Protocol for Systematic Review of Within- and Cross-Company Estimation Models[1]

Barbara Kitchenham, Emilia Mendes, Guilherme Travassos

1. Background

Early studies of cost estimation models (see for example, Kitchenham and Taylor, 1984, or Kemerer, 1987) suggested that general purpose models such as COCOMO (Boehm, 1981) and SLIM (Putnam, 1978) needed to be calibrated to specific companies before they could be used effectively. Taking this result further and following the suggestions made by DeMarco (1982), Kok et al. (1990) suggested that cost estimation models should be developed only from within-company data. However, the problem with company-specific estimation models is that it presupposes that companies are able to collect sufficient data to construct such models.

In 1999, Maxwell et al. took a new look at the issue by analysing a multi-company benchmarking database and comparing the accuracy of a within-company model with the accuracy of a cross company model. They found the within-company model to be more accurate than the cross-company model for the specific company. In the same year, Briand and his co-workers published a report suggesting that cross-company models could be as accurate as within-company models (Briand et al., 1999). The following year, he confirmed his result on a different data set (Briand et al., 2000). Two years later Wieczorek and Ruhe (2002) confirmed the same trend using the same database employed by Briand et al. (1999). 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 within-company models were superior to cross-company models (Jeffery et al., 2000 and Jeffery et al., 2001). Later Kitchenham and Mendes, undertook two studies of web-based projects (Kitchenham and Mendes, 2004, and Mendes and Kitchenham, 2004). In both studies, a within-company model was significantly better than a cross-company model. Kitchenham and Mendes noted that one difference between the study outcomes was that the studies that found cross-company estimation models to be as good or better than within-company models used databases with strict quality control procedures for data collection.

Given the importance of knowing whether or not it is possible to use cross-company estimation models to predict effort for within-company projects, we propose a systematic review of all studies comparing cross-company and within-company software estimation models in order to assess whether there are systematic reasons for the difference in study outcomes such as the quality control associated with data collection. Thus, the aim of this systematic review is to assist software companies with small data sets decide whether or not to use an estimation model obtained from a benchmarking dataset.

[1] Copyright Kitchenham, Mendes, Travassos 2006

2. Research questions

In order to determine factors that influence the outcome of studies comparing within and between company models, our primary research questions are:

- 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- and cross-company effort estimation accuracy studies?

Since some studies also compared prediction accuracy between prediction techniques and all the studies used different experimental procedures, we also had two secondary research questions:

- Question 3: Which estimation method(s) were best for constructing cross-company effort estimation models?
- Question 4: Which experimental procedure is most appropriate for studies comparing within- and cross-company estimation models?

Population: Cross-company benchmarking data bases of software projects, and Web projects.
Intervention: Effort estimation models constructed from cross-company data, used to predict single company project effort.
Comparison Intervention: effort estimation models constructed from single company data only.
Outcomes: The accuracy of the estimates/predictions made using the within- and cross-company models.
Experimental design: Observational studies using existing multi-company and within-company data bases, where their estimates for project effort are compared using single company data hold-out sample(s) (validation sets).

3. Search Strategy

3.1 Strategy used to derive search terms

The strategy used to construct search terms is as follows:

a) Derive major terms from the questions by identifying the population, intervention and outcome;
b) Identify alternative spellings and synonyms for major terms. Please also indicate if any terms were identified via consultations with experts in the field and/or subject librarians;
c) Check the keywords in any relevant papers we already have;
d) Use the Boolean OR to incorporate alternative spellings and synonyms;
e) Use the Boolean AND to link the major terms from population, intervention and outcome.

NOTE: Whenever a database does not allow the use of complex Boolean search strings we will design different search strings for each of these data bases. The search strings will be piloted and the results of the pilot reported.

Results for a)

Population: software, Web, project.
Intervention: cross-company, project, effort, estimation, model.
Comparison: single-company, project, effort, estimation, model
Outcomes: prediction, estimate, accuracy.

Results for b) Note: bold terms were included after completing step c)

Software – application, **product**
Projects – development
Web – WWW, Internet, World-Wide Web
Method – process, system, technique, methodology, procedure
Cross – multi, **multiple**
Company – organisation, organization, **organizational**, organisational
Within – single, **company-specific**
Model - **modeling**, modelling
Effort – cost, resource
Estimation – prediction, **assessment**

Results for c)
Maxwell et al (1999) keywords: Software productivity, software metrics, software project management, software development.

