How Valuable is it for a Web company to Use a Cross-company Cost Model, Compared to Using Its Own Single-company Model?

Emilia Mendes
The University of Auckland
Private Bag 92019
Auckland, New Zealand
0064 9 3737599 ext. 86137
emilia@cs.auckland.ac.nz

Gopi Dinakaran
The University of Auckland
Private Bag 92019
Auckland, New Zealand
0064 9 3737599
gdin007@ec.auckland.ac.nz

Nile Mosley
MetriQ Limited
19 Clairville Crescent, Glendowie
Auckland, New Zealand
0064 9 5750323
info@metriq.biz

ABSTRACT

Previous studies comparing the prediction accuracy of cross-company and single-company models have been inconclusive, and as such replicated studies are necessary to determine under what circumstances a company can place reliance on a cross-company model.

This paper therefore replicates a previous study by investigating how successful a cross-company cost model is: i) to estimate effort for Web projects that belong to a single company and were not used to build the cross-company model; ii) compared to a single-company cost model. Our single-company data set had data on 20 Web projects from a single company and our cross-company data set data on 30 Web projects from 25 different companies.

Our results were similar to those from the replicated study, showing that predictions based on the single-company model were significantly more accurate than those based on the cross-company model. We analysed the data using two techniques, forward stepwise regression and case-based reasoning.

Keywords

Cross-company cost model, single-company cost model, cost estimation, stepwise regression, case-based reasoning, Web projects, Web applications.

1. INTRODUCTION

One of the issues faced by Web companies is if it is worthwhile to obtain estimates for their new projects based on cross-company models, i.e. cost models built using data from other Web companies. The use of a cross-company model seems particularly useful for Web companies that do not have past projects from which to develop their own models, or that have projects in different application domains, using different technologies.

Previous studies have suggested that single-company models are needed to produce accurate effort estimates (e.g. [8],[10]). However, three main problems can occur when relying on single-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 data set is large to be of use, 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 models for effort estimation and productivity benchmarking. However, the use of cross-company models has problems of its own [2],[14]:

i) Care is 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.
iii) Uniform data collection control across different companies must be provided.
iv) The ability to partition projects (e.g. according to their completion dates) to identify those that used current development practices from those that did not.
v) 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.

Ten studies in Software and Web engineering have investigated if cross company models can be as accurate as single-company models [1],[2],[6],[7],[9],[11],[13],[14],[16],[18].

Eight studies used data from two application domains: 'business' and 'space and military'. Their findings were as follows:

> Four studies found that a cross-company model gave similar prediction accuracy to that of a single-company model [1],[2],[16],[18].
> Four studies found that a cross-company model did **not** give as accurate predications as a single-company model [6],[7],[11],[13].

Two further studies have investigated the same issues on the effectiveness of cross-company cost models using data from Web projects [9],[14] obtained from a single database. Both found that a cross-company model did **not** give as accurate predictions as a single-company model.

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

To our knowledge, the last published study that compared effort prediction accuracy between cross-company and single-company cost models based on Web project data was published over a year ago [14], and was an extended analysis of a previous study [9]. Both studies employed data on Web projects from the Tukutuku database [15]. Since then another 50 Web projects have been volunteered to this database, which may have an impact on the results observed previously.

Therefore this paper's contribution is twofold: i) to replicate Mendes and Kitchenham's study [14], using Web project data volunteered after that study was carried out; ii) to provide a procedure to help Web companies perform cost model evaluations and comparisons.

Replicated studies are necessary to determine under what circumstances a company can place reliance on a cross-company model [14], and is fundamental to establishing the validity and generalisability of results [19].

The two research questions addressed in our study are as follows:

i) How successful is a cross-company model at estimating effort for projects from a single company?
ii) How successful is a cross-company model, compared to a single-company model?

We need to investigate both research questions in combination because obtaining results where a cross-company model provides good prediction accuracy for single-company projects is not enough to say that the cross-company model is successful. A cross-company model also needs to provide accuracy similar to or better than that provided by the single-company in order to be considered successful.

Table 1 - Comparison of previous studies

	Study 1 [13]	Study 2 [1]	Study 3 [2]	Study 4 [6]	Study 5 [7]	Study 6 [18]	Study 7 [11]	Study 8 [9]	Study 9 [14]	Study 10 [16]
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 Megated: 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					

Both research questions are addressed using data on 50 Web projects, where 20 come from a single company, and 30 come from other 25 companies. These projects were volunteered to the Tukutuku database[1] after the previous study on Web projects [14] was carried out.

Like [14], we used forward stepwise regression and case-based reasoning to build cost models and obtain effort estimates. Also like [14], we measured prediction accuracy based on de facto measures of accuracy, such as the mean Magnitude of Relative Error (MMRE), median MRE, and Prediction at 25% (Pred(25)) [4]. We also used the mean and median of absolute residuals, where residuals are calculated as actual effort – estimated effort, and the Companies' effort estimates, based on an educated guess.

