Comparing Local and Global Software Effort Estimation Models – Reflections on a Systematic Review

Stephen G. MacDonell
School of Computing and Mathematical Sciences
Auckland University of Technology
Private Bag 92006
Auckland 1142, New Zealand
stephen.macdonell@aut.ac.nz

Martin J. Shepperd
School of Computing, Information Systems and Mathematics
Brunel University
Uxbridge, Middlesex UB8 3PH, UK
Martin.Shepperd@brunel.ac.uk

Abstract

BACKGROUND: the availability of multi-organisation data sets has made it possible for individual organisations to build and apply management models, even if they do not have data of their own. In the absence of any data this may be a sensible option, driven by necessity. However, if both cross-company (or global) and within-company (or local) data are available, which should be used in preference?
PROBLEM: several research papers have addressed this question but without any apparent convergence of results.
METHOD: we conduct a systematic review of empirical studies comparing global and local effort prediction systems.
RESULTS: we located 10 relevant studies: 3 supported global models, 2 were equivocal and 5 supported local models.
CONCLUSION: the studies do not have converging results. A contributing factor is that they have utilised different local and global data sets and different experimental designs thus there is substantial heterogeneity. We identify the need for common response variables and for common experimental and reporting protocols.

Keywords*: D.2.9.b Cost estimation, project effort prediction, systematic review, empirical analysis.*

1. Introduction

The collection and analysis of project cost[1] data by software organisations remains challenging, for several reasons. Some are self-evident – software projects take time, often substantial amounts of time, so the collection of data from a number of projects may require a number of years. Smaller or newly established organisations may have particular difficulty in building a sufficiently large data set in a timely manner to enable useful analysis. Established organisations that change their practices could also find it difficult to build relevant predictive models – utilising new development tools, adopting a novel process, or losing significant staff may render useful analysis difficult. The establishment and maintenance of a measurement programme also demands an ongoing investment of time and resources that organisations may not consider justified. Irrespective of the reason(s), a lack of locally collected data could preclude data-informed analysis of practice in the form of benchmarking, assessment of current activities, estimation of future tasks or consideration of improvements that might be anticipated as a result of practice changes.

In such circumstances an organisation might consider using a data set based on projects undertaken by other organisations. This could be based on the assumption that their local practices are likely to be represented by data in such a global set – from organisations of a similar size, or in the same industry sector; or in terms of projects that are similar to those that they undertake. Even if they have their own data they may believe that analysis based on a much larger and more diverse data set might enable them to develop richer and more representative models of practice, useful for benchmarking or for predictive modelling of attributes such as development effort and duration.

The primary question to be considered in this latter respect is therefore: *What evidence is there that cross-company estimation models are at least as good as within-company estimation models for predicting effort*

[1] Strictly speaking we mean effort data, however, labour is generally the largest and least predictable component of project cost.

0-7695-2886-4/07 $20.00 © 2007 IEEE
DOI 10.1109/ESEM.2007.45

for software projects? We address this question by means of meta-research – a systematic review of previously reported work. In doing so we have two aims: to consider whether the evidence is converging with respect to the relative worth of using global predictive models; and to contribute to the growing number of systematic reviews undertaken and reported in empirical software engineering.

This work is also intentionally a replication of that reported by Kitchenham *et al.* [1, 2]. In conjunction with those authors we set out to undertake independent reviews addressing the same research question. A comparative analysis of the two systematic review outcomes is the subject of a forthcoming paper.

The remainder of the paper is structured as follows. As the work is itself a systematic review of prior research we do not provide a literature review *per se* – rather, in the next section we briefly present relevant background material regarding effort estimation and the emergence of global data sets in empirical software engineering. We then describe our review in terms of the protocol used and the process and outcomes of the data extraction phase. We go on to describe the results of the aggregation and analysis of the studies collated from the literature, addressing the primary research question stated above. We then reflect on our experiences in undertaking the review, identifying in particular the questions that (for us) remain open with respect to this research technique. The paper is then closed with a short concluding section.

