An Analysis of Data Sets Used to Train and Validate Cost Prediction Systems

Carolyn Mair
Bournemouth University
United Kingdom
cmair@bmth.ac.uk

Martin Shepperd
Bournemouth University
United Kingdom
mshepper@bmth.ac.uk

Magne Jørgensen
Simula Labs
Norway
magnej@simula.no

ABSTRACT

OBJECTIVE - to build up a picture of the nature and type of data sets being used to develop and evaluate different software project effort prediction systems. We believe this to be important since there is a growing body of published work that seeks to assess different prediction approaches.
METHOD - we performed an exhaustive search from 1980 onwards from three software engineering journals for research papers that used project data sets to compare cost prediction systems.
RESULTS - this identified a total of 50 papers that used, one or more times, a total of 71 unique project data sets. We observed that some of the better known and easily accessible data sets were used repeatedly making them potentially disproportionately influential. Such data sets also tend to be amongst the oldest with potential problems of obsolescence. We also note that only about 60% of all data sets are in the public domain. Finally, extracting relevant information from research papers has been time consuming due to different styles of presentation and levels of contextural information.
CONCLUSIONS - first, the community needs to consider the quality and appropriateness of the data set being utilised; not all data sets are equal. Second, we need to assess the way results are presented in order to facilitate meta-analysis and whether a standard protocol would be appropriate.

Categories and Subject Descriptors

D.2.9 [**Software Engineering**]: Management—*Cost estimation, Time estimation*

General Terms

Economics, Management, Measurement

Keywords

Data sets, cost prediction, standardisation

PROMISE'05, May 15, 2005, St. Louis, Missouri, USA.

1. INTRODUCTION

The problem of how to generate useful software cost[1] predictions at an early stage in a project has been the subject of a considerable amount of research since the pioneering work of Benington [1] almost 50 years ago. Subsequently researchers such as Kitchenham [8] and Kemerer [5] identified the need for empirical validation of the different, and in many senses competing, prediction systems that were being proposed. This has led to some hundreds of studies that have used different (usually industrially derived), data sets in order to conduct comparative empirical studies of the relative performance of different cost prediction systems. For review articles see [2, 4].

Whilst it is clearly a positive development that cost estimation researchers are active in empirically evaluating prediction systems, this has resulted in a number of new problems. On the whole, results have tended to be inconclusive in the sense that study A using data set B finds prediction system X is to be preferred to prediction system Y, whilst study C using data set D finds the reverse. Potential explanations include use of different evaluation procedures and accuracy indicators [7] which can lead to rank reversal problems. Another, probably more significant area lies in the use of different data sets and their influence upon prediction system performance [11]. This is the motivation for this paper. We wish to investigate the nature and type of data sets being used to develop and evaluate different software project effort prediction systems. This could prove useful for future researchers considering how best to evaluate cost prediction systems. It is also a foundation for meta-analysis when researchers seek to systematically combine results from more than one study.

The remainder of this paper is organised as follows. The next section sets out the method of how we identified the research papers for our analysis. We then present our findings both by data set and by research study. We then conclude by considering the implications of these results for future empirical research studies and for those endeavouring to perform meta-analyses [9].

2. METHOD

In order to perform the analysis of data sets used to train and validate cost prediction systems, we defined the following inclusion criteria:

[1]Strictly speaking we mean effort prediction since the non-labour costs tend to be ignored in this type of research, however, cost is the more commonly used term.

1. the papers were concerned with software cost estimation, and not, for example, size or productivity estimation;
2. the data set(s) were used to evaluate prediction systems (including expert judgement);
3. the data were 'real', not simulated;
4. each dataset comprised at least 2 projects (this excluded case studies).

Given the size of the literature we decided to adopt a sampling procedure. We focused upon journals since one would expect more mature and heavily refereed research studies to be published in such outlets. Results from this search identified three journals as those which had most prolifically published relevant papers according to our criteria over the past 25 years. The selected journals were *IEEE Transactions on Software Engineering* (TSE), *Information & Software Technology* (IST) and the *Journal of Systems & Software* (JSS). All three journals have featured in other software engineering literature reviews, e.g. Glass and Chen [3]. Note that *Empirical Software Engineering* (ESE) was not included since it is not presently included within the Thomson-ISI Scientific Citation Index and is a more recent journal that contains fewer papers that fit our inclusion criteria. We estimate there are about 9 such papers which we hope to analyse in a future, more comprehensive study.

The search was based upon a personal informal bibliographic database[2] coupled with ScienceDirect and IEEE Explore using the search terms 'cost', 'estimation' and 'effort' within the three selected journals.

