Cross-company and Single-company Effort Models Using the ISBSG Database: a Further Replicated Study

Chris Lokan
School of IT&EE
UNSW@ADFA
Canberra ACT 2600, Australia
+61 2 6268 8060
c.lokan@adfa.edu.au

Emilia Mendes
Computer Science Department
University of Auckland
Private Bag 92019, Auckland, New Zealand
+64 9 373 7599 86137
emilia@cs.auckland.ac.nz

ABSTRACT
Five years ago the ISBSG database was used by Jeffery et al. [6] (S1) to compare the effort prediction accuracy between cross- and single-company effort models. Given that more than 2,000 projects were later volunteered to this database, in 2005 Mendes et al. [17] (S2) replicated S1 but obtained different results. The difference in results between both studies could have resulted from legitimate differences in data set patterns but also could have been influenced by differences in experimental procedure. S2 was unable to employ exactly the same experimental procedure used in S1, as S1's procedure was not fully documented. Therefore this paper aimed to apply S2's experimental procedure to the ISBSG database version used in S1 (release 6) to assess if differences in experimental procedure would have contributed towards different results. Our results corroborated those from S1: we found that predictions based on a single-company model were significantly more accurate than those based on a cross-company model.

Categories and Subject Descriptors
D.2.8 [**Software Engineering**]: Metrics – *process measures.*

General Terms
Management, Measurement, Experimentation.

Keywords
Effort estimation, software projects, cross-company estimation models, single-company estimation model, regression-based estimation models, replication study, experimental procedure.

1. INTRODUCTION
Previous studies have suggested that single-company data sets are needed to produce accurate effort estimates (e.g. [10],[7]). However, three main problems can occur when relying on single-company data [2]:

- The time required to accumulate enough data on past projects from a single company may be prohibitive.
- By the time the dataset is large, technologies used by the company may have changed, and older projects may no longer be representative of current practices.
- Care is necessary, as data needs to be collected in a consistent manner.

These three problems have motivated the use of cross-company data sets (datasets containing data from several companies) for effort estimation and productivity benchmarking. However, the use of cross-company data sets also has problems of its own [2], [16]:

- Care is necessary, as data needs to be collected in a consistent manner.
- Differences in processes and practices may result in trends that may differ significantly across companies.
- There is a need to guarantee uniform data collection control across different companies, compared to data collection within a single company.
- It may be necessary to partition projects (e.g. according to their completion dates) in order to identify those that used current development practices from those that did not.
- Ideally project data should represent a random sample representative of a well-defined population. Whenever this is not the case the cross-company effort model may not generalize to other projects, even if the data set is large.

To date ten studies in software engineering have investigated whether cross-company models can be as accurate as single-company models [1],[2],[5],[6],[19],[12],[15],[17], using data from conventional software or Web applications:

- Four studies found that a cross-company model gave similar prediction accuracy to that of a single-company model [1],[2],[19],[17].
- Six studies found that a cross-company model did not give as accurate predictions as a single-company model [5],[6],[12],[15],[8],[16].

A summary of these ten studies is given in Table 1.

ISESE'06, September 21–22, 2006, Rio de Janeiro, Brazil.

Table 1 – Comparison of previous studies

	Study 1 [15]	Study 2 [1]	Study 3 [2]	Study 4 [5]	Study 5 [6]	Study 6 [19]	Study 7 [12]	Study 8 [8]	Study 9 [16]	Study 10 [17]
Database	ESA	Laturi	ESA	ISBSG, Megatec	ISBSG	Laturi	Finnish	Tukutuku	Tukutuku	ISBSG
Application domain(s)	Mainly aerospace, industry, and military	MIS	Mainly aerospace, industry, military	Mixed	Mixed	MIS	IS	Mainly corporate, Information, promotional, e-commerce	Mainly corporate, Information, promotional e-commerce	Mixed
Type of application	Not Web-based	Not Web-based	Not Web-based	Not Web-based	Not Web-based	Not Web-based	Not Web-based	Web-based	Web-based	Not Web-based
Countries	Europe	Europe	Europe	ISBSG: worldwide Megatec: Australia	Worldwide	Europe	Finland	Worldwide	Worldwide	Worldwide
Total Dataset size	108	206	166	164	324	206	164	53	67	872
Single company	29	63	28	19	14	6, each 10+ projects	15	13	14	187
CC showed similar accuracy to SC	No	Yes	Yes	No	No	Yes	No	No	No	Yes

MIS - Management and information systems
IS – Information Systems
CC – Cross-company
SC – Single-company

Mendes et al.'s [17] replication of Jeffery et al.'s study [6] was unable to apply exactly the same experimental procedure used in [6], for several reasons:

- Changes in measurement rules associated with some database variables;
- Use of a very large single-company data set where the calculation of prediction accuracy based on a leave-one-out cross-validation would be too time consuming,
- Incomplete analysis process description provided in [6]. For example, it was not clear:
  - What variables were used for each estimation model.
  - If accuracy statistics were obtained based on the raw scale data.
  - If resulting estimation models were subject to sensitivity or residual analysis.
  - If the cost model was recalculated from scratch for each cross-validation exercise or a common model was re-calibrated.
  - What function points methods were considered.
  - Exactly what quality rating was used to select projects.
  - The detailed criteria used to merge some variables (e.g. organization type).