2. Brief background to the review topic

Over the years there has been a significant evolution in approaches to building cost, effort and schedule prediction models. For an extensive review see [3]. However, to summarise, the dominant philosophy of the 1970s and 1980s was that by building sophisticated, parameterised models it would be possible to take into account the differences in software development environments. A good and influential example of this type of thinking is COCOMO [4] that involves three different models (for different development modes) and 14 cost drivers to take into account different tools, environments, non-functional requirements and so forth. An obvious advantage is that the generality of the model replaces the need to collect local data. Whilst this might seem attractive, in practice there is little evidence to suggest that general models work well outside of the environments in which they were developed [5] or that at the very least recalibration is required [6].

Subsequently, most cost prediction research has assumed that it is necessary to develop local models (in terms of both parameters and structure) using local data. Model development can occur using a variety of techniques, ranging from relatively simple approaches such as least squares regression [7] to more sophisticated machine learners such as artificial neural nets [8] and case-based reasoners [9]. Machine learners work by inductively finding patterns in the training data, in other words by presenting examples of past completed projects the learner inductively builds a prediction system for future projects. Clearly, for this to be effective, it is necessary for the training examples to be representative of the future cases to be predicted. Thus local training data are important. Unfortunately (as described in the previous section) local data may not always be available, hence there is interest in using data that has been collected by other organisations.

Particularly in the last decade, there have been an increasing number of initiatives designed to collect data from multiple organisations. Primary examples of this are the commonly-named 'Finnish' data set [10] (also known as Experience, Laturi and STTF), the International Software Benchmarking Standards Group repository (ISBSGs) [11] and, in more recent times, the Tukutuku data set [12]. Other public but comparatively less diverse data sets are the NASA IV&V data set [13] and the European Space Agency data set referred to in review papers C2 and J2.

Organisations therefore have a choice – to expend effort and resources developing and maintaining their own data set or to rely on the data available from one (or more?) of these global repositories. We are therefore interested in understanding how well models developed from these global data sets perform when compared to those derived from locally collected data.

3. The review

The review was conducted by the two authors over a period of around five months in the latter half of 2005. We adopted a review process based on the emerging guidelines being advocated by Kitchenham and others [14], these being adapted primarily from those used in evidence-based medicine. Our review therefore comprised three major activities: protocol development, data extraction, and data aggregation. Each activity incorporated a degree of within-activity refinement based on discussions between the authors. There was also a degree of interplay between the activities, to some extent mimicking a software process model as waterfall-like with iteration and activity feedback. The three activities and their respective outcomes are now described in detail.

3.1. The protocol

We first developed a protocol for the review (as per the guidelines) that specified the research topic and research question of interest and a PICO definition:

Research Topic: A review of the effectiveness of within and between company software effort estimation models.

Research Question: What evidence is there that cross-company estimation models are at least as good as within-company estimation models for predicting effort for software projects?

PICO definition:

- *Population*: local and global data sets relating to non-trivial, commercial software projects; however, we note that none of the primary studies made any explicit reference to population so the above definition is inferred – a point we return to in Section 4
- *Intervention*: effort estimation modelling – using global data
- *Comparison Intervention*: effort estimation modelling – using local data
- *Outcomes*: more accurate models, reduced bias in effort estimation

We also limited our review to the consideration of experimental designs - specifically empirical analyses - that met certain inclusion criteria:

- data from 5 or more projects per company (since we consider it extremely difficult to construct any meaningful prediction system with fewer than 5 training cases) for at least 2 companies in the global data set
- comparisons of single-organisation models to global models (i.e. not to general cost-estimation models such as COCOMO)
- substantially software projects (i.e. not hardware or co-design)
- commercial projects (i.e. not student projects)
- demonstrably peer reviewed (i.e. more than review of abstracts; exclude Technical Reports, student work)
- published in English, within the last 10 years (1995-2005) because software development practices have changed substantially over time.