Pred(*n*) measures the percentage of estimates that are within *n*% of the actual values, and *n* is usually set at 25%. MRE is the basis for calculating MMRE and MdMRE, and defined as:

$$\text{MRE} = \frac{|e - \hat{e}|}{e} \qquad (1)$$

where e represents actual effort and $\hat{e}$ estimated effort.

The difference between MMRE and MdMRE is that the former is sensitive to predictions containing extreme MRE values.

Within the scope of our investigation a Web project can either represent a Web hypermedia or Web software application [3]. The former is characterised by the authoring of information using nodes (chunks of information), links (relations between nodes), anchors, access structures (for navigation) and its delivery over the Web. Technologies commonly used for developing such applications are HTML, JavaScript and multimedia. In addition, typical developers are writers, artists and organisations that wish to publish information on the Web and/or CD-ROMs without the need to use programming languages such as Java. Conversely, the latter represents software applications that depend on the Web or use the Web's infrastructure for execution. Typical applications include legacy information systems such as databases, booking systems, knowledge bases etc. Typically they employ development technologies (e.g., DCOM, ActiveX etc), database systems, and development solutions (e.g. J2EE). Typical developers are young programmers fresh from a Computer Science or Software Engineering degree, managed by more senior staff.

The remainder of the paper is organised as follows: Section 2 describes the research method employed in this study, and results

[1] http://www.cs.auckland.ac.nz/tukutuku

are presented in Section 3. Section 4 looks at the same issues presented in Section 3 however employing case-based reasoning as our technique for obtaining effort estimates. A discussion of the results is provided in Section 5, and conclusions are given in Section 6.

2. RESEARCH METHOD

2.1 Data Set Description

The analysis presented in this paper was based on Web projects from the Tukutuku database [15]. These projects represent industrial Web applications developed by Web companies worldwide. This database is part of the Tukutuku project, which aims to collect data about Web projects, to be used to develop Web cost estimation models and to benchmark productivity across and within Web Companies.

The Tukutuku database has data on 117 projects where:

Projects come from 9 different countries, mainly New Zealand (64%), Brazil (11%), United States (7%), Canada (5%), and England (4%).

Development types are enhancement projects (56%), and new developments (44%).

The applications are mainly Functional (27%), Corporate (17%), and eCommerce (15%).

The languages used are mainly HTML (88%), Javascript (DHTML/DOM) (76%), PHP (50%), Various Graphics Tools (39%), ASP (VBScript, .Net) (18%), and Perl (15%).

The analysis presented in this paper used data from 50 Web projects where 20 projects come from a single company, and the remaining 30 come from another 25 companies. These 50 projects were volunteered after study [14] was carried out.

Each Web project in the database provided 44 variables to characterise a Web application and its development process (see Table 2).

The size measures and cost drivers employed represent early Web size measures and cost drivers obtained from the results of a survey investigation [15], using data from 133 on-line Web forms aimed at giving quotes on Web development projects. In addition, these measures and cost drivers have also been confirmed by an established Web company and a second survey involving 33 Web companies in New Zealand. Consequently it is our belief that the 45 variables identified are measures that are meaningful to Web companies and are constructed from information their customers can provide at a very early stage in project development.

Table 2 - Variables for the Tukutuku database

Variable Name	Scale	Description
Country	Nominal	Country company belongs to
Established	Ordinal	Amount of time company has been established
Services	Nominal	Services Company provides
ClientInd	Nominal	Industries representative of clients
TypeProj	Nominal	Type of project (New, Enhancement)
AppDom	Nominal	Application domain
Languages	Nominal	Implementation languages used
nlang	Ratio	Number of different languages used
DocProc?	Nominal	Project followed defined and documented process
ProcImpr?	Nominal	Development team involved in a process improvement programme
Metrics?	Nominal	Development team part of a software metrics programme
devteam	Ratio	Size of development team
teamexp	Ratio	Average team experience with the development language(s) employed
Webpages	Ratio	Number of web pages
newWP	Ratio	Number of New Web pages
Wpcustom	Ratio	Web pages given by the customer
Wpout	Ratio	Web pages developed by third party
WpOwnCo	Ratio	Web pages reused from own company
txtTyped	Ratio	Number text pages typed (~600 words)
txtElec	Ratio	Number text pages electronic format
txtScan	Ratio	Number text pages scanned
imgNew	Ratio	Number new images
Img3rdP	Ratio	Number images developed by third party (not the customer)
imgScan	Ratio	Number images scanned
imgLib	Ratio	Number images reused from a library
imgOwnCo	Ratio	Number of images reused by own company
Animnew	Ratio	Number new animations
AnimLib	Ratio	Number animations reused from a library
AVNew	Ratio	Number new audio/video files
AVLib	Ratio	Number reused audio/video files
TotDiffPro	Ratio	Number <> products application offers
HEffDev	Ratio	Minimum number of hours to develop a single function/feature by one experienced developer that is considered high (above average).[2]
HEffAdpt	Ratio	Minimum number of hours to adapt a single function/feature by one experienced developer that is considered high (above average).[3]
hfots	Ratio	Number of reused High effort features/functions without adaptation
hfotsA	Ratio	Number of reused High effort features/functions adapted
hnew	Ratio	Number of new High effort features/functions
tothigh	Ratio	Total Number of high effort features/functions
fots	Ratio	Number of reused Low effort features/functions without adaptation
fotsa	Ratio	Number of reused Low effort features/functions adapted
new	Ratio	Number of new Low effort features/functions
totnhigh	Ratio	Total Number of low effort features/functions
toteffor	Ratio	Actual effort to develop the Web application
esteff	Ratio	Estimated effort to develop the Web application
accuracy	Nominal	Procedure used to record effort data