Briand et al. (1999) keywords: cost estimation, classification and regression trees, analogy, analysis of variance, least-squares regression; IEEE indexing terms: software cost estimation, statistical analysis, CART, data-driven, multi-organizational database, ordinary least squares regression, software cost estimation, software cost **modeling**, stepwise ANOVA.

Briand et al. (2000) keywords: Cost estimation, Classification and Regression Trees, Analogy, Analysis of Variance, Ordinary Least-Squares Regression, replication; IEEE indexing terms: software cost estimation, common software cost **modeling**, cost models, least-squares regression, replicated **assessment**, software costs, software organizations, software **product**

Jeffery et al. (2000) keywords: Software cost estimation; Cost **modeling** techniques; Accuracy comparison; Analogy-based estimation; Ordinary least-squares regression

Jeffery et al. (2001) IEEE index terms: data analysis, least squares approximations, software cost estimation, software metrics, CART, CART-variant, ISBSG data set, International Software Standards Benchmarking Group, analogy based estimation, **company-specific** data collection, **company-specific** models, cost estimation, estimation accuracy, large-scale industrial data set, **modeling** techniques, multi-company data, multi-**organizational** data, ordinary least squares regression, public domain metrics, robust regression, software cost estimates, software development effort estimation, stepwise ANOVA .

Wieczorek and Ruhe (2002) IEEE indexing terms: software cost estimation, software development management, statistical analysis, **company-specific** data, cost factors, cost models, data set description, **multiple** company data, multiple-organizational data, software cost estimation, statistical estimation methods.

Kitchenham and Mendes (2004): effort estimation, Web projects, cross-company estimation models, within-company estimation model, regression-based estimation models.

Mendes and Kitchenham (2004) second paper keywords: effort estimation, Web projects, cross-company, estimation models, within-company estimation model, regression-based estimation models, replication study, case-based reasoning. IEEE index terms: Internet, case-based reasoning, project management, regression analysis, software cost estimation, software development management, software metrics, Web projects, case-based reasoning, cross-company estimation models, effort estimation, forward stepwise regression, regression-based estimation models, replication study, within-company estimation model

Mendes et al. (2005) [15] keywords**:** effort estimation, software projects, cross-company estimation models, within-company estimation model, regression-based estimation models, replication study.

Results for d)

1. (software OR application OR product OR Web OR WWW OR Internet OR World-Wide Web OR project OR development)
2. (method OR process OR system OR technique OR methodology OR procedure)
3. (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)
4. (model OR modeling OR modelling)
5. (effort OR cost OR resource)
6. (estimation OR prediction OR assessment)

Results for e)

(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)

3.2 The Search Process

3.2.1 The Initial Search Phase

The initial phase our search process involves identifying candidate primary sources based on our own knowledge and searches of electronic databases using the search strings defined in Section 3.1. The electronic searches will be based on:

Databases

- IEEExplore
- ACM Digital library
- Science Direct
- El Compendex
- Web of Science
- INSPEC

Individual journals

- Empirical Software Engineering
- Information and Software Technology
- Software Process Improvement and Practice
- Management Science

In addition we will ensure coverage of conferences in which the publications we know about have appeared:

- Conferences:
    - International Metrics Symposium
    - International Conference on Software Engineering
    - Evaluation and Assessment in Software Engineering (manual search)

We will access the coverage of the search process in terms of the number of papers it identifies that we already know about.

3.2.2 The Secondary Search Phase

The second phase of our search process will support the electronic search activity using two methods:

- To review the references of each of the primary sources identified in the first phase looking for any other candidate primary sources.
- To contact researchers who authored the primary sources in the first phase, or who we believe could be working on the topic.

These activities are defined below.

3.2.2.1 Reviewing reference lists

We will check all references in selected papers for all other relevant reports/papers. This process will be repeated until no further reports/papers seem relevant. Whenever we find other relevant papers their publication source (conference, journals) will be added to our current search list as long as it is indexed electronically. However, since the first papers on this topic appeared in 1999, we can restrict our search to the years 1999-2005.

3.2.2.2 Contacting researchers

We will contact the following researchers with a list of the papers that we know they have published on the topic and enquire whether they have any unpublished papers or technical reports:

- Khaled El-Emam
- Magne Jørgensen
- Martin Shepperd
- Katrina Maxwell
- Lionel Briand
- Ross Jeffery

3.2.3 Search Process Documentation

The search will be documented in the format shown in Table 1.