These constraints and the contradictory results obtained in [17] led to the following research questions:

- Question 1: Using the data available to [6], and the experimental procedure of [17], how successful is a cross-company model at estimating effort for projects from a single company, where the single-company projects were not used for model building?
- Question 2: Using the data available to [6] and the experimental procedure of [17], how successful is a cross-company model, compared to a single-company model?
- Question 3: If we were to apply the same experimental procedure described in [17] to the ISBSG database used in [6], would we obtain similar results to those found in [6]?

All models presented in this paper were built using forward stepwise regression with the statistical language R and SPSS v12.1. All remaining analyses were carried out using SPSS v12.1. Statistical significance was set at 0.05.

Although [6] employed numerous estimation techniques we chose to employ a single technique to build the effort models, since it is not our aim to also compare the estimation accuracy between different techniques. The technique we chose was stepwise regression since it is the single technique employed in all previous studies, and either provided the best accuracy or was amongst the best.

Prediction accuracy was measured using MMRE, Pred(25), and Median MRE.

The remainder of the paper is organized as follows: Section 2 describes the research method employed, and its results are presented in Section 3. Section 4 discusses differences and similarities between [6], [17], and this study, and suggests reasons for the different results found in each study. Section 5 investigates the use of different experimental procedures. Finally, conclusions and comments on future work are given in Section 6.

2. RESEARCH METHOD

2.1 Data set Description

The data set used in our investigation represents software projects from the ISBSG database Release 6 (Release 6), as this was the same database and release used in [6].

Release 6 had data on 789 projects (14 from a single company and 775 from different companies). The original data set was reduced to comply with the criteria used in [17], as follows:

- Remove projects if their size was not measured using IFPUG version 4.
- Remove projects whose normalized effort differs from recorded effort. This should mean that the reported effort is the actual effort across the whole life cycle.
- Remove projects if they were not assigned a high data quality rating (A or B) by ISBSG.
- Remove projects with resource levels different from 1 (development team effort only).

This left 9 single-company projects (the same single-company analyzed in [6]) and 89 cross-company projects. Both data sets were much smaller than those used in [6] (14 single-company projects and 310 cross-company projects); however, our cross-company data set size was still adequate for statistical analysis and larger than other cross-company data sets used in previous studies. The problem was in relation to our single-company data set as its size was much smaller than any other single-company data sets used in previous studies and its use for statistical analysis could be problematic. Therefore we decided to consider, only for the single-company data set, projects where size was also measured using IFPUG version 3 and where normalized effort differed from recorded effort by no more than 11%. After applying these criteria our single-company data set increased to 12 projects. We applied the independent samples t-test and the Mann-Whitney U test to check if the distributions of size and effort would differ significantly between IFPUG version 3 and version 4 projects. No significant differences were found.

The set of variables was selected using similar exclusion criteria to those employed in [17]:

- Variables that had more than 40% of their values missing were excluded.
- Variables that contained estimated values (eg normalized effort), rather than actual values, were excluded.
- Variables that contained redundant information were excluded, e.g. size in lines of code, since size in function points is already included.

After applying the exclusion criteria used in [17] the original set of 21 variables was reduced to eight: *effort* (dependent variable), *size* (unadjusted function points), *organization type*, *business area type*, *application type*, *development type*, *platform*, and *language type*. In addition to using these exclusion criteria, we also removed organization type, business area type, and application type from the data sets because they had too many levels and would have required far too many dummy variables, which would rapidly reduce the degrees of freedom for analysis.

Table 2 presents the variables used in this study.

Table 2 – Variables used in this study

Variable	Scale	Description
Effort	Ratio	Normalized project effort in person hours
Ufp	Ratio	Application size in unadjusted function points
LangType	Nominal	Language type (e.g. 3GL, 4GL)
DevType	Nominal	Describes whether the development was a new development, enhancement or re-development
Platform	Nominal	Development platform (mainframe, midrange, PC)

Summary statistics for the ratio-scale and nominal variables are presented in Tables 3 and 4, respectively. The project delivery rate ("PDR", calculated as Effort/Ufp) is also included to provide an additional way to compare cross- to single-company projects. This measure is often used to measure productivity, where high values indicate low productivity.

Table 3 shows clear differences between single- and cross-company projects regarding their size, effort and project delivery rate. Median size is fairly similar; otherwise the cross-company projects tend to be larger, require more effort, have higher (worse) PDR, and have greater variance than single-company projects. Similar trends were observed in [6].