Details from each paper were catalogued according to information availability within each journal paper. For each paper we identified those data sets that were utilised. And for each data set we collected the following:

- data set name
- version (if any)
- public availability
- contact person (useful for resolving queries concerning the data set)
- start and completion date
- nationality
- number of organisations
- application domain (business sector)
- number of projects
- project type (new or enhancement or mixed)
- number of features
- presence of missing values

Additional information was also collected since this pilot is in fact part of a larger study to conduct a meta-analysis of all empirical cost prediction results, however, this is beyond the scope of this paper. We also note that this exercise was far from straightforward and often involved reference to other papers, analysis of the data directly (when available) and discussions with those responsible for collecting the data.

[2]The database formed part of the Magne Jørgensen's (Simula Labs, Norway) BEST project.

Group	Count	%
Y	42	64.6
N	10	14.1
?	13	20.0
Total	65	100.0

Table 1: Availability of Software Project Cost Data Sets

3. FINDINGS

Next we consider our findings, first in terms of the data sets (some of which are used more than once) and then in terms of research study, many of which use more than one data set, i.e. there is a many to many relationship.

3.1 Data Sets

As indicated our search for empirical studies from the three journals identified a total of 71 distinct data sets, though many of them were used more than once. However, it became apparent that a further 6 data sets were composite in the sense of being generated by concatenating two or more other data sets. We therefore decided to deal with primitive data sets only. This meant that there were 65 data sets for the analysis.

Of these data sets, over 60% are available to researchers (see Table 1). However there was some variability in the exact nature of this availability. For this reason we split Available into three further categories:

- Published where the complete data set is either contained in the paper or has subsequently been de facto published via websites, bundled with software tools or extensively shared (31 data sets).
- Restricted availability where the authors (or other researchers) have consented either explicitly or through practice, to share the data set with other researchers (6 data sets).
- Limited usage where, generally the owners, attach conditions to the usage of the data. Typically, the owner derives commercial advantage through the data, thus researchers may be allowed to use the data but prevented from publishing or sharing the data with others (5 limited use).

In other cases, particularly for older data sets, we were unclear whether the data is available. Overall, something over a quarter of data sets used are not easily available which has clear implications for replication and transparency. It is something of a moot point as to whether studies using confidential data should be published since software development organisations are subject to commercial pressures and we do not wish to hinder the flow of data made available for research. One possibility is, of course, the use of sanitisation procedures though this is at the expense of making the research context less precise and the resultant danger that data is used inappropriately.

These data sets varied in age [3] from 1979 onwards (see Figure 1). Of the 65 data sets, only 21 have exact start

[3]By age we mean the date of the last completed project as opposed to when the research was actually published.

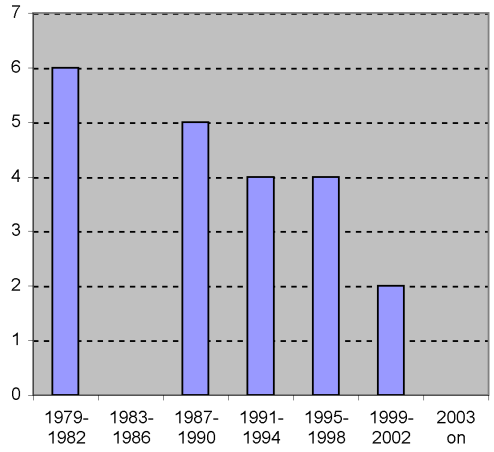

Figure 1: Project Data Sets By Age

Group	Count	%
single organisation	37	56.9
multi-organisation	18	27.7
?	10	15.4

Table 2: Single / Multi-Organisation Data Sets

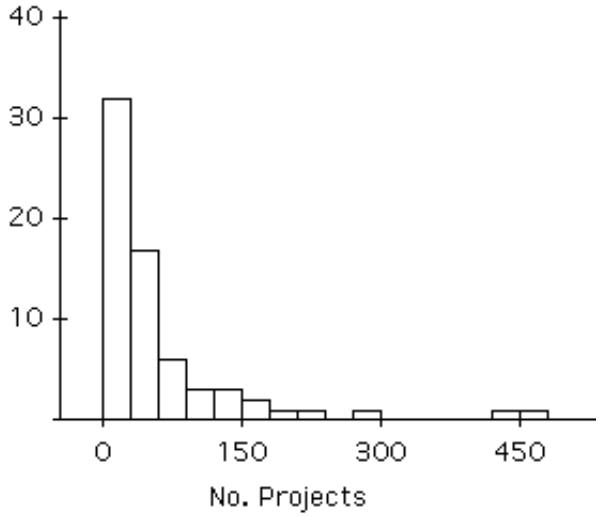

Figure 2: Histogram of Data Set Size (Number of Projects)

and end dates detailed in any study that has used them. Some other studies reported collection dates, often relative to publication. Whilst better than nothing this doesn't provide information on when the projects actually completed (which for some data sets can span a considerable period of time). Of course one can also estimate dates by simply assuming the completion date to be prior to the publication date of the paper in which they were used. However, this does not indicate how long prior to publication date the projects were completed.