Table 1 Search process documentation

Data Source	Documentation
Electronic databases	Name of database: IEEExplore Search strategy: (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) Search characteristics for each database: (X) allows for nested Boolean searches ( ) allows only for simple Boolean searches (X) indexes full-text (X) indexes abstract (X) indexes title ( ) indexes literature written in the following languages: English. Date, time and location of search: Years covered by search for each database: 1999 to 2005
Electronic Journals individual search	Name of journal Search strategy for each journal Search characteristics for each journal: ( ) allows for nested Boolean searches ( ) allows for simple Boolean searches ( ) indexes full-text ( ) indexes abstract ( ) indexes title ( ) indexes literature written in the following languages: ____, ____, ____, ____ etc. Date, time and location of search: Years covered by search for each journal
Journal Hand Searches	Name of journal Years searched Any issues not searched
Conference proceedings Hand Searches	Title of proceedings Name of conference (if different) Title translation (if necessary) Journal name (if published as part of a journal)
Efforts to identify unpublished studies	Research groups and researchers contacted (Names and contact details) Research web sites searched (Date, time and URL)
Other sources	Date, time Searched URL Any specific conditions pertaining to the search

References will be stored in an excel spreadsheet. Each reference will be indexed by first author's surname + year of publication. Whenever there is more than one value for the same surname + year of publication a letter will be added to the index. We will also identify if a reference is primary, i.e. retrieved from our search, or secondary, i.e., identified from a paper's reference list.

4. Study selection criteria and procedures for Including and Excluding Primary Studies

Criteria for including study: any study that compares predictions of cross-company models with within-company models *based on analysis of project data*.

Criteria for excluding study: We will exclude studies where projects were only collected from a small number of different sources (e.g. 2 or 3 companies), and where models derived from a single company data set were compared with predictions from a general cost estimation model.

Preliminary selection process:

The three researchers will apply the search strategy to identify potential primary studies. Each researcher will use a different set of databases/journals/conference proceedings. All researchers will check titles and abstracts of all potential primary studies against inclusion criteria. Results will be checked and any disagreements discussed and resolved. If resolution is impossible the study will be included.

Final selection process:

Copies of all papers included as a result of the initial study will be reviewed by at least two of the researchers. A random selection of papers (up to 20%) will be reviewed by all researchers. Note if fewer than 10 papers are identified each reviewer will review every paper. This review will finalise the selection of papers to be included in the data extraction process. Any disagreements in papers jointly reviewed will be discussed and resolved. If resolution of the dispute is not possible, the paper will be included.

5. Study quality assessment checklists

The criteria used to determine the overall quality of the primary studies includes six top-level questions and an additional quality issue. The overall quality score for a paper will range from 0 to 7, representing very poor and excellent quality, respectively. Top-level questions without sub-questions will be answered Yes/No/Partially, corresponding to scores 1, 0, and 0.5 respectively. Whenever a top-level question has sub-questions, scores will be attributed to each sub-question such that the overall score for the top-level question will range between 1 and 0. For example, question 1 has five sub-questions, thus each “Yes”, “No”, and “Partially” for a sub-question contributes scores of 0.2, 0, and 0.1 respectively. Note that in most cases, “Partially” means that we had some reason to infer the issue was addressed correctly but we could not confirm our inference from what was reported in the paper.

The six main questions are:

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 we consider is the size of the within-company data set, measured according to the criteria presented below. Whenever a study uses more than one within-company data set, the average score will be 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 is 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 will assess each paper assigned to them against each criterion. For each paper where partial or no information is available, we will e-mail the first author with a request for the missing information.

6. Data extraction strategy

6.1 Required Data

For each paper remaining after the selection process has been completed. The researchers will extract the data shown in Table 2.

Table 2 Data Extraction Form Completed for Maxwell et al, 1998

Data item	Value	Additional notes
Data Extractor		
Data Checker		
Study Identifier	S1	
Application domain	Space, military and industrial	
Name of database	European Space Agency (ESA)	
Number of projects in database (including within-company projects)	108	
Number of cross-company projects	60	
Number of projects in	29	

