS1:520;CCD S2: 539; CCR1:520; CCR2:539)	Hold-outs(SCSD1: 161; SCSD2:75; SCR1:139; SCR2: 68)	Min: 26 Max: 134211	IFPUG Unadjusted FPs	Yes
S18	Yes (Nasa93) No (Maxwell)	9.39	**Nasa93:** 88 (CC: 87) **Maxwell:** 62 (CC: 61)	**Nasa93:** SC1: 12 SC2: 37; SC3: 39 **Maxwell:** 8	NA	NA	No
S20	No (Tukutuku)	10.0	195 (CC1: 164, CC2: 177)	SC1: 31 SC2: 18	Min:1.10 Max: 5000	11 different size measures	Yes
Cross-company models significantly different to within-company models							
S11	No (ISBSG)	10.75	789(89)	12	Min:140 Max:78472	IFPUG Unadjusted FPs	Yes
S12	No (Tukutuku)	9.82	83(68)	15	Min:1.10 Max:3712	7 different size measures	Yes
S14 (S3)	No (ISBSG)	10.0	789(89)	12	Min: 140 Max: 78472	IFPUG Adjusted FPs	Yes
S15 (S4CCM1)	No (Tukutuku)	9.82	83(68)	15	Min:1.10 Max: 3712	11 different size measures	Yes
S16 (CCSDS1, CCSD2, CCR1)	No (ISBSG)	10.75	4106(CCSDS1:520;CCD S2: 539; CCR1:520; CCR2:539)	Hold-outs(SCSD1:161; SCSD2:75; SCR1:139;SCR2: 68)	Min: 26 Max: 134211	IFPUG Unadjusted FPs	Yes
S17	Yes (Finnish)	10.75	856(593)	201	Min: 86 Max: 41643	Application size in FiSMA FPs	Yes
S20	No (Tukutuku)	10.0	195 (CC1: 164, CC2: 177)	SC1: 31 SC2: 18	Min:1.10 Max: 5000	11 different size measures	Yes
S21	No (ISBSG) Yes (Nasa93) Yes (Cocomo81) Yes (CocomoNasa)	9.25	ISBSG: 826(ISBSG2000 = 168, ISBSG2001 = 224 ISBSG = 826) Cocomo81: 63(63) Nasa93: 93(93)	ISBSG: ISBSG2000 = 119, ISBSG2001 = 69 ISBSG = 187 CocomoNasa = 60	Not reported	ISBSG: IFPUG Unadjusted FPs Nasa93/Cocomo81/ CocomoNasa (Lines of Code)	Yes
Inconclusive							
S19	Yes (ISBSG)	7.5	151	8	Not reported	IFPUG Unadjusted FPs	Yes

4. DISCUSSION

Results considering only the searches carried out from 2006 to 2013 showed that eight (S16 and S20 are counted once) of the 11 primary studies provided independent evidence regarding the accuracy of cross-company prediction models when compared to within-company models. Overall, a slightly higher number of studies (six versus four, counting S16 and S20 twice) found that cross-company models gave prediction accuracy significantly different from within-company models. However, the same number of studies (five versus five) found that cross-company models gave prediction accuracy not significantly worse, since in one of those studies the accuracy of cross-company models was actually significantly higher.

Although overall slightly different, the tie trend presented in [13] remained, where three studies found that cross-company models gave prediction accuracy not significantly worse than that of within-company models; and four studies found that cross-company models gave prediction accuracy significantly worse than within-company models. We contend that the evolution of the estimation techniques and procedures for building cross-company models that have been more recently used, in addition to the within-company dataset sizes may have influenced the changes in the patterns that were previously documented in the original SLR on which this research is based upon. This issue is addressed in the third research question, which will be detailed in a future publication.

Further, in relation to the rigour with which quality assurance procedures are applied to data collection, previous studies suggested that such rigour might facilitate cross-company models to be as accurate as within-company models [2] [12] [26]. Contrary to the results detailed in [13], our results provide some support to that supposition, thus suggesting that time spent in quality control when gathering data may be well worth for companies that provide cross-company datasets to be used by other companies.

The quality of the primary studies does not seem to affect study results, so corroborating [13]. However, unlike [13], quality scores did not improve for the more recent studies. The only pattern was that the same authors authored the papers that presented the highest quality score, which could be potentially explained by their previous experience carrying out SLRs.

Contrary to the results previously found in the original SLR, ours suggest that studies where within-company predictions were not significantly better than cross-company predictions or cross-company predictions significantly better than within-company predictions employed larger within-company data sets (median = 68.5). No patterns were observed regarding the number of projects in the cross-company models, and maximum effort values.

With regard to validity threats to the results, the two main validity issues herein relate to whether this extended SLR failed to include all the relevant primary studies, and whether bias has been introduced given the first author contributed to most of the primary studies used as evidence (S11, S12, S13, S16, S17, S20). The first issue was addressed via carrying out a rigorous search strategy following the same protocol defined in [13]. With regard to the second issue, the first and second author checked each other's extracted data in order to minimize any bias that the first author could have introduced as a result of being one of the authors in most of the primary studies. In addition, one can argue that the quality assessment criteria may have been biased to reflect our personal preferences with respect to experimental procedures; however the studies that presented the highest quality scores were co-authored by an author experienced in carrying out systematic literature reviews so it can be argued that weaknesses found in earlier papers were avoided. In addition, both data extraction and quality assessment were checked by more than one author.