It is also instructive to observe that the data sets varied considerably in size (the number of cases or projects - see Figure 2) and the richness of information to describe each project (the number of features or variables- see Figure 3). One suspects that the patterns that might be discovered and the prediction systems evolved for a data set of 3 features differs somewhat from a data set of 40+ features. Both histograms indicate a strong tendency towards smaller data sets. As a community, we need to consider what impact this may have upon our results and recommendations to practitioners.

Another area that has been promoting debate recently concerns the use of single or multi-organisation data. For example some large benchmark data sets such as ISBSG and Finnish contain data from many organisations whereas other data sets contain projects from a single company only. Table 2 indicates that just over half of the data sets comprise projects from a single organisation and a disturbing 15% of all data sets fail to make this information clear at all.

Finally, we look at the country of origin of these data sets (see Table 3). It is clear that Europe and North America dominate, however, it is also striking that for 6 out of 65 of the data sets we are not even provided with this, what might

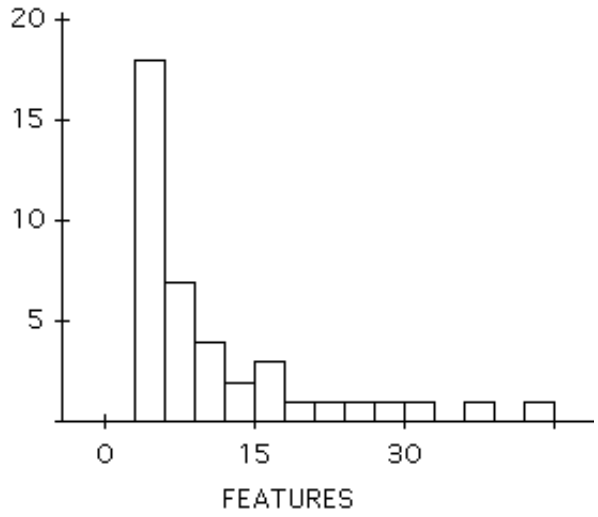

Figure 3: Histogram of Data Set Size (Number of Features)

Country	Count	%
USA	16	24.6
UK	12	18.5
Other European	11	16.9
Australian / NZ	7	10.8
Japanese	4	6.2
Canadian	3	4.6
Multi-national	4	6.2
?	6	9.2

Table 3: Software Project Cost Data Sets by Country of Origin

Journal	Count	Dates
JSS	19	1981 - 2003
TSE	18	1987 - 2004
IST	13	1994 - 2005
Total	50	

Table 5: Research Studies by Journal and Date

be regarded as, quite basic information and for several other data sets the authors had to make "informed guesses".

Finally, we provide some further descriptive information in Table 4 for the 42 data sets which have been classified as Available.

3.2 Research Studies

The systematic search described in the previous section identified a total of 50 papers that used a total of 65 unique project data sets with some data sets being used repeatedly and some in combination.

Table 5 shows the distribution of papers between the three journals identified from 1981 to present and the range of dates covered by papers satisfying our inclusion criteria. The publication trends are shown in Figure 4 and broadly indicate an increase in the number of research papers that use data sets to evaluate cost prediction systems.

Figure 5 shows that the majority of data sets are used only once. This is for two reasons. First our analysis is limited to only three journals so many studies are excluded. Second, and less expectedly is that there are many variants

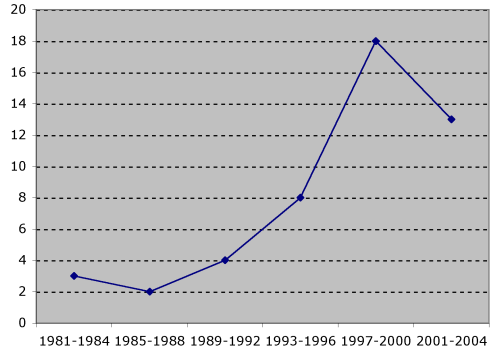

Figure 4: Line Plot of Publications Over Time

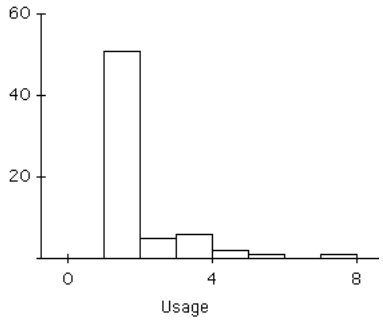

Figure 5: Histogram of Frequency of Data Set Utilisation

and versions of data sets. Examples are the ISBSG and Finnish data sets that grow over time with new versions being released often on an annual basis. Clearly it is important for researchers to be specific about which version they are using. We also observed on occasions that researchers combined two or more existing data sets or removed / added a small number of data points. Moreover there is no unambiguous naming convention so it is possible that use of synonyms has caused additional confusion.