A set of search keywords was derived by the reviewers separately examining five published papers that they were previously aware of that addressed the research question. Synonyms, variations in spelling and structure (e.g. if terms could include hyphenation) were also considered and accommodated at this point. One additional search term was added to the candidate list after discussion among the reviewers. This led to the construction of a collection of three generic search strings that when executed together would in principle lead to the retrieval of relevant primary studies:

- *Purpose*: ((cost model*) OR (cost estimat*) OR (cost predict*) OR (estimating cost) OR costimation OR (effort estimat*) OR (effort predict*) OR (estimating effort))
- *Object*: ((software project*) OR (software product*) OR (software development) OR (web project*) OR (web application*) OR (web development))
- *Context*: ((company specific) OR (company external) OR (cross company) OR (individual company) OR (multi company) OR (multi organization*) OR (multi organisation*) OR (within company))

One of the two reviewers was assigned to conduct all of the searching and the other was to verify this through a check of the search outcomes. Both reviewers agreed on an initial selection of sources to be searched, extended by the searcher with agreement from the checker after the scope of sources was considered further. A wide range of search sources was used to give as broad a coverage as possible, given that research on software development and effort estimation had been published across the research domains of business, engineering, computer science, psychology and management science. Full text/content was searched whenever it was available (i.e. in nine of the thirteen searches performed).

Abstracts of all papers retrieved were read by the searcher to determine whether they should be considered as primary studies. If this decision could not be made on the basis of abstract alone the rest of the paper was read, with papers included/discarded according to the criteria stated previously. The second reviewer provided comment on the inclusion or exclusion of a small number of borderline papers.

Aggregation of the evidence presented in the primary studies addressed questions such as: How were the data sets split into model building and testing subsets? What techniques were used to measure model accuracy? What validation approaches were used? Initial analysis was qualitative, focused on these questions along with aspects of data quality (DQ) and diversity (DD). We used one aggregator and one checker to perform this analysis.

3.2. Data extraction

Simple fragments of the above query strings were executed against the search sources in order to pilot test the larger queries (or query, in cases where the three could be concatenated). Some search interfaces were certainly easier to use and allowed for more

flexible querying than others. This testing enabled the searcher to assess the impact of wildcards, query nesting, and variations in spelling and in number (e.g. singular vs. plural variants). Once completed, the fragments were 'grown' through the addition of further terms in order to identify query size limits in the search source. This organic querying also enabled the identification of restrictions caused due to the inclusion of reserved or stopwords in the query strings. When such limits or restrictions were encountered a note was made and the queries reformulated in an attempt to overcome them.

It is worth noting that this led to us using a number of query variants in order to meet the requirements of each search source. Some sources were searched with a single query while others required a 'Search within results' sequence. The word "within" was found to be a stopword in three searches, meaning that 'within company' had to be discarded in those cases. Some sources allowed articles in press to be traversed while others did not. While we do not believe that this variation in queries had a detrimental effect on the data extraction outcomes it does highlight the disparate nature of the sources of literature in the empirical software engineering domain.

The search, which was conducted in August 2005, resulted in the retrieval of 185 potentially relevant papers including duplicates (see Table 1). We excluded studies due to inappropriateness in terms of topic (e.g. the study may have in fact been dealing with risk management but cost estimation was cited, leading to its retrieval), treatment (e.g. data from only one organisation was analysed but the possibility of multi-organisation analysis had been noted), and/or credibility (e.g. the paper may have been unrefereed). Note that approximately 85% of the papers that were excluded were rejected on the grounds of topic, in other words in order to find relevant studies we retrieved a high proportion of irrelevant work. This added a substantial burden to the workload.

Ten relevant primary studies (eight conference papers (C) and two journal papers (J)) were identified:

- C1: Briand, L.C., El Emam, K., Surmann, D., Wieczorek, I., and Maxwell, K. (1999) "An assessment and comparison of common software cost estimation modeling techniques", *Proc. 21st Intl Conf Soft Eng* pp.313-322 (Retrieved from 4 sources)
- C2: Briand, L.C., Langley, T., and Wieczorek, I. (2000) "A replicated assessment and comparison of common software cost modeling techniques", *Proc. 22nd Intl Conf Soft Eng* pp.377-386 (4 sources)
- C3: Jeffery, R., Ruhe, M., and Wieczorek, I. (2001) "Using public domain metrics to estimate software development effort", *Proc. 7th Intl Soft Metrics Symp* pp.16-27 (3 sources)
- C4: Kitchenham, B.A., and Mendes, E. (2004) "A comparison of cross-company and within-company effort estimation models for web applications", *Proc. 8th Intl Conf Empirical Assessment in Soft Eng* pp.47-55 (1 source)
- C5: Lefley, M., and Shepperd, M.J. (2003) "Using genetic programming to improve software effort estimation based on general data sets", *Proc. Genetic and Evolutionary Computation Conf* pp.2477-2487 (3 sources)
- C6: Mendes, E., and Kitchenham, B. (2004) "Further comparison of cross-company and within-company effort estimation models for web applications", *Proc. 10th Intl Soft Metrics Symp* pp.348-357 (4 sources)
- C7: Mendes, E., Mosley, N., and Counsell, S. (2003) "Early web size measures and effort prediction for web costimation", *Proc. 9th Intl Soft Metrics Symp* pp.18-39 (3 sources)
- C8: Wieczorek, I., and Ruhe, M. (2002) "How valuable is company-specific data compared to multi-company data for software cost estimation?", *Proc. 8th Intl Soft Metrics Symp* pp.237-246 (4 sources)
- J1: 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" *Info & Soft Tech* 42(14): 1009-1016 (6 sources)
- J2: Maxwell, K., van Wassenhove, L., and Dutta, S. (1999) "Performance evaluation of general and company specific models in software development effort estimation" *Mgmt Sci* 45(6): 787-803 (6 sources)

Table 1: Distribution of studies across sources

Source	Found	Discarded	Included
ACM Digital Library	15	13 (Topic: 8; Treatment: 3; Credibility: 2)	2: C1, C2
Blackwell-Synergy	5	5 (Topic: 5)	0
Compendex & Inspec	9	0	9: C1, C2, C3, C4, C6, C7, C8, J1, J2
EBSCOhost	3	1 (Topic: 1)	2: J1, J2
Expanded Academic	1	0	1: J2
IEEE Xplore	30	24 (Topic: 20; Treatment: 4)	6: C1, C2, C3, C6, C7, C8
ProQuest	24	22 (Topic: 20; Treatment: 1; Credibility: 1)	2: J1, J2
Scholar.Google	34	28 (Topic: 22; Treatment: 2; Credibility: 4)	6: C1, C2, C6, C8, J1, J2
ScienceDirect	45	44 (Topic: 41; Treatment: 2; Credibility: 1)	1: J1
Springer	11	10 (Topic: 8; Treatment: 2)	1: C5
Wiley Interscience	0	0	0
WoK Proceedings	5	0	5: C3, C5, C6, C7, C8
WoK Web of Science	3	0	3: C5, J1, J2
Totals	185	147 (Topic: 125; Treatment: 14; Credibility: 8)	38 (10 distinct)

3.3. Data aggregation and analysis

The data shown in Table 2 reveals that with the exception of papers C1 and C8 each study used a different data set, or version of a data set (since the Experience, ISBSG and Tukutuku data sets have been growing over time). The data sets also vary considerably in terms of:

- size (both the number of cases and features)
- quality (in terms of extent of missing values)
- types of predictor features available
- types of software project included (in terms of business sector, size, and country of origin)

Depending upon the definition of the population of interest this variety could be seen as positive in terms of sampling or better coverage. On the other hand it suggests that we are considering a quite heterogeneous population. Unfortunately this is not something we can formally analyse as quantitative measures of heterogeneity (see for example Higgins *et al*. [15]) require as input the individual variances for the response variables - the accuracy measures - for each primary study; no study provided this information. There is also considerable variation in the modelling methods employed, although ordinary least squares (OLS) regression is common to all studies.

In addition to high variability in the data sets used we also observe considerable variations in the analysis procedures employed by each study. This reveals itself, for example, in the range of approaches used for holding out data, ranging from the jack knife to more complex *n*-fold designs. Elsewhere it has been shown that results can be highly sensitive to these decisions and with high variance in the response variable (accuracy) the results from too few samples can be misleading [16].

Evaluation of data quality (DQ – see Table 3) comprised subjective assessments of the reported reliability of the data collection and verification procedures, the degree of completeness in the data, and whether an incentive was provided to encourage organisations to submit data. While such an incentive may have a positive impact in growing the size of the data set, it may be offset if the data is of low quality. Where answers could not be determined definitively a question mark is noted. Consideration of data set diversity (DD) accounted for the number of countries and organisations that were 'represented by' project records, the mix of application domains, the extent to which the global data set was dominated by records from a small number of organisations, and the degree to which the data and characteristics of the single organisation matched those in the global set.