within-company data set		
Size metric(s): FP (Yes/No) Version used: LOC (Yes/No) Version used: Others (Yes/No) Number:	FP: No LOC: Yes (KLOC) Others: No	
Number of companies	37	
Number of countries represented	8	European only
Were quality controls applied to data collection?	No	
If quality control, please describe		
How was accuracy measured?	Measures: R^2 (for model construction only) MMRE Pred(25) r (Correlation between estimate and actual)	
Cross-company model		
What technique(s) was used to construct the cross-company model?	A preliminary productivity analysis was used to identify factors for inclusion in the effort estimation model. Generalised linear models (using SAS). Multiplicative and Additive models were investigated. The multiplicative model is a logarithmic model.	
If several techniques were used which was most accurate?	In all cases, accuracy assessment was based on the logarithmic models not the additive models.	It can be assumed that linear models did not work well.
What transformations if any were used?	Not clear whether the variables were transformed or the GLM was used to construct a log-linear model	Not important: the log models were used and they were presented in the raw data form – thus any accuracy metrics were based on raw data predictions.
What variables were included in the cross-company model?	KLOC, Language subset, Category subset, RELY	Category is the type of application. RELY is reliability as defined by Boehm

		(1981)
What cross-validation method was used?	A hold-out sample of 9 projects from the single company was used to assess estimate accuracy	
Was the cross-company model compared to a baseline to check if it was better than chance?	Yes	The baseline was the correlation between the estimates and the actuals for the hold-out.
What was/were the measure(s) used as benchmark?	The correlation between the prediction and the actual for the single company was tested for statistical significance. (Note it was significantly different from zero for the 20 project data set, but not the 9 project hold-out data set.)	
Within-company model		
What technique(s) was used to construct the within-company model?	A preliminary productivity analysis was used to identify factors for inclusion in the effort estimation model. Generalised linear models (using SAS). Multiplicative and Additive models were investigated. The multiplicative model is a logarithmic model.	
If several techniques were used which was most accurate?	In all cases, accuracy assessment was based on the logarithmic models not the additive models.	It can be assumed that linear models did not work well.
What transformations if any were used?	Not clear whether the variables were transformed or the GLM was used to construct a log-linear model	Not important: the log models were used and they were presented in the raw data form – thus any accuracy metrics were based on raw data predictions.
What variables were included in the within-company model?	KLOC, Language subset, Year	
What cross-validation method was used	A hold-out sample of 9 projects from the single company was used to assess estimate accuracy	
Comparison		

What was the accuracy obtained using the cross-company model?	Accuracy on main single company data set (log model):
	n=11 (9 projects omitted)
	MMRE=50%
	Pred(25)=27%
	r=0.83
	Accuracy on single company hold out data set
	n=4 (5 projects omitted)
	MMRE=36%
	Pred(25)=25%
	R=0.16 (n.s)
		Using the 79 cross-company projects, Maxwell et al. identified the best model for that dataset and the best model for the single company data. The two models were identical.
		This data indicates that for all the single company projects:
		n=15
		Pred(25)=26.7% (4 of 15)
		MMRE=46.3%

What was the accuracy obtained using the within-company model?	Accuracy on main single company data set (log model):
	n=14 (6 projects omitted)
	R^2=0.92
	MMRE=41%
	Pred(25)=36%
	r=0.99
	Accuracy on single company hold out data set
	n=6 (3 projects omitted)
	MMRE=65%
	Pred(25)=50% (3 of 6)
	r=0.96	

What measure was used to check the statistical significance of prediction accuracy (e.g. absolute residuals, MREs)?	Estimated and actual effort	
What statistical tests were used to compare the results?	r, correlation between the prediction and the actual	
What were the results of the tests?		
Data Summary		
Data base summary (all projects) for size and effort metrics.	Effort min: 7.8 MM Effort max: 4361 MM Effort mean: 284 MM Effort median: 93 MM Size min: 2000 KLOC Size max: 413000 KLOC Size mean: 51010 KLOC Size median: 22300 KLOC	KLOC: non-blank, non-comment delivered 1000 lines. For reused code Boehm's adjustment were made (Boehm, 1981). Effort was measured in

		man months, with 144 man hours per man month
With-company data summary for size and effort metrics.	Effort min: Effort max: Effort mean: Effort median: Size min: Size max: Size mean: Size median:	Not specified
Study quality		
Analysis process		
1) Is the analysis process description complete?		
1.1) Was the data validated investigated to identify outliers and to assess distributional properties before analysis?	Yes	
1.2) Was the result of the investigation used appropriately?	Partial	
1.3) Were the resulting estimation models subject to sensitivity or residual analysis?	yes	Plots of residuals vs. fitted to check for violations of LSR assumptions; Ramsay RESET test used to determine if there were omitted variables. Cook Weisberg test used to detect heteroscedasticity. Studentized residuals and Cook's D used to detect presence of influential outliers
1.4) Was the result of the sensitivity or residual analysis used appropriately?	No	No need to adapt based on Cook's D
1.5) Were accuracy statistics based on the raw data scale?	Partial	
Other aspects of the Study Quality		
2) Is it clear what projects were used to construct each model?	Yes	