2.2 Data Quality

Web companies that volunteered data for the Tukutuku database did not use any automated measurement tools or quality control procedures for data collection. Therefore the accuracy of their data cannot be determined. In order to identify guesstimates from more accurate effort data, we asked companies how their effort data was collected (see Table 3).

Table 3 - How effort data was collected

Data Collection Method	# of Projects	% of Projects

[2] this number is currently set to 15 hours based on the collected data.

[3] this number is currently set to 4 hours based on the collected data.

Hours worked per project task per day	60	51
Hours worked per project per day/week	32	27
Total hours worked each day or week	13	11
No timesheets (guesstimates)	12	10

At least for 79% of Web projects in the Tukutuku database effort values were based on more than guesstimates. In relation to the 50 projects used in this study, 67% of the 30 cross-company projects and 100% of the 20 single-company projects are also more than guesstimates. However, we are also aware that the use of timesheets does not guarantee 100% accuracy in the effort values recorded.

2.3 Modelling Techniques

Like [14], the techniques used to build both cross-company and single-company models were forward stepwise regression (SWR) and case-based reasoning (CBR). Except for CBR, all results presented here were obtained using the statistical software SPSS 10.1.3 for Windows. Finally, all the statistical significance tests used $\alpha = 0.05$.

CBR is a branch of Artificial Intelligence where knowledge of similar past cases is used to solve new cases [17]. Within the context of our investigation, cases are Web projects, and each case is represented by a set of project attributes (e.g. number of Web pages). The similarity between cases can be measured in different ways, however, like [14], the similarity measure used in this study is the Euclidean distance. In addition, all the project attributes considered by the similarity function had equal influence upon the selection of the most similar project(s).

Stepwise regression [12] is a statistical technique whereby a prediction model (Equation) is built, and represents the relationship between independent (e.g. Webpages) and dependent variables (e.g. toteffort). This technique builds the model by adding, at each stage, the independent variable with the highest association to the dependent variable, taking into account all variables currently in the model. It aims to find the set of independent variables (predictors) that best explains the variation in the dependent variable (response).

To build the cross-company (CC) and single-company (SC) cost models we used a similar set of variables to that employed in [14], and also excluded variables based on the following criteria:

More than 50% of instances of a variable were zero.

The variable was categorical (nominal and ordinal).

The variable was related to another variable, in which case both could not be included in the same model. To measure the strength of the association between variables we used the Spearman's rank correlation statistical test.

The motivation for Mendes and Kitchenham [14] to exclude categorical variables from their analysis was that the Tukutuku categorical variables had many levels, thus requiring a large number of dummy variables which rapidly reduce the degrees of freedom for analysis.

Table 4 shows the variables used by Mendes and Kitchenham [14], and the variables used in our analysis. Several variables could not be employed in our study since they presented too many zeros.

Table 4 – Variables used in the studies

	Our study	
Mendes and Kitchenham	**SC data set**	**CC data set**
nlang	nlang	nlang
devTeam	devTeam	devTeam
teamexp	teamexp	teamexp
newWP	Webpages	Webpages
ImgNew	ImgNew	ImgNew
ImgLib	Too many zeros	Too many zeros
Img3rdP	Too many zeros	Too many zeros
hfotsa	Too many zeros	Too many zeros
tothigh	Too many zeros	Too many zeros
fotsa	fotsa	fotsa
totnhigh	totnhigh	totnhigh
toteffort	toteffort	toteffort
esteffort	esteffort	esteffort

Summary statistics for the variables used in our study are presented in Table 5. Table 5 suggests that there are clear differences between the single-company projects and cross-company projects regarding their effort and application size in: number of Web pages, new images and low-effort functions/features.

