A Replicated Comparison of Cross-Company and Within-Company Effort Estimation Models Using the ISBSG Database

Emilia Mendes[*] Chris Lokan[†]
Robert Harrison[‡] Christopher Triggs[**]

[*] University of Auckland, emilia@cs.auckland.ac.nz
[†] University of New South Wales,
[‡] University of Auckland, rhar152@ec.auckland.ac.nz
[**] University of Auckland, triggs@stat.auckland.ac.nz

A Replicated Comparison of Cross-company and Within-company Effort Estimation Models using the ISBSG Database

Emilia Mendes[1]
emilia@cs.auckland.ac.nz

Chris Lokan[2]
cjl@itee.adfa.edu.au

Robert Harrison[3]
rhar152@ec.auckland.ac.nz

Chris Triggs[3]
triggs@stat.auckland.ac.nz

[1]Computer Science Department
[3] Statistics Department
The University of Auckland
Private Bag 92019, Auckland,
[2]School of Information Technology and Electrical Engineering
UNSW@ADFA, Australian Defence Force Academy
Canberra ACT 2600, Australia

Abstract

Four years ago was the last time the ISBSG database was used to compare the effort prediction accuracy between cross-company and within-company cost models. Since then more than 2,000 projects have been volunteered to this database, which may have changed the trends previously observed. This paper therefore replicates a previous study by investigating how successful a cross-company cost model is: i) to estimate effort for projects that belong to a single company and were not used to build the cross-company model; ii) compared to a within-company cost model. Our within-company data set had data on 184 software projects from a single company and our cross-company data set employed data on 672 software projects.

Our results did not corroborate those from the previous study, showing that predictions based on the within-company model were not significantly more accurate than those based on the cross-company model. We analysed the data using forward stepwise regression.

Keywords: effort estimation, software projects, cross-company estimation models, within-company estimation model, regression-based estimation models, replication study.

1. Introduction

Previous studies have suggested that within-company data sets are needed to produce accurate effort estimates (e.g. [11],[8]). However, three main problems can occur when relying on within-company data [2]:

i) the time required to accumulate enough data on past projects from a single company may be prohibitive.
ii) 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.
iii) 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]:

i) care is also necessary as data needs to be collected in a consistent manner.
ii) differences in processes and practices may result in trends that may differ significantly across companies.

Other researchers have also suggested additional difficulties, such as [17]:

- To guarantee uniform data collection control across different companies, compared to data collection within a single company.
- To be able 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.
- To ensure the project data represents a random sample representative of a well-defined population. Whenever this is not the case the cross-company effort model may not generalise to other projects, even if the data set is large.

Nine studies in Software engineering have investigated whether cross company models can be as accurate as within company models [1],[2],[6],[7],[20],[14],[16].

Seven used data from two application domains: ‘business’ and ‘space and military’. Their findings were as follows:

1530-1435/05 $20.00 © 2005 IEEE

- Three studies found that a cross-company model gave similar prediction accuracy to that of a within-company model [1],[2],[20].
- Four studies found that a cross-company model did **not** give as accurate predications as a within-company model [6],[7],[14],[16].

Two further studies have recently investigated the same issues on the effectiveness of cross-company effort models, with data from Web projects [10],[17] obtained from a single database. Both found that a cross-company model did **not** give as accurate predications as a within-company model.

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

Table 1 - Comparison of previous studies

	Study 1 [16]	Study 2 [1]	Study 3 [2]	Study 4 [6]	Study 5 [7]	Study 6 [20]	Study 7 [14]	Study 8 [10]	Study 9 [17]
Database	ESA	Laturi	ESA	ISBSG, Megatec	ISBSG	Laturi	Finnish	Tukutuku	Tukutuku
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
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
Countries	Europe	Europe	Europe	ISBSG: worldwide Megated: Australia	worldwide	Europe	Finland	worldwide	worldwide
Total Dataset size	108	206	166	164	324	206	164	53	67
Single company	29	63	28	19	14	6, each 10+ projects	15	13	14
CC showed similar accuracy to WC	No	Yes	Yes	No	No	Yes	No	No	No

MIS - Management and information systems CC – Cross-company
IS – Information Systems WC – Within-company

To our knowledge, the last published study that used the International Software Benchmarking Standards Group (ISBSG) database to compare the effort prediction accuracy between cross-company and within-company cost models was published four years ago by Jeffery et al. [7]. Since then more than 2,000 software projects have been volunteered to this database, which may have an impact on the results observed previously.