We then used a more quantitative approach to analyse the evidence favouring one modeling method over another (Table 4). For each primary study we considered the number of statistical tests that indicated that a local or global model was more accurate ('For Local' or 'For Global'), or where the tests were inconclusive ('Indifferent'). The '#Total comparisons' figure is, in general, the number of tests ('#Tests', equivalent to the number of modeling methods used) multiplied by the number of 'Accuracy measures' employed, then by the number of local organisations under scrutiny ('#Single orgns', normally one). In addition, five of the ten studies undertook '#Further comparisons'. In some cases these were comparisons against benchmark predictions or adjusted models, or applied variations to the accuracy measures employed.

Table 2: Details of model building and validation in each primary study

Code	Data source	Data set sizes	Data sampling/split notes	Predictor variables	Modeling methods	Validation method
C1	Experience	206 total, 119-n multi, 63 single; n=63, 13, 12, 11, 10 and 10 for 6 orgns	Database contained 206 project records. Chose to consider those from orgns with 10 or more in the DB i.e. six orgns, 119 project records. 63 of the 119 were from the target single orgn.	EFPs, Org type, App type, Target hw, 15 prod factors	OLS regression, stepwise ANOVA, CART, analogy, combinations	Hold-out, 6 orgns; Hold-out, 6 cross
C2	European Space Agency	166 total, 60-n/39 multi, 29/25 single; n=29, ?, ?, ? for 4 orgns	Database contained 166 project records. Chose to consider those from orgns with 8 or more in the DB i.e. four orgns, 60 project records. 29 of the 60 were from the target single orgn. Numbers of predictions made were 39 and 25 respectively, due to missing data in holdout samples.	Adj KLOC, Env type, Team size, 7 COCOMO factors	OLS regression, stepwise ANOVA, CART, analogy, combinations	Hold-out, 4 orgns; Hold-out, 3 cross
C3	ISBSG	789 total, 324-n multi, 14/12 single; n=14	Database contained 789 project records. Chose to consider those records that were rated high quality and that addressed resource levels – devmt and support i.e. 324 project records. 14 of the 324 were from the target single orgn. Number of predictions made was 12, due to missing data in holdout sample.	FPs, Org type, Lang type, Domain, Team size, Target hw	OLS regression, stepwise ANOVA, CART, analogy, robust regression	Hold-out, 2 levels; Leave-one-out
C4	Tukutuku	53 total, 53-n multi, 13 single; n=13	Database contained 53 project records. 13 of the 53 were from the target single orgn.	Team size, Team exp, 11 counts of pages, functions, images, animations	OLS regression	Hold-out, 1 cross; Leave-one-out
C5	Experience	407 total, 149 multi, 63 single; n=48	Database contained 407 project records. Chose to consider those from projects completed before 15 Oct 1991 (with an additional 15 to be completed by a single orgn) i.e. 149 project records. 48 of the 149 were from the target single orgn. Note: all 149 used in training for multi-orgn test.	83 features	OLS regression, analogy, ANN, GP	Hold-out, 1 cross; Hold-out, 1 cross
C6	Tukutuku	67 total, 67/67-n multi, 14 single; n=14	Database contained 67 project records. 14 of the 67 were from the target single orgn. Note: all 67 used in first round of training for multi-orgn test.	Team size, Team exp, Num langs, 8 counts of pages, functions, images	OLS regression, analogy	Leave-one-out/ Hold-out, 1 cross; Leave-one-out
C7	Tukutuku	36 total, 36-n multi, 12 single; n=12	Database contained 36 project records. 12 of the 36 were from the target single orgn. Note: testing performed separately i.e. multi against multi, local against local.	24 counts of pages, functions, images, animations	OLS regression, analogy	Hold-out, 20-cross; hold-out, 20 cross
C8	Experience	206 total, 206-n multi, n single; n=63, 13, 12, 11, 10 and 10 for 6 orgns	Database contained 206 project records. Concentrated on those from orgns with 10 or more in the DB i.e. six orgns. 63, 13, 12, 11, 10 and 10 of the 206 were from the target single orgns to compare against.	EFPs, Org type, App type, Target hw, 15 prod factors	OLS regression, stepwise ANOVA, analogy	Hold-out, 6 orgns; Hold-out, 1 cross/Leave-one-out
J1	ISBSG + Megatec	451 repos, 145 external, 19 single	Database contained 451 project records. Chose to consider those that 'matched' the target orgn's data and profile i.e. 145 records. Single orgn (external to repository) had 19 project records.	FPs, Dev type, Lang type, Target hw, Team size, PDR	OLS regression, analogy	"Hold-out", 1 cross; Leave-one-out
J2	European Space Agency	108 total, 108-n multi, 29/4,6 single; n=29	Database contained 108 project records. 29 of the 108 were from the target single orgn. Numbers of predictions made were 4 and 6 respectively, due to missing data in holdout sample.	KLOC, Country, Company, Start year, Lang type, Env type, Team size, 7 COCOMO factors	OLS regression	Hold-out, 1 cross; Hold-out, 1 cross