3) Is it clear how accuracy was measured?	Yes	
4) Is it clear what cross-validation method was used?	Yes	
5) Were all model construction methods fully defined?	Yes	
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?	No	
6.2) Was the comparison based on a completely independent hold out sample or n-fold cross-validation for the within-company model?	Yes	
7) Size of the data set	Good quality	

6.2 Data extraction process

For each paper a researcher will be 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 will be 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 will be held in word tables, one file per paper, using the table format shown in Table 2. After the extracted data has been checked a single word file containing the final agreed data will be constructed. No inter-rater agreement statistics will be calculated since our process in intended to achieve 100% agreement, i.e. whenever we are unable to understand what was reported in the primary study we will approach the authors for clarification.

7. Synthesis of the extracted data:

7.1 Question 1

Question 1 is “What evidence is there that cross-company estimation models are no significantly worse than within-company estimation models for predicting effort for software/Web projects?”

Results will be tabulated as shown in Table 3. We should consider providing forest plots to provide a clearer summary of the results. However, for a forest plot we would need not just an accuracy statistic but a measure of the standard deviation of that statistic. The candidate measures are MMRE, Pred(25), and *r* (correlation between actual and estimate).

However, to generate common significance tests and determine confidence intervals for MMRE or *r* for all the reported studies, we would need information that is seldom reported such as the actual and estimate for each project. Therefore we will also ask those who have published previous studies to provide us with the absolute residuals for the validation sets used. This does not violate the confidentiality of the data sets since only a single value is required and only for a small subset of the data. This information will be used to carry out a meta analysis based on the difference between the absolute residuals.

We also wish to make a recommendation as part of this protocol that such data be published in all future papers on the topics so that proper meta analysis can be performed.

7.2 Question 2

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

This question will be addressed by tabulating the studies as shown in Table 4. We will report the studies in subgroups depending on whether the study suggested that cross-company models were at least as good as with-company models, or the study suggested that within-company models were significantly better than cross-company models.

7.3 Question 3

Question 3 is “Which estimation method(s) were best for constructing Cross-company effort estimation models?”. This will be investigated by tabulating the studies as shown in Table 5. In addition, we will also provide another two summary tables as shown in Tables 7 and 8.

7.4 Question 4

Question 4 is “Which experimental protocol is most appropriate for studies comparing within- and cross-company estimation models?”. This will be investigated by indicating the studies, as shown in Table 8a, 8b and 8c.

8 Schedule for Review

Table 3 Summary of evidence of accuracy of within- and cross-company estimates (completed for the Maxwell et al, 1998 study)

Study	Database	Basis for Predictions (Cross-validation for within-company model)	Statistical tests comparing Within (WC) to Cross-company (CC)
Cross-company model not significantly worse than within-company model			
Cross-company model significantly worse than within-company model			
Inconclusive			
S1	ESA	Independent hold-out (9 projects)	Correlation analysis between actual and estimate, no formal statistical significance test

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							
Cross-company model significantly worse than within-company models							
Inconclusive							
S1	Partially (ESA)	5.77	108 (60)	29	Min: 1123.2 Max: 627984	KLOC	Yes
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							

Table 5 Estimation methods

Study	DB	Cross-company predictions			Within-company predictions		
		MMRE	**Pred(25)**	**MdMRE**	**MMRE**	**Pred(25)**	**MdMRE**
Cross-company models not significantly worse than within-company models							
Cross-company model significantly worse than within-company models							
Inconclusive							
S1	ESA	GLM: 36% (4 pjs)	GLM: 25% 11.1% (adjusted for missing predictions		GLM: 65%	GLM: 50% 33% (adjusted for missing predictions	
General Linear Model (**GLM**)							

Table 6 Best Estimation Method

Study	Cross-Company	Within-Company

Table 7 Effectiveness of Different Techniques

Summary	Frequency method evaluated	Frequency best method for cross-company models	Frequency best method for within-company models