Table 3 –Characteristics for the ratio-scaled variables

Single-company data – 12 projects					
Variable	Mean	Median	St. Dev.	Min.	Max.
Ufp	289	285	118	120	517
Effort	679	714	295	212	1,238
PDR	2.4	2.1	0.9	1.3	4.1
Cross-company data – 89 projects					
Ufp	489	203	1465	10	13580
Effort	4310	1624	10730	140	78472
PDR	10.3	8.5	7.4	0.7	39.3

Table 4 – Characteristics for the nominal variables

Single company – 12 projects			
Category	Levels	Mean Effort	#Projs
LangType	4GL	679	12
DevType	Enhancement	323	2
	New development	751	10
Platform	Midrange	893	1
	PC	741	11
Cross-company – 89 projects			
LangType	3GL	6369	38
	4GL	2243	29
	Application generator	2152	2
DevType	Enhancement	1262	32
	New development	3775	52
	Re-development	29387	5
Platform	Mainframe	5102	68
	Midrange	2535	10
	PC	1318	7

Table 4 summarizes the mean effort and number of projects for categorical variables. Overall the cross-company data set presents more levels per categorical variable than the single-company data set, which is of no surprise. As in [17], redevelopment and enhancement projects were merged in the single-company data set, as there was only one of each and their effort values were similar.

We were unable to do the same for the cross-company data set because of the large difference in average effort between re-development and enhancement projects. The average effort for projects developed on a PC is smaller than on other platforms, in both data sets. In general there are similar trends between both data sets.

2.2 Modelling Techniques

Before building the cross- and single-company effort models using stepwise regression it is important to make sure that assumptions related to using multivariate regression are not violated [14]. For example, skewed numerical variables need to be transformed such that they resemble more closely a normal distribution. Independent variables should present a reasonable relationship with effort (our dependent variable), and variables used in the same model should be independent from each other. The One-Sample Kolmogorov-Smirnov Test (K-S test) was used to check if the two numerical variables, *Ufp* and *Effort*, were normally distributed. In the cross-company data set they are not, so they were both transformed to a natural logarithmic scale to approximate a normal distribution [11]. Once transformed, their distributions were re-checked; the K-S test confirmed that they were both normally distributed. The transformed variables' names are *leffort* and *lUfp*. In the single-company data set, no transformation was necessary.

A scatter plot was used as a visual way of investigating the relationship between *effort* and *ufp* in the single-company data set (see Figure 1), and between *leffort* and *lUfp* in the cross-company data set (see Figure 2).

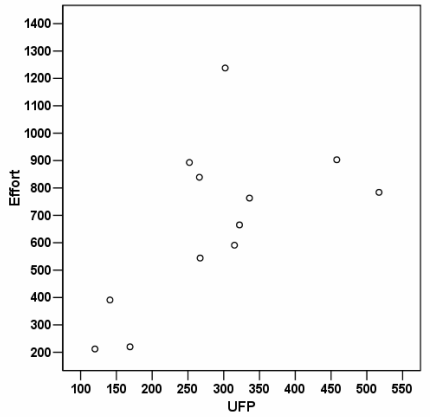

Figure 1 – effort and ufp for single-company data set

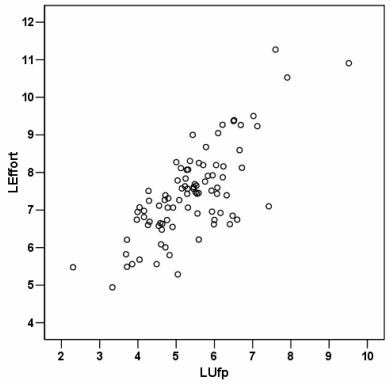

Figure 2 – lnEffort and lnUfp for cross-company data set

In the cross-company data set, the nominal variables *Platform*, *DevType*, and *LangType* had three levels each. Each was replaced by two dummy variables, where each variable was coded 0 and 1. In the single-company data set, *DevType* and *Platform* had two levels each thus each variable was replaced by one dummy variable; since *LangType* had only one level (4GL) it was not included in the stepwise regression. The final set of variables used for each data set is presented in Table 5.

Table 5 – Variables used in the stepwise regression

Variable	Meaning
Single-company	
Effort	Effort in hours
Ufp	Size in unadjusted function points
DevTypeNew	Dummy variable where 'new development' is coded as 1 and 'enhancement' is coded as 0
PlatformPC	Dummy variable where 'PC' platform is coded as 1 and 'midrange' platform is coded as 0
Cross-company	
leffort	Natural logarithm of effort.
lufp	Natural logarithm of ufp.
DevTypeNew	Dummy variable where 'new development' type is coded as 1 and others are coded as 0
DevTypeEnh	Dummy variable where 'enhancement' type is coded as 1 and others are coded as 0
PlatformMF	Dummy variable where 'mainframe' platform is coded as 1 and others are coded as 0
PlatformMR	Dummy variable where 'Midrange' platform is coded as 1 and others are coded as 0
LangType3GL	Dummy variable where '3GL' language type is coded as 1 and others are coded as 0
LangType4GL	Dummy variable where '4GL' language type is coded as 1 and others are coded as 0

2.3 Analysis Methods

To verify the **stability** of each cost model the following steps were used [8]:

- Use of a residual plot showing residuals vs. fitted values to investigate if the residuals are random and normally distributed.
- Calculate Cook's distance values [4] for all projects to identify influential data points. Any projects with distances higher than $3 \times (4/n)$, where n represents the total number of projects, are immediately removed from the data analysis [14]. Those with distances higher than 4/n but smaller than $(3 \times (4/n))$ are removed in order to test the model stability, by observing the effect of their removal on the model. If the model coefficients remain stable and the goodness of fit improves, the influential projects are retained in the data analysis.