Therefore this paper's contribution is to replicate Jeffery et al.'s work [7], using project data volunteered after the ISBSG Release 6. The research questions addressed are as follows:

i) How successful is a cross-company model at estimating effort for projects from a single company, when the model is built from a data set that does not include that company;
ii) How successful is a cross-company model, compared to a within-company model?

Both issues are addressed using data on 872 software projects. 187 come from a single company, and 685 come from other companies.

All models used in this investigation were built using forward stepwise regression using the statistical language R[1] and SPSS v10.1. All remaining analyses were carried out using SPSS v10.1. Statistical significance was set at 0.05.

We wish to make it clear that we chose to use a single technique to build the effort models, since it is not our aim to also compare different estimation techniques regarding their estimation accuracy. The choice of stepwise regression was motivated by it being the single technique employed in all nine previous studies, which either provided the best accuracy or was amongst the best.

As in [7], prediction accuracy was measured using MMRE, Pred(25), and Median MRE.

The remainder of the paper is organised as follows: Section 2 describes the research method employed in this study. Results are presented in Section 3 and discussed in Section 4. Finally, conclusions and comments on future work are given in Section 5.

[1] R is an open source statistical programming language based on the S and S/Plus programming languages.

2. Research Method

2.1 Data set Description

The analysis presented in this paper was based on software projects from Release 9 of the ISBSG database, where only those projects volunteered after Release 6 were considered.

The purpose of the ISBSG repository is to provide organisations with a broad range of project data from many industries and many business areas. The data can be used for effort estimation, awareness of trends, comparing platforms and languages, and productivity benchmarking[2]. Release 9 had data on 3,024 projects where:

- Projects come from 20 different countries, mainly Japan (28%), the United States (26%), Australia (24%), the Netherlands (7%), and Canada (6%).
- Development types are enhancement projects (57%), new developments (41%), and re-developments (2%).
- The applications are mainly Management Information Systems (18%) and transaction/production systems (40%).
- More than 70 different programming languages are represented. By category: 3GLs (68%), 4GLs (27%), and application generators (5%). Major languages are Cobol/Cobol II (22%), C/C++ (14%), Visual Basic (11%), Java/J2EE (8%), PL/I (6%), Oracle (5%), SQL (5%), Natural (3%), and Access (2%).

In order to make our analysis meaningful we had to remove projects according to the following criteria:

- Remove projects included in Release 6 as we were only interested in projects added to the database since Jeffery et al's study [7].
- As in [7], remove projects if their size was measured in lines of code or if their size was measured in an outdated version of function points (size measured with an older version is not comparable with size measured with IFPUG version 4.0 or later).
- Remove projects whose normalised effort differs from recorded effort. This should mean that the reported effort is the actual effort across the whole life cycle.
- As in [7], 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). Resource level measures the set of people whose time is included in the effort data reported. Jeffery et al. [7], employed resource levels 1 and 2, however since Release 6 ISBSG no longer records resource level 2 in the same way therefore we only included resource level 1.

We analysed 872 projects: 187 from one company and 685 from other companies.

ISBSG's rules about confidentiality mean that we do not know the identity of the single company. However we can note that the 187 projects come from more than one industry/business area, and, as in the overall ISBSG database, transaction processing systems are most common.

The ISBSG database provides data on 88 variables. We reduced the number of variables to 21 that we believed could potentially have an impact on effort. This subset was further reduced based on the same exclusion criteria employed in [7]:

- Variables that had more than 40% of their values missing were excluded.
- Variables that contained estimated values (eg normalised work 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.

Unfortunately, most variables had more than 40% of their values missing, which largely contributed to reducing the set of 21 variables to four. These variables are presented in Table 2. LangType had 26% of its values missing for the cross-company subset, and 5% missing for the within-company subset. We decided to apply an imputation technique called k-Nearest Neighbour (k-NN) due to its simplicity and reported good results [3].