Table 3: Data quality and diversity in each of the primary studies

Code	DQ: collection, verification	DQ: completeness	DQ: incentive	DD: country	DD: organisations	DD: domain	DD: organisation dominance	Degree of match single to multi
C1	High	?	Yes (financial, analytical)	None (Finland)	26	Low (Business)	High (63 of 206 projects from one orgn)	High
C2	High	Medium (90 of 166 projects data complete)	Yes (analytical)	Medium (10 Europe)	69	Medium (Aerospace, military, industrial, business)	Medium (29 of 166 projects from one orgn)	Medium
C3	Medium	Low (Only used variables with > 60% complete data)	Yes (analytical)	Medium-High (Up to 20 countries, 6 major)	?	Low (Business)	?	Low-Medium
C4	Low	? (Some variables excluded because data missing)	Yes (analytical)	Medium (8 countries)	24	? (Web hypermedia/ software in "mixed" domains)	Medium (13 of 53 projects from one orgn)	?
C5	Assumed High	? (Some variables excluded because data missing)	Assumed Yes (analytical)	None (Finland)	?	Assumed Low (Business)	Medium (63 of 407 projects from one orgn)	?
C6	Low	? (Several variables excluded because data missing)	Assumed Yes (analytical)	Assumed Medium	Assumed 25	? (Web hypermedia/ software in "mixed" domains)	Medium-High (27 of 67 projects from two orgns)	Low-Medium
C7	Low	Medium (Two projects with data missing, imputed)	Yes (analytical)	Low (5 countries)	17	? (Web hypermedia/ software in "mixed" domains)	Medium-High (12 of 36 projects from two orgns)	Low-Medium
C8	High	?	Yes (financial, analytical)	None (Finland)	26	Low (Business)	High (63 of 206 projects from one orgn, further 56 from 5 orgns)	Varies
J1	Medium/High	Low (Data for many of 38 vars not complete)/?	Yes (analytical)	Medium (14 countries, few major)	?	Low (Business)/Low (Business)	?	Low-Medium
J2	High	Assumed Medium	Yes (analytical)	Low (8 Europe)	37	Medium (Aerospace, military, industrial, business)	Medium (29 of 108 projects from one orgn)	Low-Medium

Table 4: Analysis of statistical evidence

Code	Accuracy measures	#Tests	#Single orgns	#Further comparisons	#Total comparisons	For Local	For Global	Indifferent	Significant?	Authors favour?	Review outcome?
C1	MMRE, MdMRE, pred(.25)	8	1	0	24	13	11	0	No	Global	Global
C2	MMRE, MdMRE, pred(.25)	8	1	0	8 (only MdMRE results reported)	6	2	0	No	Global	Global
C3	MMRE, MdMRE, pred(.25), Rsq	7	1	21	49	39	9	1	Yes	Local	Local
C4	MMRE, MdMRE, pred(.25), MAD, MdAD	1	1	5	10	10	0	0	Yes	Local	Local
C5	MMRE, BMMRE, pred(.25), rho, AMSE, worst case error	5	1	0	30	12	15	3	?	Local	Inconclusive
C6	MMRE, MdMRE, pred(.25), MAD, MdAD	2	1	10	20	15	5	0	Yes	Local	Local
C7	MMRE, pred(.25)	4	1	1	9	9	0	0	?Yes	Local	Local
C8	MMRE, MdMRE, pred(.25)	3	6	0	54	25	27	2	?No	Global	Global
J1	MMRE, MdMRE, pred(.25)	5	1	0	15	14	0	1	Yes	Local	Local
J2	MMRE, pred(.25), rho	1	1	3	6	4	2	0	?	Local	Inconclusive

On this basis three studies (C1,2,8) were interpreted as favouring global models. In passing it is gratifying to note that C1 and C8 were consistent since they used the same data set, although their approaches differed. Five studies (C3,4,6,7, J1) were interpreted as favouring local models. Two studies (C5, J2) were interpreted as inconclusive due to the absence of significance testing.