The prediction **accuracy** of models was checked by omitting a group of projects and predicting the effort for the group of omitted projects. The rationale was to use different sets of projects to build and to validate a model. Finally the prediction accuracy of each model was always tested on the raw data (not log-transformed data) and the same statistics used in [17] and [6] were employed (e.g. MMRE, Median MRE, and Pred(25).

3. RESULTS

3.1 Cross-Company Data

The best cross-company model, based on the full set of 89 projects, selected two significant independent variables: *lUfp* and *PlatformMF*. Its adjusted R^2 was 0.62. The residual plot for the 89 projects showed several projects that seemed to have very large residuals. This was also confirmed using Cook's distance. Eight projects had their Cook's distance above the cut-off point (4/89); of these two had values greater than 0.1348 (3 times the cut-off value). These two projects were permanently removed from the analysis.

To check the model's stability, a new model was generated without the six projects that presented high Cook's distance, giving an adjusted R^2 of 0.716. In the new model the independent variables remained significant and the coefficients had similar values to those in the previous model. Therefore, the six high influence data points were not permanently removed. The final equation for the cross-company data set, based on 87 projects, is described in Table 6. Its adjusted R^2 was 0.622.

Table 6 – Best cross-company Model to calculate leffort

Independent Variables	Coefficient	Std. Error	t	p>|t|
(constant)	2.655	0.403	6.594	0.00
lufp	0.798	0.071	11.283	0.00
PlatformMF	0.630	0.177	3.559	0.01

When transformed back to the raw data scale, this gives the equation:

$$effort = 14.224\, ufp^{0.798}\, e^{0.63 PlatformMF} \quad (1)$$

The residual plot and the P-P plot for the final model are presented in Figures 3 and 4, respectively. P-P Plots (Probability plots) are normally employed to verify whether the distribution of a variable matches a given distribution, in which case data points gather around a straight line. The distribution that has been checked here is the normal distribution, and Figure 4 suggests that the residuals are normally distributed, and the One-Sample Kolmogorov-Smirnov Test confirmed that they were.

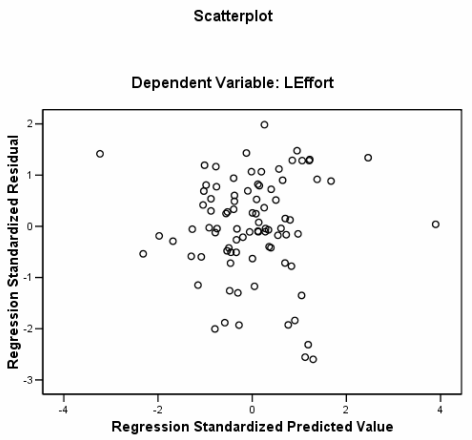

Figure 3 – Residuals for best cross-company model

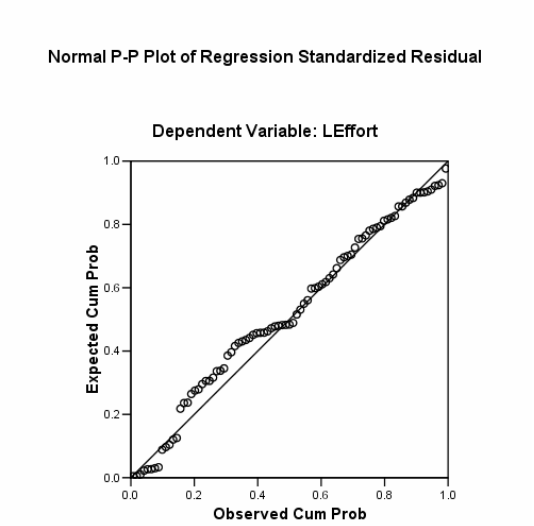

Figure 4 – Normal P-P plot for best cross-company model

3.2 Cross-Company Model applied to Single-Company data

The cross-company model represented by Equation 1 was used to estimate effort for the 12 single-company projects, which were used as a validation set. The prediction accuracy statistics are presented in Table 7.

Table 7 – accuracy statistics for CC model on SC data

Accuracy for our CC model	
MMRE	114%
MdMRE	111%
Pred(25)	8.3%
Accuracy for Jeffery et al.'s CC model	
MMRE	90%
MdMRE	68%
Pred(25)	0%

The accuracy for our model is very low (high mean and median MREs and low Pred(25) values). This is a similar pattern to that observed in [6], although their MMRE and MdMRE were slightly better than ours.

3.3 Single-Company Data

The best single-company model involved only one significant independent variable: *Ufp*. Its adjusted R^2 was 0.312. The model selected none of the dummy variables and only explains 31.2% of the variation in effort, suggesting that there are other contributing variables missing from the final model.