Table 2 Variables used in this study

Variable	Scale	Description
Effort	Ratio	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

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

Table 3 suggests that there are clear differences between the within-company projects and cross-company projects regarding their effort and application size in unadjusted function points (ufps). Size, effort, and PDR for the cross-company projects have slightly greater

[2] www.isbsg.org

variance than for within-company projects. On average within-company projects are about twice the size in ufps, compared to cross-company projects. Effort too is greater for within-company projects, though (for mean values at least) by not as much. Thee mean, Standard deviation, and maximum PDR, for within-company projects, are all lower (better) than for cross-company projects. Both present similar median PDR, however trends suggest that the productivity of within-company projects is better than that for cross-company projects. This is a similar trend to that observed in [7].

Table 3 Project Characteristics for the ratio-scaled variables

Within-company data – 184 projects					
Variables	Mean	Median	Std. Dev.	Min.	Max.
Ufp	587.5	293.5	792	16	6294
Effort	4706.5	2418	6717	140	57687
PDR	12.87	7.26	16.67	0.53	165.93
Cross-company data – 672 projects					
Ufp	292	118	809	3	16148
Effort	3710	1249	7415	14	73920
PDR	18.84	9.26	30	0.49	315.63

Table 4 summarises the mean effort and number of projects for the categorical variables used in this study. Overall the cross-company data set presents more levels per categorical variable than the within-company data set, which is no surprise. On average, for both cross-company and within-company projects, new development projects used higher effort than enhancement projects. The average effort for projects that used a 3gl or 4gl tends to be similar, and this trend is observed on both cross-company and within-company data sets. In general there are similar trends between both data sets.

Table 4 Project Characteristics for the nominal variables

Within company – 184 projects			
Category	Levels	Mean Effort	#projects
LangType	3gl	4585.52	155
	4gl	5353.45	29
DevType	Enhancement	3261.31	84
	New development	4706.55	100
Cross-company – 672 projects			
LangType	3gl	3810.8	487
	4gl	3447	185
DevType	Enhancement	2461.9	522
	New development	8056.4	150

The original cross-company data set (685 projects) had a few using language types 2gl and 5gl, and also very few ‘redevelopment’ projects. When a situation like this arises, one option is to merge levels with a small number of data points to others with larger number of data points. We adopted this approach for ‘redevelopment’ projects, by merging them with ‘enhancement’ projects, reducing the number of DevType levels to two. In relation to LangType merging levels would be meaningless. Since their possible impact on effort is negligible, we removed these five data points, leaving the original cross-company dataset with 680 projects.

2.2 Modelling Techniques

Before building the cross-company and within-company regression models using stepwise regression it is important to make sure that numerical variables are normally distributed, independent variables have a reasonable relationship with effort (our dependent variable), and that variables used in the same model are independent from each other. The One-Sample Kolmogorov-Smirnov Test (K-S test) was used on both data sets to check if the two numerical variables, Ufp and Effort, were randomly distributed. As they were not, they were both transformed to a natural logarithmic scale to approximate a normal distribution [13]. 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.

To investigate the relationship between leffort and the independent variables, three techniques were used: scatter plots and 2-tailed Pearson’s correlation test for numerical variables, and One-Way ANOVA to check the relationship between nominal independent variables and leffort. A scatter plot was used as a visual way of investigating the relationship between leffort and lufp (see Figures 1 and 2). The Pearson’s correlation test confirmed a significant relationship between these two variables on both data sets.

The One-Way ANOVA results (see Table 5) helped to reduce the number of nominal variables to one for the within-company data set and two for the cross-company data set.

Table 5 Results for One-Way ANOVA between nominal variables and leffort

	Variables	Relationship with leffort
Within company	LangType	No relationship
	DevType	Significant relationship
Cross company	LangType	Significant relationship
	DevType	Significant relationship

Since each nominal variable had two levels we replaced them each by dummy variables, coded 0 and 1. The final set of variables used for each data set is presented in Table 6.