Limitations were self-identified in some primary studies, particularly in relation to data quality, model construction and experimental design. In several cases the study authors themselves expressed reservations about the outcomes and applicability of their work. Some studies acknowledged that they were effectively pointing out which approach was "less bad" than the other. Overall there was a lack of strong evidence in the primary studies - individually and collectively - and no feasibility of meta-analysis, not least because different response variables were employed.

4. Reflections on the review

Many questions arose during the review, particularly during the data extraction phase (but with implications for the review protocol). One observation is that there is no definitive collection of literature sources that should be considered in conducting a review. Prior reviews and meta-analyses have considered a (different) range of sources [1-3].

As noted in [17] we need to improve how we write our papers, adopting a consistent form. Structured abstracts, meaningful titles and keyword schemes could also be valuable [18]. This is evident from the fact that 85% of the retrieved papers that were excluded were rejected on the grounds of topic. In spite of us using what we believed to be a concise and focused query many irrelevant papers were retrieved. If more meaningful titles, keywords and the like are used then the precision of the searching process could be much improved. The use of standardised response variables would allow meta-analysis to be performed, potentially significant in determining whether there are any underlying patterns.

How much should we rely on searches of (full text) databases to identify our primary studies? The search engines had limitations, and certainly some were easier to use than others. Searches had to be adapted for each portal, meaning that pilot testing of searches, with consequent refinement of the protocol (including the research question), proved to be important.

Also, should we include coverage of unpublished work - so-called 'grey' literature? In our case we did not consider such work. We believe, however, that this did not mean that we were biased to positive results because we assume that for the research question at hand there was no good/right or bad/wrong answer. On the positive side, restricting ourselves to peer-reviewed studies should have in principle ensured that we considered only work of high quality.

How far should we go back? What should be the duration for a review? Such questions are complicated not only by the publication date of a study but also the age of the data considered within the various data sets and the fact that the latter is not always known.

In comparing our actions to those recommended in the guidelines contained in [17] for reviews it is clear that not all steps were followed. In particular we made some decisions based on prior knowledge that limited the scope of the review based on the fact that this was a relatively small-scale exercise, undertaken by two reviewers and expected to reveal fewer than twenty relevant studies:

- we did not have the protocol reviewed by a panel of experts (presumably other researchers in the field) – some verification of the protocol did occur, however, through early exchanges with the other review team.
- we developed the search strategy on our own, without the assistance of librarians – while we felt confident in our ability to develop a strategy that would lead us to uncover all relevant works we cannot say categorically that this was the case.
- we did not pilot the entire review process – in fact in some respects the review as a whole was something of a pilot, oriented as much to learning lessons about reviews as identifying an answer to the research question. As we expected to uncover only a small number of primary studies a full pilot seemed unwarranted. We did pilot the search activity and refined the protocol and our searching as a result.
- we did not discuss the composition of the set of studies discarded/included with an expert panel, or approach the authors of the original studies to identify overlooked work – we remain uncertain of the impact of this decision.

The lack of strong evidence gained from the review and the questions and comments above could lead us to question whether we are being somewhat premature in conducting systematic reviews. Is the body of literature sufficiently mature and of sufficiently high quality to support reviews and meta-analyses? We certainly encountered some difficulties in comparing and combining the studies due to methodological and reporting differences. Most obviously, no meta-analysis is possible due to use of many different response variables. In this respect a simple and useful recommendation is that researchers report residuals, in addition to any other accuracy statistics that they may choose to employ. We also need more research into the impact of different validation procedures. For

example, how much difference does it make to use a jackknife compared to say an *n*-fold procedure? How should a value be selected for *n*? Until such time as this is better understood it would be better that researchers limit themselves to a restricted range of procedures to better enable comparability.