Table 5 – Summary statistics for variables

Single-company data – 20 projects					
Variables	Mean	Median	Std. Dev.	Min.	Max.
nlang	3.00	3.00	0.00	3	3
devTeam	1.75	2.00	0.72	1	3
teamexp	2.00	2.00	0.00	2	2
Webpages	19.70	9.00	32.26	1	134
ImgNew	1.25	1	1.48	0	5
fotsa	2.50	2.00	2.04	0	8
totnhigh	3.90	2.50	4.09	1	16
toteffort	7.83	4.71	7.16	1.1	22
esteffort	6.63	4.00	6.60	1	22
Cross-company data – 30 projects					
nlang	3.20	4	1.77	1	7
devTeam	3.97	3	4.40	1	23
teamexp	3.62	2.5	2.36	1	10
Webpages	38.57	22.5	38.12	0	135
ImgNew	74.53	22	198.81	0	1000
fotsa	2.37	1	3.81	0	16
totnhigh	5.30	4	5.77	0	20
toteffort	163.14	49	230.56	2	1000
esteffort	470.43	46.5	1815.17	2	10020

The average size of applications for the single-company data set is around 20 Web pages, 1 image and 4 low-effort features. However, their corresponding medians are 9, 1 and 2.5 respectively, which indicates that half the Web applications in the data set were constructed with less than 9 Web pages, 1 image and 2.5 low-effort features. In contrast, the average size of applications for the cross-company data set is around 39 Web pages, 75 new images, and 5 low-effort features. Their corresponding medians are 22.5, 22 and 4 respectively, which indicates that half the Web applications in the data set were constructed with up to 22.5 Web pages, 22 new images, and 4 low-effort features. On average cross-company projects are about twice the size in Web pages, new images and number of low-effort functions/features, compared to single-company projects.

The median total effort for cross-company applications (49 person hours) is also much larger than that for single-company applications (4.7 person hours). In addition, single-company projects present much higher productivity (size/effort) than cross-company projects.

The summary statistics in Table 5 also show that there is least one application in the cross-company data set with no Web pages. When such situation occurs it is customary to check the value with the data provider. We contacted the company that volunteered this project data but did not receive a reply. We also noticed another inconsistency in the data for this same project thus our course of action was to remove this project from the cross-company data set.

Whenever variables were highly skewed they were transformed before being used in the forward stepwise procedure. This was done in order to comply with the assumptions underlying stepwise regression [12] (e.g. residuals should be independent and normally distributed; relationship between dependent and independent variables should be linear). The transformation we employed was to take the natural log (ln), which makes larger values smaller and brings the data values closer to each other [12]. A new variable containing the transformed values was created for each original variable that needed to be transformed. All new variables are identified as *Lvarname*, e.g. *Lnlang* represents the transformed variable *nlang*.

In addition, whenever a variable needed to be transformed but had zero values, the natural logarithmic transformation was applied to the variable's value after adding 1.

Table 6 - Variables used in the stepwise regression

SC data set	CC data set	Meaning
devTeam	Lnlang	Natural log. of number of different languages used
Webpages	LdevTeam	Natural log. of size of development team
fotsa	Lteamexp	Natural log. of average team experience
totnhigh	LWebpages	Natural log. of number of Web pages
toteffort	Lfotsa	Natural log. of (1+reused low-effort features adapted)
	Ltotnhigh	Natural log. of (1+number of low-effort features/functions)
	Ltoteffort	Natural log. of total effort to develop a Web application.
	LImgNew	Natural log. of (1+number of new images in the application)

The set of variables used in the stepwise regression and case-based reasoning is presented in Table 6. *Nlang* and *teamexp* were not included in the stepwise procedure to build the single-company model because they presented constant values for all the projects. None of the SC variables needed to be transformed.

The variable *toteffort* was used as the dependent variable when building the best single-company model, however *Ltoteffort* was the one employed as dependent variable when building the cross-company model.

2.4 Steps to Follow to Answer Our Research Questions

This Section details the steps that need to be carried out to answer each of the two research questions this study investigated. These were the same steps used in [14]. Both questions are also presented to provide a context for each set of steps.

Question 1: How successful is a cross-company model at estimating effort for projects from a single company?

Steps to follow:

1) Apply forward stepwise regression to build a cross-company cost model using the cross-company data set. Not applicable to CBR.
2) If model is not linear, transform the model back to the raw data scale. Not applicable to CBR.
3) Use the model in step 2 to estimate effort for each of the 20 single-company projects. The single-company projects are the validation set used to obtain effort estimates. The estimated effort obtained for each project is also used to calculate accuracy statistics (e.g. MRE). The equivalent for CBR is to use the cross-company data set as a case base to estimate effort for each of the 20 single-company projects.
4) The overall model accuracy is aggregated from the validation set (e.g. MMRE, MdMRE). Same for CBR.

These steps are used to simulate a situation where a single company uses a cross-company model to estimate effort for its new projects.

Question 2: How successful is a cross-company model, compared to a single-company model?