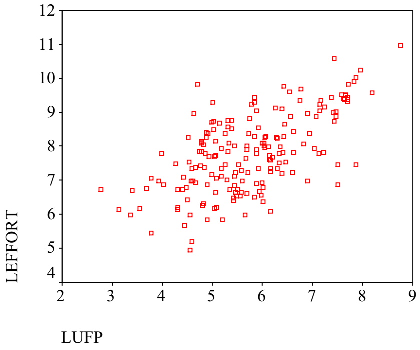


Figure 1 – Scatter plot of leffort and lufp for within-company data set.

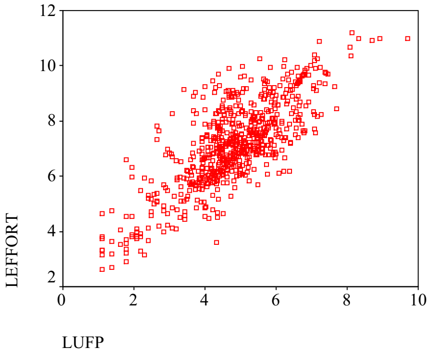


Figure 2 - Scatter plot of leffort and lufp for cross-company data set

Table 6. Variables used in the stepwise regression

Variable	Meaning
	Within company
leffort	Natural logarithm of teffort.
lufp	Natural logarithm of ufp.
Newdev	Dummy variable where 'new development' type is coded as 1 and 'enhancement' type is coded as 0
	Cross-company
leffort	Natural logarithm of teffort.
lufp	Natural logarithm of ufp.
Newdev	Dummy variable where 'new development' type is coded as 1 and 'enhancement' type is coded as 0
Fourthgl	Dummy variable where '4gl' language type is coded as 1 and '3gl' language type is coded as 0

2.3 Analysis Methods

To verify the **stability** of each cost model we used the following steps [10]:

S1. Use of a residual plot showing residuals vs. fitted values to investigate if the residuals are random and normally distributed.

S2. Calculate Cook's distance values [5] for all projects to identify influential data points. Any projects with distances higher than 3 × (4/*n*), where *n* represents the total number of projects, are immediately removed from the data analysis [15]. Those with distances higher than 4/n but smaller than (3 × (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 highly 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 and we employed the same statistics as Jeffery et al. [7] (e.g. MMRE, Median MRE, and Pred(25). However, in addition to calculating these statistics using the regression-based estimated effort, we also calculated these same statistics using the median effort for the cross-company and within-company data sets. This was done to have a benchmark for comparison.

3. Results

3.1 Results based on Cross-Company Data

The best cross-company fitting model is described in Table 7. Its adjusted R^2 was 0.591.

Table 7 Best Fitting Model to calculate leffort

Independent Variables	Coefficient	Std. Error	t	p>|t|
(constant)	2.849	0.143	19.894	0.000
lufp	0.897	0.029	31.132	0.000

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

$$\ln(\text{effort}) = 2.849 + 0.897 \times \ln(\text{ufp}) \qquad (1)$$

which, when transformed back to the raw data scale, gives the Equation:

$$\text{effort} = 17.27 \times \text{ufp}^{0.897} \qquad (2)$$

None of the dummy variables was selected by the model. In addition, this model only explains 59.1% of the variation in effort, suggesting that there are other contributing variables missing from this model.

Checking the model

The residual plot for the 680 projects showed several projects that seemed to have very large residuals. This trend was also confirmed using Cook's distance. 35 projects had their Cook's distance above the cut-off point (4/480), and of these 35 eight had values greater than 0.018 (3 times the cut-off value). These eight projects were permanently removed from the analysis. After re-fitting the model using 672 projects we found that 31 projects presented Cook's distance above the cut-off point.

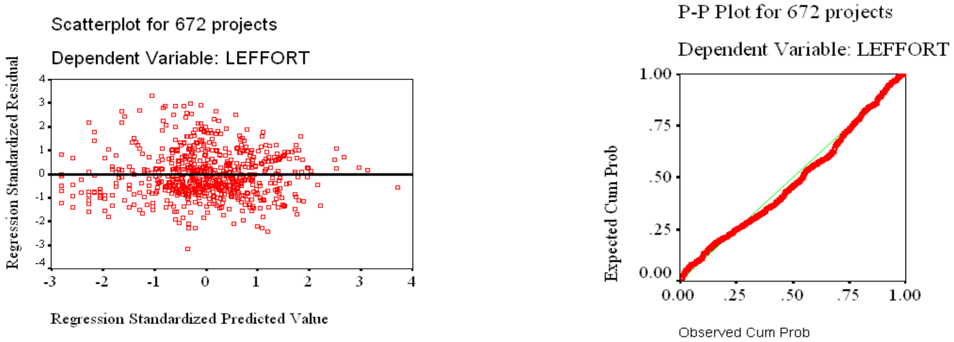
To check the model's stability, a new model was generated without the 31 projects that presented high Cook's distance, giving an adjusted R^2 of 0.646 (see table 8).