Table 8a Study procedure factors – Model construction options

Options for data preparation	**Pros**	**Cons**	**Used in Studies**
Data set transformed in a standard way independent of construction method	Easiest approach.	Risks using an inappropriate transformation.	
Data set transformed appropriately for each model construction method	Theoretically the best option.	More time consuming	
Options for sensitivity analysis	**Pros**	**Cons**	**Used in studies**
Performed	Good practice because it reduces possibility of results being biased as a result of atypical data values.		
Not performed	Simplest option when evaluating many different estimation methods.	Bad practice. Results may be biased by atypical data values.	
Options for sensitivity analysis methods	**Pros**	**Cons**	**Used in studies**
Module residual analysis	Identifies projects that have a large residual. Re-analyzing the data with those projects omitted tests the resilience of the model. Can be undertaken for any prediction model, statistical or non-statistical.		
Influence analysis	Identifies projects that have large residuals and have a large influence on the model.	Currently only feasible for regression.	
Comparison with naïve model	Provides assurance that the model is better than a simple baseline model.	Researchers may disagree about the baseline model.	
Comparison with random model	Provides assurance that the model is better than simple guesswork.	This is a minimal criterion for model validation.	
Options for prediction validation	**Pros**	**Cons**	**Used in Study**
Independent hold-out sample	Theoretically the best option particularly if there is a prior justification for the hold-out e.g. using projects started after a certain date as the hold-out.	Not feasible for small data sets	
N-fold cross-validation where N<sample size (restricted to ensure one prediction per project)	A reasonable option if there is no obvious hold-out criteria. With a small data set hold-out		

	samples could be at least 2 projects.		
N-fold cross-validation where N<sample size (allowing multiple predictions for each project)	Reduces bias in estimates of mean and variance of absolute residuals when comparing different estimation methods (see **Error! Reference source not found.**)	Complicates the analysis because an additional procedure is needed to determine the prediction to be used in any statistical test. If the average is used, this is biased unless each project had an equal number of predictions.	
N-fold cross-validation where N=sample size	The easiest option practically, usually supported by options in statistical tools.	The worst option theoretically since statistics based on a leave-one-out cross-validation are functionally related to statistics based on predictions without cross-validation.	
Options for basis of statistical significance testing	**Pros**	**Cons**	
MRE		The metric is inherently biased	
Absolute residual	The metric is unbiased.		
Options for statistical significance testing	**Pros**	**Cons**	**Used in studies**
Performed	Gives an objective assessment of whether one model is better than another.		
Not performed		Does not allow a definitive assessment of whether or not one model is better than the other.	

Table 8b Study procedure factors – Model construction options

Option for within-company selection	**Pros**	**Cons**	**Used in Study**
Part of the cross company data set	Will have collected data according to the database standards.		
Independent data set	More representative of companies that want to utilize that cross-company data. Easier for experiments since it is easier to vary data set properties to investigate which factors affect the quality of estimates. (There are probably more within-company data sets than cross-company data sets.)	May not have collected appropriate data.	
Options for cross-company model construction	**Pros**	**Cons**	**Used in Studies**
Stepwise approach independent of within company model		There is a risk of producing a model that cannot be used on the single company data (because input variables may not have been collected).	
Re-calibration of stepwise model obtained from all data (within- and cross-company data)	Ensures that the model can be used on the single company data. Realistic approach for a company that has a reasonable amount of their own data.	The cross-company model is not independent of the within-company model.	
Stepwise approach based on measures collected on the within- company data set that are also collected by the cross-company data set	Ensures that the model can be used on the single company data. Realistic approach for a company that has a reasonable amount of their own data. The cross-company model is only dependent on the within-company model with respect to the choice of metrics not the functional form of the model.		
Cross-company model includes within-company projects	Realistic approach for companies with any data	The cross-company model is not independent of the within-company model.	
Options for within-company model	**Pros**	**Cons**	**Used in studies**

construction			
Stepwise based on data available in benchmarking databases	Suitable if the single company is part of the cross-company data set.		
Stepwise based on data collected in the company	Suitable if the single company is not part of the cross-company data set.		
Options for model construction method	**Pros**	**Cons**	**Used in studies**
Regression (OLS, Stepwise, Robust)	The most commonly used method. All statistical tools support regression.		
ANOVA (effort or productivity)		Not automated. In most cases equivalent to regression.	
CART (effort or productivity)		Requires a specialist tool.	
Analogy			
Genetic programming		May be difficult for non-experts	