One project had Cook's distance above the cut-off point (4/12). Re-fitting the model without this data point improved adjusted R^2 to 0.42, and the coefficients were broadly similar, so we retained this data point.

The final equation for the single-company data set, based on 12 projects, is described in Table 8.

Table 8 – Best Single-company Model to calculate effort

Independent Variables	Coefficient	Std. Error	t	p>|t|
(constant)	253.4	194.8	1.208	0.255
ufp	1.538	0.629	2.445	0.035

The Equation as read from the final model's output is:

$$effort = 253.4 + 1.538ufp \quad (2)$$

The residual plot and the P-P plot for the final model are presented in Figures 5 and 6, respectively. Figure 6 suggests that the residuals are normally distributed, and the One-Sample Kolmogorov-Smirnov Test confirmed that they were.

To assess the accuracy of the predictions for the single-company model a 20-fold cross-validation was applied to the data set, using the raw scale and a 66% split. This means that 20 times a randomly generated set of 4 projects (34%) was omitted from the data set, and an equation similar to Equation 2 was calculated using the remaining 8 projects (66%). Then the estimated effort was calculated for all the projects that had been omitted from the data set, and statistics such as MRE and absolute residual were also obtained.

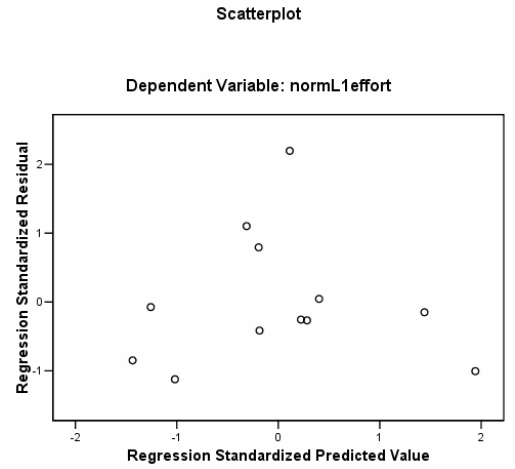

Figure 5 – Residuals for best single-company model

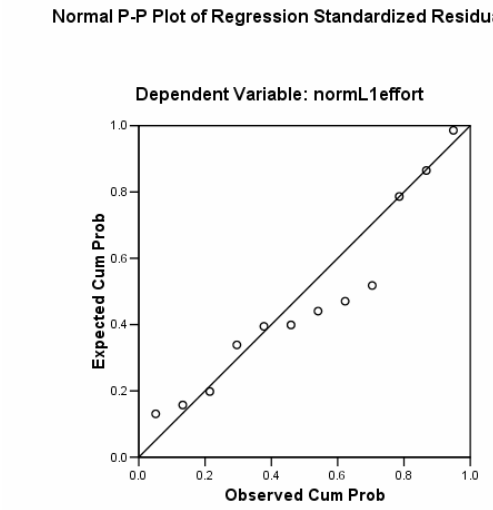

Figure 6 – Normal P-P plot for best single-company model

The prediction accuracy statistics are presented in Table 9. The model's prediction accuracy is still short of the typical target of MMRE=25% and Pred(25)=75%, but it is better than that for the cross-company model.

Table 9 –accuracy statistics for single-company data

Accuracy for our SC model	
MMRE	44.9%
MdMRE	25.2%
Pred(25)	50.0%
Accuracy for Jeffery et al.'s SC model	
MMRE	25.4%
MdMRE	22.8%
Pred(25)	58%

The accuracy for our model is low based on MMRE and Pred(25). The only measure similar to that obtained in [6] was MdMRE. These results seem surprising, given that our single-company data set has almost the same projects as in [6]. The difference comes from some size and effort values being different (see Sections 4.1 and 4.3).

3.4 Answering Research Questions

The first research question (see Section 1) was addressed by the results from Section 3.2. The accuracy of estimates obtained for the 12 single-company projects using the cross-company model (see Equation 1) does not indicate good prediction accuracy. MMRE is 114%, which is poor (25% is considered "good" [3], and Pred(25) is also poor (8.33%, when 75% indicates a good prediction model). Our results corroborate those found in [6].

To address the second research question the absolute residuals from using the 12 single-company projects with the single-company models (see Section 3.3) were compared to those obtained using the same projects with the cross-company model (see Section 3.2). The comparison was done using the Mann-Whitney Test for two independent samples. The results indicated that absolute residuals for the single-company projects using single-company models were significantly different from absolute residuals obtained for the single-company projects using a cross-company model.

The results for the second research question suggest that the single company will obtain better effort estimates using a model based on its own historical data, compared to estimates obtained from a cross-company model. This confirms the results presented in [6].

The answers obtained for our first and second research questions can also be used to answer our third research question. Thus, applying the same experimental procedure described in [17] to the ISBSG database release used in [6] provided similar results to those found in [6].

The results we obtained suggest that, despite differences between experimental procedures and to some extent data sets used in [17] and in [6], these differences were not significantly large to influence the outcome of our analysis. These findings also suggest that the results obtained in [17] most possibly were due to differences in patterns between projects from releases 6 and 9, rather than differences from experimental procedures.