Steps to follow:

1) Apply forward stepwise regression to build a single-company cost model using the single-company data set. Not applicable to CBR.
2) Obtain the prediction accuracy of estimates for the model obtained in 1) using a leave-one-out cross-validation. Cross-validation is the splitting of a data set into training and validation sets. Training sets are used to build models and validation sets are used to validate models. A leave-one-out cross-validation means that the original data set is divided into *n* different subsets (*n* is the size of the original data set) of training and validation sets, where each validation set has one project. The equivalent for CBR is to use the single-company data set as a case base, after removing one project, and then to estimate effort for the project that has been removed. This step is iterated 19 times, each time removing a different project.
3) The overall model accuracy is aggregated across the 19 validation sets. Same for CBR.
4) Compare the accuracy obtained in Step 3 to that obtained for the cross-company model. Same for CBR.

Steps 1 to 3 simulate a situation where a single company builds a model using its own data set, and then uses this model to estimate effort for new projects.

2.5 Analysis Methods

To verify the stability of each cost model built using forward stepwise regression the following steps were employed [9]:

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 [5] 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 [12]. 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 adjusted R^2 (goodness of fit) improves, the highly influential projects are retained in the data analysis.

3. RESULTS

3.1 Results Based on Cross-company Data

The best cross-company regression model is described in Table 7. Its adjusted R^2 is 0.43, thus explaining only 43% of the variation in effort and suggesting that there are other contributing variables missing from this model. This model is much worse than that built in [14], which provided an adjusted R^2 of 0.63. However, both models are exponential.

Table 7 - Best Fitting Model to calculate Ltoteffort

Independent Variables	Coefficient	Std. Error	t	p>|t|
(constant)	-0.208	1.082	-0.192	0.849
LWebpages	0.801	0.261	3.069	0.005
Lnlang	0.923	0.349	2.645	0.014
LImgNew	0.316	0.127	2.478	0.020

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

$$\text{Ltoteffort} = -0.208 + 0.801\text{LWebpages} + 0.923\text{Lnlang} + 0.316\text{LImgNew} \quad (2)$$

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

$$\text{toteffort} = 0.812 \times \text{Webpages}^{0.801} \times \text{nlang}^{0.923} \times \text{ImgNew}^{0.316} \quad (3)$$

Finally, the variables selected by this model are different from those selected in [14], which were *LnewWP*, *devTeam*, and *Ltothigh*.

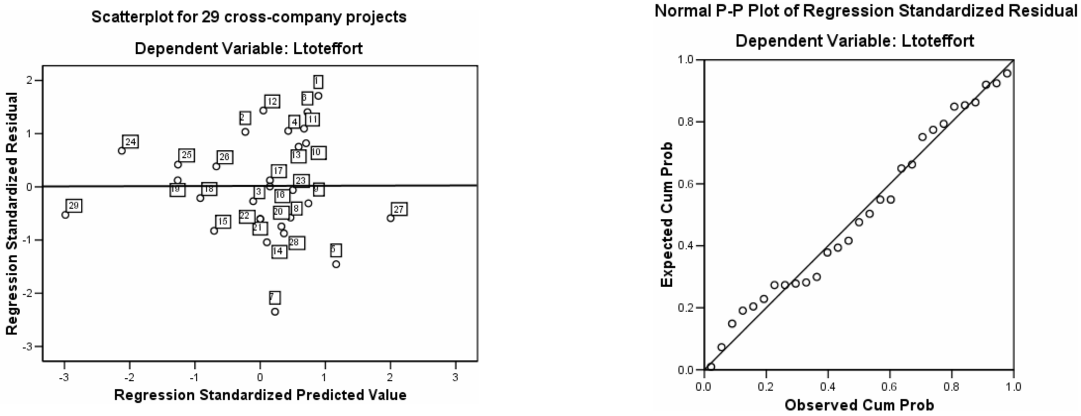
Checking the Model

The residual plot (see Figure 1(a)) for the 29 projects showed that 2 projects seemed to have large residuals. This trend was also confirmed using Cook's distance, where these two projects presented a Cook's distance greater than 4/29. To check the model's stability, a new model was generated without those two projects, giving an adjusted R^2 of 0.51. In the new model the independent variables remained significant and the coefficients had similar values to those in the previous model. Therefore, the two high influence data points were not permanently removed.

The residual plot and the P-P plot (Probability plot) for the final model are presented in Figure 1(a) and Figure 1(b) respectively. P-P Plots are normally employed to verify if 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 1(b) suggests that the residuals are normally distributed.

Measuring Prediction Accuracy

To assess the accuracy of the predictions for the cross-company model we used as validation set the 20 projects from the single-company data set.



(a) (b)

Figure 1 – Residual and P-P plots for best fitting cross-company model

The prediction accuracy statistics are presented in Table 9, where we can see that the predictions using the stepwise regression model are very poor, assuming that a good model should present MMRE and MdMRE close to 25% and Pred(25) of at least 75% [4]. In addition, the cross-company model's prediction accuracy was significantly worse than both the estimate accuracy provided using expert opinion and predictions based on the median of the data set (4.71). This pattern was confirmed by comparing the absolute residuals using the Wilcoxon matched-paired signed rank test, a statistical test that compares the distributions of two variables. If both variables present very different distributions they are assumed to be significantly different from one another.