In the new model the independent variable remained significant and the coefficients have similar values to those in the previous model (see Table 7). Therefore, the 31 high influence data points were not removed.

Table 8 Best Fitting Model after removing 31 projects

Independent Variables	Coefficient	Std. Error	t	p>|t|
(constant)	2.622	0.137	19.140	0.000
lufp	0.937	0.027	34.140	0.000

The residual plot and the P-P plot for the final model are presented in Figure 3(a) and Figure 3(b) 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 which has been checked here is the normal distribution, and Figure 3(b) suggests that the residuals are normally distributed.



(a) **(b)**

Figure 3 – Residual and P-P plots for best fitting cross-company model

Measuring Prediction Accuracy

To assess the accuracy of the predictions for the cross-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 224 projects (33%) was omitted from the data set, and an Equation, similar to that shown by Equation 1, was calculated using the remaining 448 projects (66%). This Equation was then transformed back to the raw scale, giving an Equation similar to that shown by Equation 2. The estimated effort was calculated for all the projects omitted from the data set, and statistics such as MRE and absolute residual were also obtained.

This cross-validation approach was different from [7]. Jeffery et al. [7] employed a leave-one-out cross validation however it was only applied to their within-company data (14 projects). Our within-company data set is much larger than theirs and in addition we also use cross-validation to measure the prediction accuracy of the cross-company model. In addition, when using n-fold cross-validation, their analysis has been limited to a maximum of 14 training sets, which according to recent studies, may lead to untrustworthy results [13]. According to [13] ideally 20 sets or more should be deployed, so we have employed a 20-fold cross-validation.

The prediction accuracy statistics are presented in Table 9. We can see that the model's prediction accuracy was poor. However, its accuracy was significantly different from that based on the median effort of the data set (median = 1249.5) using the Wilcoxon matched-paired signed rank test on absolute residuals. The residuals obtained using the regression model were generally smaller than those obtained using the median effort, indicating that estimates based on a regression model provided better accuracy than those based on the median effort.

The differences between values obtained for medians and means for the MREs suggest that the data set contains several outliers. The box plots of absolute residuals using estimates based on regression model also show the existence of numerous outliers (see Figure 4).

Table 9 Prediction accuracy statistics for the cross-company data set

Prediction Accuracy	Estimates based on regression model
MMRE	97.8%
MdMRE	61.7%
Pred(25)	21%
Prediction Accuracy	Estimates based on median effort
MMRE	297%
MdMRE	75%
Pred(25)	16.8%

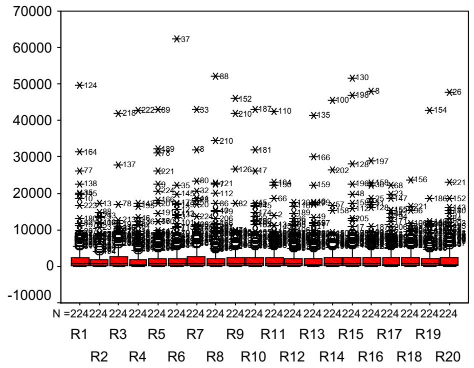


Figure 4 – Box plots for absolute residuals using estimates based on regression model

3.2 Results based on Within-Company Data

The best within-company fitting model is described in Table 10. Its adjusted R^2 was 0.388.