Table 8c Study procedure factors – Reporting options

Options for accuracy statistics	**Pros**	**Cons**	**Used in studies**
Pred(25)	Simple measure. Can be adjusted correctly to allow for failure to make a prediction.		S1, S2, S4, S8, S9, S10
MMRE		Ratio-based measures are unstable and can lead to incorrect assessments (see **Error! Reference source not found.**).	S1, S2, S4, S7, S8, S9, S10
MdMRE	Used in other disciplines (e.g. economics).	Ratio-based measures are unstable and can lead to incorrect assessments. (see **Error! Reference source not found.**).	S2, S3, S4, S5, S6, S8, S9, S10
BalancedMRE		Ratio-based measures are unstable and can lead to incorrect assessments. (see **Error! Reference source not found.**).	S7
Mean Absolute residual	Not as unstable or biased as ratio-based accuracy statistics.	Inappropriate for non-Normal distributions. Does not have an obvious baseline value.	S8, S9
Median absolute residual	Not as unstable or biased as ratio-based accuracy statistics.	Does not have an obvious baseline value.	S8, S9
Options for information	**Pros**	**Cons**	**Used in studies**

reported			
Selected accuracy statistics for within-company and cross-company predictions	Simplest option	This level of information is unsuitable for meta-analysis.	All studies
Mean difference between MRE for within- and cross-company predictions		This level of information is unsuitable for meta-analysis.	None
Mean difference between absolute residuals for within- and cross-company predictions		This level of information is unsuitable for meta-analysis.	None
Mean difference between MRE with standard error	Minimal data sufficient for restricted meta-analysis.	MRE is a biased statistic which would bias any meta-analysis.	None
Mean difference between absolute residuals with standard error	Minimal data required for restricted meta-analysis. MAR is unbiased.		None
Effort actual and predicted for each single company project	Sufficient data for meta-analysis. Makes testing a new model construction method easier (assuming the raw data is available to researchers) – the new method can be easily compared with previous results.	Single-company effort values may be commercially sensitive.	None
Residuals for each method for single company projects	Sufficient data for meta-analysis. Actual effort values remain confidential. Makes testing a new model construction method easier (assuming the raw data is available to researchers).		None

8 References

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

[2] Briand, L.C., K. El-Emam, K. Maxwell, D. Surmann, I. Wieczorek. An assessment and comparison of common cost estimation models. Proceedings of the 21st International Conference on Software Engineering, ICSE 99, 1999, pp 313-322.

[3] Briand, L.C., T. Langley, I. Wieczorek. A replicated assessment of common software cost estimation techniques. Proceedings of the 22nd International Conference on Software Engineering, ICSE 20, 2000, pp 377-386.

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

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

[6] Jeffery, R., M. Ruhe and I. Wieczorek. Using public domain metrics to estimate software development effort. Proceedings 7th International Software Metrics Symposium, London, IEEE Computer Society Press, 2001, pp 16-27.

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

[8] Kitchenham, B.A. and N.R. Taylor. Software cost models. ICL Technical Journal, May 1984, pp73-102.

[9] Kitchenham, B.A., and E. Mendes. A Comparison of Cross-company and Within-company Effort Estimation Models for Web Applications, Proceedings 8th International Conference on Empirical Assessment in Software Engineering EASE 2004, Computer Society Press, 2004, pp 47-55.

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

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

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

[13] Mendes, E. and B.A. Kitchenham. Further Comparison of Cross-Company and Within Company Effort Estimation Models for Web Applications. Proceedings 10th International Symposium on Software Metrics. Metrics 2004, Chicago, Illinois September 11-17th 2004, IEEE Computer Society, ISBN 0-7695-21290, pp 348-357, 2004.

[14] Wieczorek, I. and M. Ruhe. How valuable is company-specific data compared to multi-company data for software cost estimation? Proceedings 8th International Software Metrics Symposium, Ottawa, IEEE Computer Society Press, June 2002, pp 237-246.

[15] 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.