The estimates based on the median model were also significantly worse than those provided using expert opinion. What these results suggest is that a single company is better off using their own experts' estimates than the median effort or a regression-based cross-company model. Both mean and median absolute residuals confirm this trend. These results partially corroborate those obtained in [14], where Mendes and Kitchenham found that the cross-company model presented significantly worse prediction than estimates based on expert opinion, however similar predictions to those using a median model.

The differences between values obtained for medians and means for the MREs suggest that the data set contains several outliers.

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

Prediction Accuracy	Estimates based on regression model	Estimates made by company personnel	Estimates based on median model
Mean MRE (MMRE)	312%	28.5%	92%
Median MRE (MdMRE)	105%	23.4%	69%
Pred(25)	0%	55%	25%
Mean absolute residual	21	2.4	5.3
Median absolute residual	11	1.4	3

3.2 Results Based on Single-company Data

The best single-company fitting model is described in Table 10. Its adjusted R^2 was 0.727.

Table 10 - Best Fitting Model to calculate toteffort

Independent Variables	Coefficient	Std. Error	t	p>ltl
(constant)	-7.663	2.315	-3.310	0.004
devTeam	7.206	1.245	5.786	0.000
fotsa	1.152	0.438	2.634	0.017

Checking the model

The residual plot for the 20 projects showed one project that seemed to have a large residual (see Table 11). This trend was also confirmed using Cook's distance, where this project had its Cook's distance above the cut-off point (4/20).

Table 11 – Project with Cook's distance > 0.2

devTeam	Webpages	ImgNew	fotsa	totnhigh	toteffort
3	30	4	1	1	21

To check the model's stability, a new model was built without this project, giving an adjusted R^2 of 0.703, which is smaller than that for the previous model. In the new model the independent variables remained significant; however the coefficients had very different values to those in the previous model, indicating that the high influence data point had to be permanently removed.

The best single-company model, after removing the high-influence project, is described in Table 12. Its adjusted R^2 is 0.703, indicating that the model explains 70% of the variation in effort. This model is not as good as the one described in [14] (adjusted R^2 of 0.95), however both have selected one common variable – *fotsa*. In addition, our model is linear whereas the one in [14] was exponential.

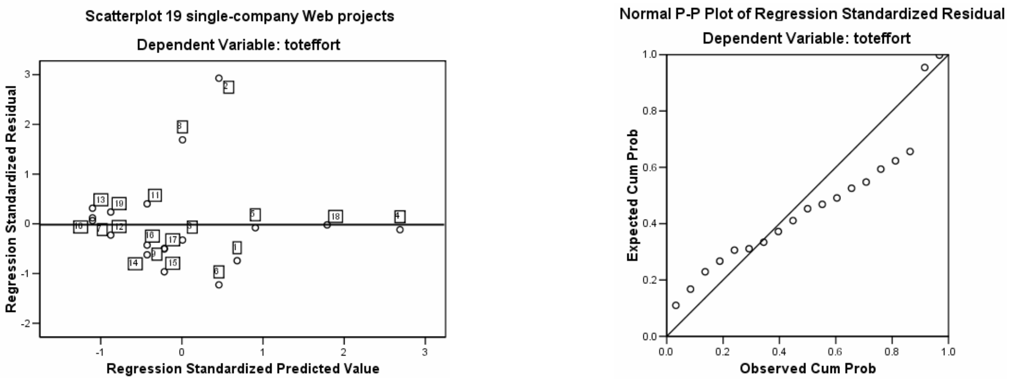
The Equation as read from the final model's output is:

$$\text{Toteffort} = -6.720 + 0.6311\text{devTeam} + 1.277\text{fotsa} \tag{4}$$

The residual plot and the P-P plot for the final single-company model are presented in Figure 2(a) and Figure 2(b) respectively. Although the P-P plot shows that data points do not completely gather closely around a straight line, the residuals are normally distributed, based on the Kolmogorov-Smirnov test (K-S test).

Table 12 - Best Fitting Model after removing the high-influence project

Independent Variables	Coefficient	Std. Error	t	p>ltl
(constant)	-6.720	2.319	-2.897	0.011
devTeam	6.311	1.340	4.708	0.000
fotsa	1.277	0.430	2.969	0.009



(a)

(b)

Figure 2 – Residual and P-P plot for best fitting single-company model

Measuring Prediction Accuracy

To assess the accuracy of the predictions for the single-company model we employed a 19-fold cross-validation to the data set, where 18 projects at a time were in the training set and one project in the validation set. This means that for 19 times, a project was omitted from the data set, and an Equation, similar to that shown by Equation 4, was calculated using the remaining 18 projects. At each time the estimated effort was calculated for the project 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 13, where we can see that the single-company model's prediction accuracy was not significantly different from both the estimate accuracy provided using expert opinion and predictions based on the median of the data set (4.12). This result was confirmed using the Wilcoxon matched-paired signed rank test on absolute residuals.