Table 10 Best Fitting Model to calculate leffort

Independent Variables	Coefficient	Std. Error	t	p>|t|
(constant)	4.162	0.343	12.136	0.000
lufp	0.635	0.059	10.824	0.000

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

$$\ln(\text{effort}) = 4.162 + 0.635 \times \ln(\text{ufp}) \quad (3)$$

which, when transformed back to the raw data scale, gives the Equation:

$$\text{effort} = 64.2 \times \text{ufp}^{0.635} \quad (4)$$

None of the dummy variables was selected by the model. In addition, this model only explains 38.8% of the variation in effort, even worse than the cross-company model, suggesting that there are other contributing variables missing from this model.

Checking the model

The residual plot for the 187 projects showed some projects that seemed to have very large residuals. This trend was also confirmed using Cook's distance. 11 projects had their Cook's distance above the cut-off point (4/187), and of these 11, three had values greater than 0.06 (3 times the cut-off value). These three projects were permanently removed from the analysis. After re-fitting the model using 184 projects we found that eight projects presented Cook's distance above the cut-off point.

To check the model's stability, a new model was generated without the eight projects that presented high Cook's distance, giving an adjusted R^2 of 0.431 (see table 11). In the new model the independent variable remained significant and the coefficients have similar values to those in the previous model (see Table 11). Therefore, the eight high influence data points were not removed.

Table 11 Best Fitting Model after removing eight projects

Independent Variables	Coefficient	Std. Error	t	p>|t|
(constant)	4.076	0.328	12.410	0.000
lufp	0.652	0.057	11.520	0.000

The residual plot and the P-P plot for the final model are presented in Figure 5(a) and Figure 5(b) respectively. Figure 5(b) suggests that the residuals are normally distributed.

Measuring Prediction Accuracy

To assess the accuracy of the predictions for the within-company model we also employed a 20-fold cross-validation to the data set, using the raw scale and a 66% split. This means that 20 times a randomly generated set of 62 projects (33%) was omitted from the data set, and an Equation, similar to that shown by Equation 3, was calculated using the remaining 122 projects (66%).

This Equation was then transformed back to the raw scale, giving an Equation similar to that shown by Equation 4. Then the estimated effort was calculated for all the projects that had been omitted from the data set, and likewise, statistics such as MRE and absolute residual were also obtained.

The prediction accuracy statistics are presented in Table 12, where we can see that the model's prediction accuracy was not very good. However, its accuracy was significantly different from predictions based on the median effort of the data set (median = 2418) using the Wilcoxon matched-paired signed rank test on absolute residuals.

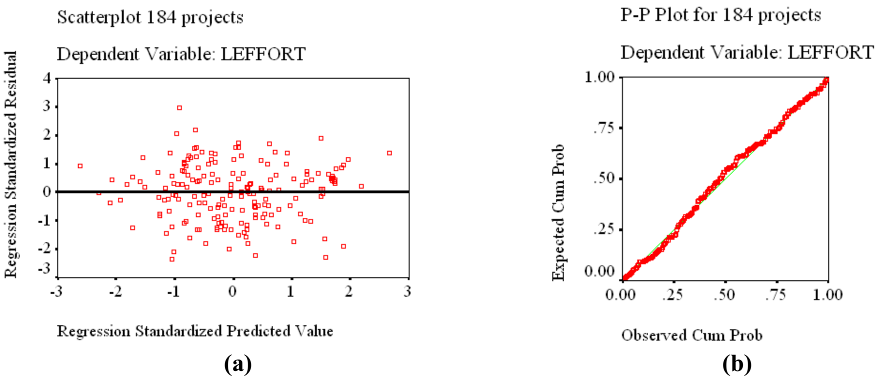


Figure 5 – Residual and P-P plot for best fitting within-company model

Table 12 Prediction accuracy statistics for the within-company data set

Prediction Accuracy	Estimates based on regression model
MMRE	102%
MdMRE	60%
Pred(25)	20.8%
Prediction Accuracy	Estimates based on median effort
MMRE	137.5%
MdMRE	52%
Pred(25)	14%

The residuals obtained using the regression model were generally smaller than those obtained using the median effort, indicating again that estimates based on a regression model provided better accuracy than those based on the median effort.