Secondly, are the studies too diverse or heterogeneous to meaningfully combine? We have chosen a research question that has received quite extensive attention from very capable researchers. That said, we note that this research question may be problematic. Large organizations may have multiple divisions each with very different practices, but their projects would be reported as coming from a single organization, confounding the review treatment. Furthermore, as we noted in the protocol there has been no explicit consideration of what population each primary study is addressing. The result is considerable variation in terms of project size, application, development method(s) and infrastructure. It may be that the primary studies should be partitioned to reduce heterogeneity.

On balance, however, we believe that the positives of a review outweigh the negatives. If we wish to advance our empirical efforts then, as here, we can learn valuable lessons through systematic reviews. This should inform the conduct and reporting of subsequent studies, so that it is easier to undertake quality reviews and more definitive outcomes can be achieved.

5. Conclusions

While there was found to be a tendency for the more recent (and perhaps higher quality?) primary studies to support local models it would be inappropriate to state at this stage that the evidence is converging on that outcome. Moreover, we encountered several challenges in combining and interpreting results. These conclusions point to the need for not only more primary studies (addressing appropriate research questions) but also higher quality primary studies conducted using agreed standards and with discipline-wide reporting protocols.

6. Acknowledgements

This work was funded by the UK Engineering and Physical Sciences Research Council - Grant EP/D003504 and by AUT University. We would also like to thank the reviewers of the paper for their constructive suggestions.

7. References

[1] Kitchenham, B., E. Mendes and G.H. Travassos, "A systematic review of cross- vs. within-company cost estimation studies", *Proc. 10th Intl Conf Empirical Assessment in Soft Eng* 2006.

[2] Kitchenham, B.A., E. Mendes and G.H. Travassos, "Cross versus within-company cost estimation studies: a systematic review". *IEEE Trans Soft Eng*, 33(5): 316-329, 2007.

[3] Jørgensen, M. and M. Shepperd, "A systematic review of software development cost estimation studies", *IEEE Trans Soft Eng*, 33(1): 33-53, 2007.

[4] Boehm, B.W., *Software Engineering Economics*. 1981, Englewood Cliffs, N.J.: Prentice-Hall.

[5] Kemerer, C.F., "An empirical validation of software cost estimation models", *Commun ACM*, 30(5): 416-429, 1987.

[6] Gulezian, R., "Reformulating and calibrating COCOMO", *J. Systems Software*, 16: 235-242, 1991.

[7] Kok, P., B.A. Kitchenham and J. Kirakowski, "The MERMAID approach to software cost estimation", in *Esprit Technical Week*, 1990.

[8] Srinivasan, K. and D. Fisher, "Machine learning approaches to estimating development effort", *IEEE Trans Soft Eng*, 21(2): 126-137, 1995.

[9] Shepperd, M.J. and C. Schofield, "Estimating software project effort using analogies", *IEEE Trans Soft Eng*, 23(11): 736-743, 1997.

[10] Experience Pro - Software Technology Transfer Finland http://www.sttf.fi/eng/indexEnglish.htm

[11] International Software Benchmarking Standards Group http://www.isbsg.org

[12] Web Cost Models - Tukutuku Benchmarking Project http://www.cs.auckland.ac.nz/tukutuku/

[13] NASA Metrics Data Program http://mdp.ivv.nasa.gov/

[14] Kitchenham, B., "Procedures for performing systematic reviews", Joint Technical Report: Keele University TR/SE-0401, NICTA 0400011T.1, 2004.

[15] Higgins, J.P.T., S.G. Thompson, J.J. Deeks and A.G. Altman, "Measuring inconsistency in meta-analyses", *British Medical Journal*, 327: 557-560, 2003.

[16] Kirsopp, C. and M. Shepperd, "Making inferences with small numbers of training sets", *IEE Proc - Software*, 149(5): 123-130, 2002.

[17] Kitchenham, B.A. "Systematic reviews", Keynote presentation at the *10th Intl Soft Metrics Symp*, 2004.

[18] Glass, R. L., I. Vessey and V. Ramesh, "Research in software engineering: an analysis of the literature", *J. Sys Soft*, 44: 491-506, 2002