However, the estimates based on the median model were significantly worse than those provided using expert opinion. What these results suggest is that effort estimates for the single-company projects based on the single-company data will be similar when using either a regression-based cost model, or experts' estimates, or the median effort for past projects. Both mean and median absolute residuals confirm this trend.

These results do not corroborate those obtained in [14], where Mendes and Kitchenham found that the single-company model presented significantly better prediction than estimates based on expert opinion or based on the median effort.

Table 13 - Prediction accuracy statistics for the single-company data set

Prediction Accuracy	Estimates based on regression model	Estimates made by company personnel	Estimates based on median model
Mean MRE (MMRE)	51.13%	27.71%	80.20%
Median MRE (MdMRE)	51.81%	21.81%	68.16%
Pred(25)	26.32%	57.89%	15.79%
Mean absolute residual	2.49	2.06	4.72
Median absolute residual	1.76	1.15	2.12

3.3 Comparing Accuracy between the Cross-company and Single-company models

To compare the accuracy between the cross-company and single-company models we used the absolute residuals for the 20 single-company projects employed to validate the regression-based cross-company model (see Section 3.1) and the absolute residuals for each of the 19 single-company validation sets used to validate the regression-based single-company model (see Section 3.2). Their box plots are presented in Figure 2, where ResidualsCC and ResidualsSC are the residuals for the cross-company and single-company models respectively. The box plots show that the spread of the distribution for ResidualCC is much wider than that for ResidualSC. In addition, ResidualCC has at least 60% of its values greater than ResidualSC's values, indicating that residuals based on the cross-company model were much worse than residuals based on the single-company model.

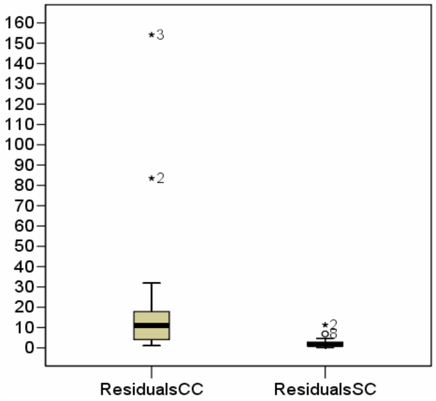


Figure 2 – Box plots for absolute residuals

Apart from using box plots, we also applied the Mann-Whitney Test for two independent samples to check if both sets of residuals came from the same population.

Results confirmed that the absolute residuals for the single-company model are significantly better (smaller) than the absolute residuals for the cross-company model ($\alpha<0.05$). These results corroborate those obtained in [14], and also corroborate findings previously published [6],[7], where similarly to the Tukutuku data set, the data was collected without using rigorous quality-assurance procedures.

4. OBTAINING EFFORT ESTIMATES USING CASE-BASED REASONING

There is no clear answer to date as to what is the best technique to employ to obtain effort estimates, for given a data set. Shepperd and Kadoda suggested that data set characteristics should have a strong influence on the choice of techniques to employ to obtain effort estimates [17]. The less "messy" the data set, i.e., small number of outliers, small amount of collinearity, strong relationship between independent and dependent variables, the higher the chances that regression analysis will give the best estimation accuracy. Conversely, very "messy" data sets should use case-based reasoning (CBR) to obtain more accurate effort estimates. The Tukutuku data set presents some level of collinearity, outliers, and a non-linear relationship between independent variables and the dependent variable for the cross-company models obtained in [14] and in this study. Thus, like [14], we also investigated the use of case-based reasoning to obtain effort estimates.

Like [14], we also used CBR-works, a commercial case-based reasoning tool, to obtain our effort estimates. Estimates were based on the average effort of the two most similar projects in the case base, identified on the basis of Euclidean distance, with no different weights for attributes or adaptation of the estimated effort.

Our results for CBR are summarised in Table 14, as follows:

CBR cross-company model provided predictions significantly worse than those for the regression-based cross-company model ($p<0.05$). In [14] no significant differences were found.

CBR single-company model provided predictions significantly worse than those for the regression-based single-company model ($p<0.05$). Our results corroborate those in [14].

CBR cross-company model presented significantly worse predictions than the CBR single-company model ($p<0.05$). Mendes and Kitchenham [14] found the opposite.