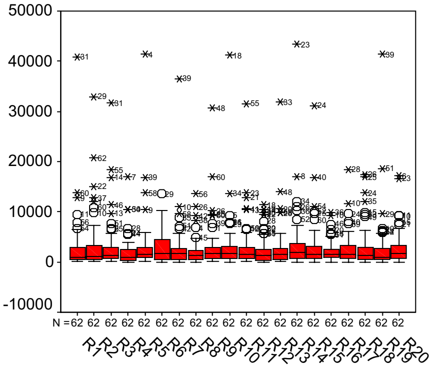


Figure 6 – Box plots for absolute residuals using estimates based on within-company regression model

The box plots of absolute residuals using estimates based on regression model also shows the existence of numerous outliers (see Figure 6).

3.3 Using Within-company data as validation set for Cross-company model

We used the cross-company model represented by Equation 2 to estimate effort for all the 184 within-company projects, which were used as our validation set. The prediction accuracy statistics were calculated using the regression model and also based on the median effort for the within-company data set (median = 2418) (see Table 13). Now the accuracy of estimates based on the regression model were not significantly different from predictions based on the median effort of the data set using the Wilcoxon matched-paired signed rank test on absolute residuals. These results show that the median effort would provide as good predictions as those obtained using the cross-company model.

Table 13 Prediction accuracy statistics for the within-company data set using cross-company model

Prediction Accuracy	Estimates based on regression model
MMRE	123%
MdMRE	61%
Pred(25)	20.6%
Prediction Accuracy	Estimates based on median effort
MMRE	131%
MdMRE	72%
Pred(25)	16.8%

The residuals obtained using the cross-company model were highly skewed and presented numerous outliers.

3.4 Answering our Research Questions

The research questions addressed in this study are as follows:

1. How successful is a cross-company model at estimating effort for projects from a single company, when the model is built from a data set that does not include that company;

2. How successful is a cross-company model, compared to a within-company model.

Our first research question is addressed by the results from Section 3.3. The accuracy of estimates obtained for the 184 within-company projects using the cross-company model (see Equation 2) does not indicate good prediction accuracy. MMRE is 123%, which is poor (25% is considered "good" [4], and Pred(25) is also poor (21%, when 75% indicates a good prediction model). The absolute residuals obtained using the cross-company model were not significantly different from residuals obtained using the median effort. This suggests that there is no advantage to the single company obtaining their effort estimations using the cross-company model.

To address our second research question we compared the absolute residuals for using the 184 within-company projects with the within-company models (see Section 3.2) to those obtained using the same projects with the cross-company model (see Section 3.3). The comparison was done using the Mann-Whitney Test for two independent samples. The results indicated that absolute residuals for the within-company projects using within-company models were not significantly different from absolute residuals obtained for the within-company projects using a cross-company model. Figure 7 shows absolute residuals for both groups. RESWCCC and RESWCWC stand for Residuals for within-company projects using a cross-company model and residuals for within-company projects using a within-company model, respectively.

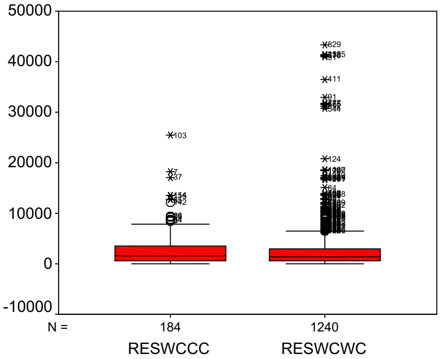


Figure 7 – Box plots for absolute residuals

4. Discussion

The results for our second research question suggest that the single company will not obtain better effort estimates using a model based on its own historical data, compared to estimates obtained from a cross-company model, or estimations based on a median effort. These results do not converge with those presented in [7], however they do corroborate findings from previous studies, using different data sets [1],[2],[20].

Previous studies have endeavoured to explain possible circumstances under which cross-company models are likely to be as accurate as within-company models. Data collection following rigorous quality assurance procedures seems to be a strong candidate [1],[10],[17],[20]. The evidence we have of a replication study supporting the claim that strong quality assurance procedures facilitate similar prediction accuracy between cross-company and within-company models is provided by [20] using the Laturi database. This database is the only one used so far in previous studies that applied rigorous quality assurance mechanisms to their data collection from the start. Both original [1] and replicated [20] studies consistently showed no differences in prediction accuracy between cross-company and within-company models. The ESA database did not initially have strict quality assurance mechanisms [16], however once such mechanisms were incorporated results also showed no differences in prediction accuracy between cross-company and within-company models [2].