4. DISCUSSION

Table 10 summarizes the accuracy statistics found in [6], [17], and this study. It is apparent that the results seen in [6] are generally the best. When the cross-company model is applied to

the single-company data our results showed worse MdMRE and Pred(25) than [6] and [17]. When the single-company model is applied to the single-company data, [17] presents worse results than [6] or this study.

Table 10 – accuracy statistics from the three studies

Model	Data	Study	R^2	Mean MRE	Md MRE	P25
Cross-company	Single-company	[6]	0.50	0.90	0.68	0.00
		[17]	0.55	1.23	0.61	0.21
		This	0.62	1.14	1.11	0.08
Single-company	Single-company	[6]	0.47	0.25	0.23	0.58
		[17]	0.39	1.02	0.60	0.21
		This	0.31	0.45	0.25	0.50

The following sub-sections detail further differences in data preparation, experimental design, attributes and projects between this study and [6].

4.1 Data Preparation

The filtering criteria used in this study differed from those used in [6] and, for the single-company data set, also differed slightly from the criteria employed in [17]. Differences from [17] are pragmatic, aiming to retain as much single-company data as possible while minimizing the risks introduced from uncertainty in the data. Differences from [6] are more fundamental, aiming generally to support sounder comparisons between cross- and single-company projects:

Project size: Jeffery et al. [6] did not explicitly mention that they did not make any distinction between different types of function points, i.e. IFPUG function points of various definitions, feature points, Mark II function points, full function points, and others were all analyzed together. Different definitions are incompatible, and no reliable multipliers exist for converting between different definitions. The analysis in [17] was restricted to projects where size was measured using IFPUG version 4.0 or later. In this study we considered only IFPUG version 4+ for the cross-company data set, but both IFPUG versions 3 and 4+ for the single-company data set. We believe our choice is justified because in the single-company data set they appear indistinguishable. In addition, combining them allowed us to analyze 12 single-company projects as one group rather than as two groups of only half the size.

Project effort: Some time after the release of the ISBSG version 6, ISBSG introduced the concept of effort "normalization". Some projects record effort for only part of the development life cycle; these should not be compared directly with projects that report effort for other phases or all phases. Normalized effort is ISBSG's estimate of full life cycle effort, for projects that report effort for only part of the life cycle. Normalization improves the comparability of effort values between projects, at the cost of introducing uncertainty in the effort values. If normalization makes only a small difference, the risk is small. At some threshold, which should be determined by the researcher or practitioner according to his or her own needs, the uncertainty introduced by using normalized effort outweighs the benefit of improved comparability, and the project should be removed from the data set.

In [17] and in [6] no normalized effort was employed. In this study we avoided the risk of normalization altogether for cross-company projects, by only considering projects that were not affected by normalization, but for the single-company projects we accepted projects for which normalized effort differed from recorded effort by up to 11%; this was a pragmatic decision that allowed us to retain 12 single-company projects.

Data quality rating: In [6], it was not made clear if only projects with data quality rating A were analyzed, or if projects with quality rating B were also included; recent study of the data showed that it was only data quality rating A. In [17] and here, data quality rating A and B were used. ISBSG considers both A and B suitable for sound analysis, so this expands the data sets (both cross-company and single-company) with small risk.

Effort resource level: Jeffery et al. [6] analyzed 201 projects with effort resource level 1 (development team only). They also analyzed separately 123 projects with effort resource level 2 (support effort added). Because ISBSG has changed the definition of effort resource level 2 since then, we only considered projects with effort resource level 1.

4.2 Modelling and Analysis Methods

One clear difference between [6] and this study is in the use of dummy variables. In addition to dummy variables for Platform and Language type, Jeffery et al. [6] merged levels for categorical variables, and included dummy variables to capture Organization Type and Business area type, each of which had 11 levels. We felt that a dataset of about 100 projects is insufficient for that many independent variables, since at least 10 projects per independent variable is desirable (when stepwise regression is used, 40 projects per independent variable is preferable) [18], so we did not define dummy variables for those two extra attributes. In addition, even if we had been able to use that number of dummy variables, we were unaware of how the categorical levels were merged and thus it would be difficult to repeat the same procedure.

A second difference is in the method of cross-validation. Jeffery et al. [6] used a leave-one-out cross-validation however it was only applied to their single-company data (14 projects). The single-company data set used in [17] was much larger, making a manual leave-one-out cross-validation too time consuming. Further, when using cross-validation the analysis presented in [6] was limited to a maximum of 14 training sets, which according to recent studies, may lead to untrustworthy results [11]. According to [11] ideally 20 sets or more should be deployed, so this was the number of training and validation sets used in [17] (20-fold cross-validation). As this study aims to repeat the experimental procedure of [17], we also used a 20-fold cross-validation.

These differences could have two effects. Most importantly, the different set of independent variables used in [6] could lead to a model with different accuracy. Second, the different cross-validation procedures could lead to differences in the values of the accuracy statistics. The second issue is investigated in Section 5. The first issue is investigated now.