Table 14 - Summary Results for CBR and Regression models

	Predictions based on CBR		Predictions based on Regression	
Prediction statistics	**Cross-company model (CCCM)**	**Single-company model (CSCM)**	**Cross-company model (RCCM)**	**Single-company model (RSCM)**
Number of predictions	20	19	20	19
MMRE	312%	51.13%	6601.31%	349.84%
Median MRE	105%	51.81%	3981.63%	247.57%
Pred(25)	0%	26.32%	0%	15%
Mean absolute residual	21	2.49	294.97	9.25
Median absolute residual	11	1.76	98.67	10.25

5. DISCUSSION

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 single-company model.

Our first research question is addressed by the results from Sections 3.1 and 4. The accuracy of estimates obtained for the 20 single-company projects using the regression-based cross-company model (see Equation 3) does not indicate good prediction accuracy. MMRE is 312%, which is poor (25% is considered "good" [4]), and Pred(25) is extremely poor (0%, when 75% indicates a good prediction model). The same pattern is present for predictions obtained using CBR: MMRE is 6601% and Pred(25) is 0%, both extremely poor. The absolute residuals obtained using the regression-based cross-company model were significantly worse than residuals obtained using both the median effort and expert opinion. This suggests that there is no advantage to a single company that does not have past projects from which to develop their own models, to use a cross-company model to obtain effort estimates. It can rely on the expert-based estimates. Our results corroborate those by Mendes and Kitchenham [14].

To address our second research question we compared the absolute residuals for the 19 single-company projects with the single-company model (see Sections 3.2 and 4) to those obtained using 20 single-company projects with the cross-company model (see Sections 3.3 and 4). The comparison was done using the Mann-Whitney Test for two independent samples. Results for both the regression and CBR models indicated that absolute residuals for the single-company projects using the single-company model were significantly lower (better) than absolute residuals obtained for the single-company projects using a cross-company model. These results suggest that expert-based estimation could be used for estimation until it is possible for a Web company to build its own single-company model, which can be used by itself or in combination with expert-based estimations. This is even more appropriate for Web companies that develop Web applications of the same type, using the same technologies and staff [9]. Our results using regression models corroborate those by Mendes and Kitchenham [14], however differ using CBR.

Our CBR results may be explained by a higher homogeneity in the single-company projects, a factor to which CBR seems more sensitive to than other modelling techniques. Given that the single-company projects exhibit much lower variance in size and effort than the cross-company projects, the selection of an inadequate 'most-similar' project is very likely to have less damaging consequences in the prediction than it would have for the cross-company projects.

Although few studies have reported that cross-company models present similar accuracy to single-company models [1],[2],[16],[18], ours and others did not corroborate those findings [6],[7],[11],[13],[9],[14]. Such contradictory results stress the need to establish under which circumstances a company can rely on a cross-company model.

Previous studies have suggested that data collection following rigorous quality assurance procedures make it likely for cross-company models to be as accurate as single-company models [1],[9],[14],[18]. This claim is supported by [18], the only replicated study to date that used a cross-company database of software projects where rigorous quality assurance mechanisms were applied to their data collection from the start. Both original [1] and replicated [18] studies consistently showed no differences in prediction accuracy between cross-company and single-company models. The Tukutuku database does not have strict data quality assurance procedures, which may explain the recurring poor performance of cross-company models across studies.

6. CONCLUSIONS

We found that the cross-company model provided poor predictions for the single-company projects and much worse predictions than the single-company model. These results suggest that the cross-company model was not successful either at estimating effort for projects from a single company, or in comparison, to a single-company model. Mendes and Kitchenham [14] obtained the same results for models built using stepwise regression, using a different Tukutuku data set to ours. However, despite providing poor predictions for the single-company projects, Mendes and Kitchenham's case-based reasoning cross-company model did not give worse predictions than the single-company model. One possible explanation is that their cross-company data set was larger than ours, an aspect which favours CBR in general.

Given the results of the research to date, we only advise the use of cross-company models whenever data is obtained using rigorous quality control procedures.

As part of our future work we aim to add more rigorous quality control procedures for gathering data on Web projects for the

Tukutuku database, and then to replicate our study using further data.

7. ACKNOWLEDGMENTS

We would like to thank all those companies that have volunteered data on their projects to the Tukutuku database.

8. REFERENCES

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

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

[3] Christodoulou, S. P., P. A. Zafiris, T. S. Papatheodorou, WWW2000: The Developer's view and a practitioner's approach to Web Engineering, in Proceedings of Second ICSE Workshop on Web Engineering, 4 and 5 June 2000, Limerick, Ireland, 2000, 75-92.

[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, 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, 1009-1016.

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

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

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

[10] Kitchenham, B.A. and N.R. Taylor. Software cost models. ICL Technical Journal, May 1984, 73-102.

[11] 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, 2003, 2477-2487.

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

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

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

[15] Mendes, E., N. Mosley, and S. Counsell, Investigating Early Web Size Measures for Web Cost Estimation, in Proceedings of EASE'2003 Conference, Keele, April, 2003, 1-22.

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

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

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

[19] Wilcoxon, F. Individual comparisons by ranking methods. Biometrics, 1, 1945, 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.