We noted that the most heavily used data sets (COCOMO, Desharnais, Kemerer and Albrecht and Gaffney) are amongst the oldest data sets dating from the 1970s or 80s. In one sense this is to be expected since these data sets have had the most opportunity for use. However, when conducting meta-analyses or other forms of overall analysis we do need to be somewhat cautious about their age in an industry characterised by rapid change.

4. DISCUSSION

In this study of 50 published empirical studies of cost prediction systems from three software engineering journals we have uncovered some interesting characteristics of data sets that are used to train and evaluate software cost prediction systems.

We observed that some of the better known and publicly accessible data sets were used repeatedly making them potentially disproportionately influential. Such data sets also tend to be amongst the oldest with potential problems of obsolescence. We also note that only about 60% of all data sets are in the public domain and this can be particularly problematic when the data set description is incomplete or limited.

Data sets varied considerably in terms of size, number of features, age, nationality, number of organisations, treatment of missing data and so forth. This means we need to be much more systematic in exploring the relation between data set characteristics and prediction system performance. We also need to avoid using data sets that are no longer representative of modern software development practices and current data collection opportunities. Since availability of data sets is clearly a factor we need to consider making some of the more modern and complex data sets widely available. For this reason initiatives such as the PROMISE [10] are very welcome. Having said this, there is the danger that more complex data sets are more easily misunderstood, so detailed protocols and dialogue with those associated with

Dataset name	Study	Published	Projects	Features	Missing?	Organisations	Country
Abran-Robillard	TSE 22 12 895-910	Y	21	31	Y	S	Canada
Albrecht-Gaffney	TSE 9 6 639-648	Y	24	5	N	S	USA
Bailey-Basili	IST 36 5275-282	Y	18				USA
Belady-Lehmann	IST 36 5275-282	Y	33				USA
Shepperd-Cartwright	TSE 27 11 1014-1022	Y	39	3	N	M	UK
BT software houses	JSS 5 4 267-278	Y	12	4		S	UK
BT systemX	JSS 5 4 267-278	Y	11	4		S	UK
COCOMO	JSS 16 235-242	Y	63	42		M	USA
CSC	JSS 64 1 57-77	Y	145	11	Y	S	USA
Desharnais	TSE 23 12 736-743	Y	81	9	Y	S	Canada
Dolado	JSS 37 161-173	Y	24	4	N	S	Rest of Europe
Finnish (Mermaid)	TSE 23 11 736-743	Y	38	29		M	Rest of Europe
Hastings-Sajeev	TSE 27 4 337-350	Y	8	7	N	M	Australia / NZ
Heiat-Heiat	JSS 39 1 7-14	Y	35	4	Y	M	USA
ICL	JSS 5 4 267-278	Y	10	4		M	UK
Jørgensen97	JSS 68 3 253-262	Y	20	4	N		Rest of Europe
Jørgensen04-X	TSE 30 4 209-217	Y	47	4	N	S	Rest of Europe
Jørgensen04-Y	TSE 30 4 209-217	Y	23	4	N	S	Rest of Europe
Kemerer	IST 44 15 911-922	Y	15	3	N	S	USA
MERMAID 1	TSE 23 12 736-743	Y	30	18	Y		Rest of Europe
MERMAID 2	IST 44 13-24	Y	28	17	Y		UK
Misic-Tesic	JSS 41 2 133-143	Y	7	16	N	S	Rest of Europe
Miyazaki et al.1	JSS 27 3-16	Y	48	8	N	M	Japan
Miyazaki et al.2	JSS 27 3-16	Y	10	4	N	S	Japan
Miyazaki et al.3	JSS 27 3-16	Y	11	3	N	S	Japan
Miyazaki et al.5	JSS 27 3-16	Y	34	5	N	S	Japan
Moser-etal	JSS 49 33-42	Y	37	4	N	M	Switzerland
Telecom 1	TSE 23 11 736-743	Y	18	5	N	S	UK
Wingfield	IST 36 5 275-282	Y	15			S	USA
WSD1	JSS 2 2 97-103	Y	33	8	N	S	USA
WSD2	JSS 2 2 97-103	Y	30	8	N	S	Multi
Dolado-academic	IST 43 61-72	R	48	4		M	Rest of Europe
Jeffery-Stathis	IST 42 14 1009-1016	R	19		Y	M	Australia / NZ
Jørgensen95	TSE 21 8 674-681	R	109	11		S	Rest of Europe
MacDonell-Shepperd	JSS 66 2 91-98	R	77	26	N	S	Australia / NZ
Uni of Otago	IST 45 389-404	R	70		N	S	Australia / NZ
Yourdon	IST 36 5 275-282	R	17				USA
COCOMO II.1998	TSE 25 4 573-583	L	161	22		M	USA
ASMA R5	IST 39 7 469-476	L	136	7	Y	M	Australia / NZ
Experience	TSE 27 10 890-908	L	206	17	N	M	Rest of Europe
ISBSG R5	IST 42 14 1009-1016	L	451	38	Y	M	Multi
ISBSG 421	IST 42 649660	L	421	7	Y	M	Multi