To gain further insight into the impact of different methods, we sought to apply the modelling methods of [6] to our 12 within-company and 89 cross-company projects. As noted in Section 1, [17] was unable to repeat exactly the analysis of [6] because some details of their analysis were not clear. However, due in part to extra project attributes provided to one of the present authors by ISBSG, to repeat most of the analysis was now possible.

We attempted to re-create the dummy variables for Organization type and Business area type that were used in [6]. For Business area type we could do this exactly; for Organization type we could not quite do so, though for most projects it was possible. We added these dummy variables to the stepwise regression procedure, to repeat as best we could the modelling approach of [6]. Only one of these dummy variables (Business area = Engineering) was significant. The Wilcoxon matched pairs test for differences in absolute residuals found no significant differences between the accuracy of the model including Business area type and the simpler models we had found earlier (Section 3).

This suggests three things:

- The simpler models should be preferred.
- Differences between this study, [6], and [17] are not caused by the different experimental procedures used in these studies.
- The results reported in [6] appear better than in this study (see Table 10), but the difference is not reliable.

4.3 Different Attributes

Researchers generally prefer to use unadjusted function points rather than adjusted function points, because the adjustment process has theoretical problems and generally does not improve the accuracy of effort estimates based on function points [13]. Unadjusted function points were used in [17], and therefore in this study, assuming they had also been used in [6] (it was not explicitly documented in [6] if the size they used was measured using adjusted or unadjusted function points). However, study of the data makes it clear that [6] used adjusted function points.

To see whether the different size measure made any difference, we repeated the analysis of Section 3.2 using adjusted function points as the size measure for our single company. We compared the absolute residuals with those from the model using unadjusted function points. The Wilcoxon matched pairs test for differences in absolute residuals found no significant differences. The same was done in the cross-company data set, with the same result.

4.4 Different Projects

This study corroborated the results obtained in [6], where, using ISBSG Release 6, predictions for projects from a single company were significantly better when using an effort model based on that company's own project data. These results contradict those in [17], where different single- and cross-company projects (only those added since Release 6) from the ISBSG database were used. As we have previously seen, differences between the experimental procedures used in [6] and [17] did not impact the results in our study; thus, it seems that the contradictory results obtained between studies [6] and ours, and study [17] are very likely to be due to differences in the data.

Tables 11 and 12 show summary statistics for the single- and cross-company data sets used in this study, and in [6] and [17]. The size and effort of the single-company projects used in this study and in [6] (SC1) are much smaller than those of single-company projects used in [17] (SC2); SC1 projects present a small variation in size and effort, compared to SC2; finally, SC1 projects presented much higher productivity than SC2 projects.

Table 11 – Comparison of single-company data sets

Single-company data – 12 projects (this study and [6]) SC1					
Variable	Mean	Median	St. Dev.	Min.	Max.
Ufp	289	285	118	120	517
Effort	679	714	295	212	1238
PDR	2.4	2.1	0.9	1.3	4.1
Single-company data – 184 projects (study [17]) SC2					
Ufp	588	294	792	16	6294
Effort	4707	2418	6717	140	57687
PDR	12.9	7.3	16.7	0.5	165.9

The size of the cross-company projects used in this study and in [6] (CC1) was larger than the size of the cross-company projects used in [17] (CC2); however their effort and productivity are very similar.

SC1 presented much higher productivity than CC1, which may explain why the single-company model was so much better than the cross-company model at estimating effort for the single-company projects. A different trend is observed between SC2 and CC2, where SC2 presented similar productivity to CC2. Another observation is that SC2 projects were on average larger in size and effort than CC2 projects, and conversely, SC1 projects were on average smaller in size and effort than CC1 projects.

Table 12 – Comparison of cross-company data sets

Cross-company data – 89 projects (this study and [6]) CC1					
Variable	Mean	Median	St. Dev.	Min.	Max.
Ufp	489	203	1465	10	13580
Effort	4310	1624	10730	140	78472
PDR	10.3	8.5	7.4	0.7	39.3
Cross-company data – 672 projects (study [17]) CC2					
Ufp	292	118	809	3	16148
Effort	3710	1249	7415	14	73920
PDR	18.8	9.3	30	0.5	315.6

In addition, SC2 projects covered a range of application types, business types, languages and platforms. In that sense it was broadly similar to the ISBSG database as a whole. In contrast, SC1 projects are noticeably homogeneous. They are similar in size, and all use the same programming language. PDR is fairly stable for them all, and rather better than average for the ISBSG database.

The ability to form tightly focused cross-company comparison data sets depends on having relevant data available. In [17], and to a lesser extent in this study, the large amount of missing data meant that few variables could be included in the models. Low adjusted R^2 values indicate that there are other influential variables not included in the models, which may largely affect the results.

The advantage of using single-company data over cross-company data for the company studied here and in [6] comes most probably from having single-company projects much more productive and similar to each other than to the rest of the database.