Appendix A1

Computer Science/Software Engineering Journals searched using ScienceDirect (this may be a subset of the entire set of indexed journals since our Institutions may not subscribe to all the publications)

Ad Hoc Networks
Advanced Engineering Informatics
Advances in Engineering Software
AEU - International Journal of Electronics and Communications
Applied Soft Computing
Artificial Intelligence
Artificial Intelligence in Engineering
Artificial Intelligence in Medicine
Biometric Technology Today
Card Technology Today
Cognitive Science
Cognitive Systems Research
Computational Biology and Chemistry
Computational Geometry
Computational Statistics & Data Analysis
Computer-Aided Design
Computer Aided Geometric Design
Computer Audit Update
Computer Communications
Non-Computer Compacts
Computer Fraud & Security
Computer Fraud & Security Bulletin
Computer Graphics and Image Processing
Computer Languages
Computer Languages, Systems & Structures
Computer Law & Security Report
Computer Methods in Applied Mechanics and Engineering
Computer Methods and Programs in Biomedicine
Computer Networks
Non-Computer Networks (1976)
Computer Networks and ISDN Systems
Computer Physics Communications
Computer Physics Reports
Computer Programs in Biomedicine
Computer Speech & Language
Computer Standards & Interfaces
Computer Vision and Image Understanding
Computer Vision, Graphics, and Image Processing
Computerized Medical Imaging and Graphics
Computers and Biomedical Research
Computers & Chemistry
Computers and Electronics in Agriculture
Computers & Geosciences
Computers and Geotechnics
Computers & Graphics

Computers & Security
Non-Computers and Standards
Computers & Structures
Computers & Urban Society
Computers in Biology and Medicine
Computers in Human Behavior
Computers in Industry
Computing Systems in Engineering
CVGIP: Graphical Models and Image Processing
CVGIP: Image Understanding
Data & Knowledge Engineering
Non-Data Processing
Decision Support Systems
Design Studies
Differential Geometry and its Applications
Digital Investigation
Digital Signal Processing
Discrete Applied Mathematics
Displays
Non-Education and Computing
Electronic Commerce Research and Applications
Electronic Notes in Theoretical Computer Science
Engineering Analysis with Boundary Elements
Engineering Applications of Artificial Intelligence
Environmental Software
Non-Estuarine and Coastal Marine Science
Non-Euromicro Newsletter
Expert Systems with Applications
Finite Elements in Analysis and Design
Future Generation Computer Systems
Fuzzy Sets and Systems
Graphical Models
Graphical Models and Image Processing
Image and Vision Computing
IMPACT of Computing in Science and Engineering
Information and Computation
Non-Information and Control
Information Fusion
Information & Management
Information and Organization
Information Processing Letters
Information Processing & Management
Information Sciences
Information Sciences - Applications
Information Security Technical Report
Information and Software Technology
Information Storage and Retrieval
Information Systems
Infosecurity Today
Integration, the VLSI Journal

Intelligent Data Analysis
Interacting with Computers
Non-Interfaces in Computing
International Journal of Approximate Reasoning
International Journal of Electrical Power & Energy Systems
International Journal of Human-Computer Studies
International Journal of Man-Machine Studies
ISPRS Journal of Photogrammetry and Remote Sensing
Journal of Algorithms
Journal of Biomedical Informatics
Journal of Computational Physics
Journal of Computer and System Sciences
Journal of the Franklin Institute
The Journal of Logic and Algebraic Programming
Journal of Microcomputer Applications
Journal of Molecular Graphics
Journal of Molecular Structure: THEOCHEM
Journal of Network and Computer Applications
Journal of Parallel and Distributed Computing
The Journal of Strategic Information Systems
Journal of Systems and Software
Journal of Systems Architecture
Journal of Visual Communication and Image Representation
Journal of Visual Languages & Computing
Knowledge-Based Systems
Laboratory Automation & Information Management
Mechanical Systems and Signal Processing
Medical Image Analysis
Microelectronic Engineering
Microelectronics Journal
Microelectronics Reliability
Microprocessing and Microprogramming
Microprocessors
Microprocessors and Microsystems
Network Security
Neural Networks
Neurocomputing
Optical Fiber Technology
Optical Switching and Networking
Parallel Computing
Pattern Recognition
Pattern Recognition Letters
Performance Evaluation
Philips Journal of Research
Photogrammetria
Real-Time Imaging
Robotics
Robotics and Autonomous Systems
Science of Computer Programming
Signal Processing

Signal Processing: Image Communication
Speech Communication
Telecommunications Policy
Telematics and Informatics
Tetrahedron Computer Methodology
Theoretical Computer Science
Transportation Research Part C: Emerging Technologies
Non-USSR Computational Mathematics and Mathematical Physics
Web Semantics: Science, Services and Agents on the World Wide Web
World Patent Information

Computer Science/Software Engineering publications searched using IEEExplore

To add at once we have a final version of the protocol end as it is a very long list

Computer Science/Software Engineering publications indexed by ACM digital library

To add at once we have a final version of the protocol end as it is a very long list