The ISBSG and Tukutuku databases do not have strict data quality assurance procedures. Except for our study, all previous studies using these databases consistently show differences between prediction accuracy using cross-company and within-company models [6],[7],[10],[17]. We believe one of the reasons for our results is the amount of missing data, which hindered the inclusion of important variables in our models, in particular for the within-company model. Its adjusted R^2 indicated that there were other influential variables not included in the model, which may have largely affected the results used to address our second research question.

Another factor that may have influenced the results obtained in the different studies is the process used to construct the various models, as it appears the way models are constructed can affect the results [17]. For example, were the variables used in cross-company models selected by analysing the full data set, where parameters are recalibrated after removing the within-company data? Or were they selected after removing the within-company data? Mendes and Kitchenham [17] used both approaches and obtained cross-company models with different prediction accuracy. This suggests that the way to construct the models can affect the results. Unfortunately, except for [20],[17], and [10], previous studies did not provide details on the methods employed for building the cross-company and within-company models, making it impossible to consider the impact of the model construction process.

Perhaps an additional explanation for the results for our second research question is the similarity of application domains between our within-company and cross-company data sets. It seems that when more specific information cannot be used to select a tightly focused cross-company data set, how well a cross-company model performs will depend on how broadly similar the cross-company and within-company projects are. Similar results using data sets of similar application domains has also been obtained in [1], [2].

Neither the within-company model nor the cross-company model performed well, in terms of MMRE and Pred(25).

5. Conclusions

For the data set used in this study, we found that the predictions obtained for a single company using a cross-company model were similar in accuracy to those this company would obtain using its own within-company model. To our knowledge, this is the first study to show these results using a database with no strict quality assurance procedures.

Our results contradict those presented in [7], however corroborate those from previous studies [1][2][20].

Future work in this area will concentrate in two areas. One is to narrow our attention to specific domains: data sets will be smaller but also more homogeneous. The second is to consider other validation approaches, particularly chronological splitting where models are built from older projects and validated on newer projects, as in [14].

Acknowledgments

We would like to thank the ISBSG group for making Release 9 available for our research and all those companies that have volunteered data on their projects. I would also like to thanks Dr. N. Mosley[3] for his comments on a previous draft of this paper.

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] Cartwright, M. H., Shepperd, M. J., Song, Q. Dealing with Missing Software Project Data. Proceedings 9th International Symposium on Software Metrics. Metrics 2003, Sydney September 3-5th 2003, IEEE Computer Society, pp 154-165, 2003.

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

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

[6] 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.

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

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

[9] Kitchenham, B.A. A procedure for analysing unbalanced data sets. IEEE Trans. Software Engineering. 24(4), 1998, pp 278-301.

[10] Kitchenham, B.A., and E. Mendes. A Comparison of Cross-company and Within-company Effort Estimation Models for Web Applications, Proceedings EASE 2004, 2004, pp 47-55.

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

[12] Kitchenham, B.A., L.M. Pickard, S.G. MacDonell and M.J. Shepperd. What accuracy statistics really measure. IEE Proceedings - Software, 148(3), June 2001, pp 81-85.

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

[14] 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.

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

[16] 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.

[17] 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.

[18] Mendes, E., N. Mosley, and S. Counsell, A Replicated Assessment of the Use of Adaptation Rules to Improve Web Cost Estimation, Proceedings of ISESE'2003 Conference, Rome, September, 2003, pp 100-109.

[19] Shepperd, M.J., and G. Kadoda, Using Simulation to Evaluate Prediction Techniques, Proc. IEEE 7th International Software Metrics Symposium, London, UK, 2001, pp. 349-358.

[20] 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.

[21] Wilcoxon, F. Individual comparisons by ranking methods. Biometrics, 1, 1945, pp 80-83.

Note. The data set can be made available to the reviewers for independent assessment of the statistical analyses presented in this paper but cannot be published for confidentiality reasons. Please contact Emilia Mendes at emilia@cs.auckland.ac.nz.

[3] MetriQ (NZ) Limited, http://www.metriq.biz