5. DIFFERENT EXPERIMENTAL PROCEDURES

In this study, similar to [17], we used as basis for predictions a 20-fold cross-validation for the single-company data set. Results did not seem to differ from those based on a leave-one-out cross-

validation, as in [6]. However, other studies have used different cross-validation combinations:

- 3-fold cross-validation [2], which is effectively a leave-four-out cross-validation.
- 6-fold cross-validation [1], which is effectively a leave-two-out cross-validation.
- Independent hold-out [14],[12].

Therefore in this Section we present the results from employing these different cross-validation combinations, in addition to a leave-three-out cross-validation (same as 4-fold cross-validation), to the single-company data set used in this study. The statistical significance between absolute residuals was compared using the Wilcoxon matched pairs test, with significance level set at 0.05.

3-fold cross-validation: the single-company data set was split into three different training and validation sets, where the training sets had each 8 projects and the validation sets had each 4 projects.

6-fold cross-validation: the single-company data set was split into six different training and validation sets, where the training sets had each 10 projects and the validation sets had each 2 projects.

Independent hold-out sample: both cross- and single-company models use the same validation set, which is a subset of the single-company data set. Our hold-out sample had 4 projects.

All cross-validation combinations, except for the use of an independent hold-out sample, showed that predictions for the single-company projects obtained using a model built from data belonging to that same single company were significantly superior to predictions obtained using a model built from cross-company data (see Table 13). Thus, except for the independent hold-out sample technique, all other combinations corroborate the findings from this study and those from [6].

Table 13 – Different cross-validation combinations

	This study	[6]]	3-fold	4-fold	6-fold	Hold-out
MMRE	44.9%	25.4 %	68%	80.3 %	72.7 %	45.8%
MdMRE	25.2%	22.8 %	46%	45.6 %	44.7 %	46.2%
Pred(25)	50.0%	58%	8.3%	16.7 %	33.3 %	0%

The use of an independent hold-out sample improved the accuracy measures for the cross-company model (see Table 14). This approach is different from all other approaches since the validation set is a totally separate data set from the data sets used to build the single- and cross-company models. Two previous studies used independent hold-out samples [14],[12], however, their results were not based on statistical significance tests and therefore could not be used as evidence of the usefulness of using such technique [9].

Table 14 – Accuracy for Cross-company model

	This study	[6]	Independent hold-out sample
MMRE	114%	90%	73%
MdMRE	111%	68%	38%
Pred(25)	8.33%	0%	25%

Our single-company data set was small and this led to the use of a very small independent hold-out sample. Thus it is important that other larger single-company data sets are also used in order to fully investigate to what extent using an independent hold-out sample provides the most favorable choice.

6. CONCLUSIONS

This study replicated the study described in [6], using the same database and release, however applying a different experimental procedure. We found that the predictions obtained for a single company using a cross-company model were significantly less accurate than those this company would obtain using its own single-company model (though neither the single-company model nor the cross-company model performed well, in terms of MMRE and Pred(25)). This corroborates the findings in [6].

We have described the process used to prepare the data, particularly with regard to making sure that size and effort values are comparable. This is important for any empirical analysis, but particularly when working with a database like that of ISBSG, in which data comes from many sources and may have many different definitions.

We investigated several factors that might help explain why different studies of cross-company and single-company models produce different results. In this instance, the advantage of single-company data comes from the similarity between projects and their high productivity, both widely different from the cross-company projects.

Finally, we looked at other experimental procedures that have been used in previous studies, to find out whether or not they would corroborate our findings. The use of an independent hold-out sample was the only combination that contradicted our results and those in [6].

Future work in this area will concentrate on looking at different single-company data sets and different experimental procedures to investigate if the use of independent hold-out samples provides an optimal solution.

7. ACKNOWLEDGMENTS

We would like to thank the ISBSG group for making Releases 6 and 9 available for our research and all those companies that have volunteered data on their projects.

8. REFERENCES

[1] 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.

[2] 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.

[3] Conte, S. D., Dunsmore, H. E., Shen, V. Y. *Software Engineering Metrics and Models*, Benjamin-Cummins, 1986.

[4] Cook, R.D. Detection of influential observations in linear regression. Technometrics, 19, 1977, pp 15-18.

[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 Metrics'01, London, 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 E. Mendes. A Comparison of Cross-company and Single-company Effort Estimation Models for Web Applications, Proceedings EASE 2004, 2004, pp 47-55.

[9] Kitchenham, B.A., and E. Mendes, and Travassos, G. A Systematic Review of Cross- vs. Within-Company Cost Estimation Studies, EASE'06, (accepted for publication).

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

[11] Kirsopp, C. and Shepperd, M. Making Inferences with Small Numbers of Training Sets, IEE Proceedings Software, 149, pp 123-130, 2002.

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

[13] Lokan, C.J. Function Points. Advances in Computers, M.V. Zelkowitz (ed), Vol 65, pp 298-347, Elsevier, 2005.

[14] Maxwell, K. Applied Statistics for Software Managers. Software Quality Institute Series, Prentice Hall, 2002.

[15] 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, pp 787-803, 1999.

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

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

[18] Tabachnick, B.G. and Fidell, L.S. *Using Multivariate Statistics,* HarperCollins, 1996.

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