Table 4: Available Data Sets Description. Y = published, R = restricted availability and L = limited use permitted. S = single organisation and M = multiple organisations.

collection are essential.

A possible threat to our findings is the question of how representative are the studies that we have identified? Clearly it would be useful to continue this work in order to construct a more complete picture. Nonetheless we believe we have examined a considerable number of studies over a period of almost 25 years from three international, refereed and archival journals.

In addition, the process of extracting relevant information from research papers has been time consuming due to different styles of presentation and levels of contextural information. Again, we consider initiatives such as the PROMISE [10] helpful. In addition there have been occasions where we have had to make subjective judgements, for example about whether the removal of an outlier constitutes the formation of a new data set and so forth. There are also a number of blanks and unknowns within Table 4. Given the scale of the task it is also possible that there are errors and misunderstandings on our part (for which we of course apologise). Consequently we the authors would be grateful for any guidance or corrections on the part of those whose work we have attempted to analyse.

Overall we feel our pilot analysis highlights the need to give very careful consideration to three issues. The data sets we use are extremely varied so we need to consider which data sets we use for training and validation, for instance is it appropriate to use an old data set or study mixed (new and enhancement) project types? Second, given this variation, context is important so when publishing data sets it is essential to provide enough contextual information to support meaningful generalisation. Lastly, meta-analyses and systematic reviews [6] will be greatly facilitated by the use of standard protocols.

Acknowledgments

This work was funded by the UK Engineering and Physical Sciences Research Council under grant GR/S45119.

5. REFERENCES

[1] H. Benington, "Production of large computer programs," presented at *Symp. on Advanced Computer Programs for Digital Computers*, Washington, D.C., 1956.

[2] L. Briand, and I. Wieczorek, "Resource Modeling in Software Engineering," in *Encyclopedia of Software Engineering*, J.J. Marciniak, Ed., 2nd ed. New York: John Wiley, 2002.

[3] R. Glass, and T.Y. Chen, "An assessment of systems and software engineering scholars and institution (19992003)," *Journal of Systems & Software*, 76(1), pp91-97, 2005.

[4] M. Jørgensen, "A review of studies on expert estimation of software development effort," *Journal of Systems & Software*, 70(1-2), pp37-60, 2004.

[5] C. Kemerer, "An empirical validation of software cost estimation models," *Communications of the ACM*, 30, pp416-429, 1987.

[6] B. Kitchenham, "Procedures for performing systematic reviews," Keele University, UK, Technical Report TR/SE-0401 - ISSN:1353-7776, July 2004.

[7] B. Kitchenham, S. MacDonell, L. Pickard, and M. Shepperd, "What accuracy statistics really measure," *IEE Proceedings - Software Engineering*, 48, pp81-85, 2001.

[8] B. Kitchenham and N. Taylor, "Software project development cost estimation", *Journal of Systems & Software*, 5(4) pp267-278, 1985.

[9] J. Miller, "Replicating software engineering experiments: a poisoned chalice or the Holy Grail," *Information & Software Technology*, vol. 47, pp. 233-244, 2005.

[10] J. Sayyad Shirabad, and T.J. Menzies, "The PROMISE Repository of Software Engineering Databases". School of Information Technology and Engineering, University of Ottawa, Canada. Available: http://promise.site.uottawa.ca/SERepository [Last accessed 21 February, 2005].

[11] M. Shepperd, and G. Kadoda, "Comparing Software Prediction Techniques Using Simulation," *IEEE Trans. on Softw. Eng.*, 27(11), pp987-998, 2001.