The implications of these results for researchers and practitioners are as follows:

- *Researchers*: We noticed a considerable impact from applying data collection quality control on the accuracy of cross-company estimation models. Moreover, it was possible to observe an evolution in the techniques and methods that have been employed for cross-company estimation models in more recent studies, such as the use of concept drift (S21), nearest neighbour filtering (S20), and chronological split (S13, S16, S17). These techniques seem to have improved the results of cross-company models when compared to models built based on larger within-company datasets. On the other hand, employing smaller single-company datasets consistently showed improved results for within-company estimation models, probably because of project heterogeneity in larger datasets. These factors require further empirical investigation, especially given the tied scenario, with exactly half of the studies showing cross-company models significantly worse than within-company models.
- Practitioners: The main goal for practitioners, such as project managers involved in cost estimation, is estimation accuracy. Having this in mind, they should be aware that larger within-company datasets do not produce better results and that they should be selective when including past project data into their datasets, properly characterizing their projects and using such characterization to identify similar projects for building their estimation models. Concerning the use of cross-company models, they should prefer quality controlled datasets and avoid using the entire dataset without employing filtering and selection mechanisms, which considerably improved cross-company results in recent studies.

Due to space constraints, the main focus of this paper was to present the findings based on the additional evidence found after extending [13]. The integration of these findings to those in [13] are the focus of another publication.

Finally, an interesting aspect to be discussed concerns the experience of extending a previously published SLR. Although the previous SLR was published in details, including a precise description of the employed protocol, this effort would be significantly higher if we did not have one of the authors as part of the team providing direct access to the employed instruments. We believe that this is the case for most published SLRs, where a complete package for a systematic update is seldom publicly available. In our point of view, currently the support for extending SLRs, and specially for integrating results for aggregated analyses, which we did not conduct in the context of this paper, is limited. Given the importance of extending SLRs and of keeping their results up to date this issue certainly deserves further attention from the empirical software engineering community.

5. CONCLUSION AND FUTURE WORK

This paper presented the partial results from extending a previously published SLR, based on 11 recent primary studies comparing prediction accuracy of estimation models built based on cross-company and within-company data. Of these 11 primary studies, two were not independent studies, and one was inconclusive, leaving eight papers. Two of these papers (S16 and S20) presented

some findings where cross-company predictions were not significantly different from within-company predictions and other findings where cross-company predictions were significantly different from within-company predictions. In addition, four other studies found that cross-company models gave prediction accuracy significantly different from within-company models (three showing improved results for within-company datasets and one for cross-company datasets), and two studies found that cross-company models presented prediction accuracy not significantly different from within-company models.

Thus, including the results from S16 and S20 in both sides, five studies indicated that cross-company estimation models are not significantly worse (four showing no significant difference and one showing a difference in favor of cross-company models) and five indicated that they are significantly worse. This tied scenario and the practical benefits of the possibility of using cross-company data indicate the need for further research on the topic.

Our results showed that strict quality control on data collection may contribute to whether 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 pattern when examining the study related factors was that studies where cross-company predictions were significantly worse than within-company predictions employed larger within-company data sets. We believe that this newly observed behaviour might be related to the techniques and methods that have been employed in more recent studies to overcome different heterogeneity issues in cross-company datasets, such as the use of concept drift (S21), nearest neighbour filtering (S20), and chronological split (S13, S16, S17). Thus, there is some objective evidence that sample size (cross or within-company) and estimation accuracy do not go hand in hand, and that the sample for building estimation models should be carefully selected/filtered based on quality control and project similarity aspects.

The results described herein will be aggregated to the results previously published, and reported in a future publication. Finally, further details, including the analysis for Question 3 will be the subject of a journal paper.

6. ACKNOWLEDGMENTS

We would like to thank the authors of the previously conducted SLR, Barbara Kitchenham and Guilherme Travassos, who participated in the initial data extraction efforts of primary studies published before November 2006 and had direct and significant contribution on the employed research method.

7. REFERENCES

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

[2] Brereton, P., Kitchenham, B. A., Budgen, D., Turner, M. and Khalil, M. (2007) Lessons from applying the systematic literature review process within the software engineering domain, Journal of Systems and Software, vol. 80, no. 4, pp. 571–583.

[3] 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.

[4] 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.

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

[6] 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.

[7] 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.

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

[9] Kalinowski, M., Card, D. N. and Travassos, G. H. (2012) Evidence-Based Guidelines to Defect Causal Analysis, IEEE Software, vol. 29, no. 4, pp. 16–18, Jul. 2012.

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

[11] Kitchenham, B. and Charters, S. (2007) Guidelines for performing Systematic Literature Review in Software Engineering, Technical Report EBSE-2007-01, 2007.

[12] 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.

[13] Kitchenham, B.A., Mendes, E., and Travassos, G.H. (2006), 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] Kitchenham, B.A., Mendes, E., and Travassos, G.H. (2007), Cross- vs. Within-Company Cost Estimation Studies: A Systematic Review, IEEE Transactions on Software Engineering, Volume 33, Issue 5, May, pp. 316-329, DOI: 10.1109/TSE.2007.1001.

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

[16] 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.

[17] 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.

[18] 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.

[19] MacDonell, S., Shepperd, M., Kitchenham, B. and Mendes, E. (2010) How Reliable Are Systematic Reviews in Empirical Software Engineering?, IEEE Trans. Softw. Eng., vol. 36, no. 5, pp. 676–687.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

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

[26] 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.