Further Comparison of Cross-company and Within-company Effort Estimation Models for Web Applications

Emilia Mendes
Computer Science Department
The University of Auckland
Private Bag 92019, Auckland,
New Zealand
emilia@cs.auckland.ac.nz

Barbara Kitchenham
Department of Computer Science
Keele University, Staffordshire ST5 5GB, UK
barbara@cs.keele.ac.uk
and
National ICT Australia, Locked Bag 9013 Alexandria,
NSW 1435, Australia
barbara.kitchenham@nicta.com.au

Abstract

This paper extends a previous study, using data on 67 Web projects from the Tukutuku database, investigating to what extent a cross-company cost model can be successfully employed to estimate effort for projects that belong to a single company, where no projects from this company were used to build the cross-company model. Our within-company model employed data on 14 Web projects from a single Web company.

Our results were similar to those from the previous study, showing that predictions based on the within-company model were significantly more accurate than those based on the cross-company model. We also found that predictions were very poor when the within-company cost model was used to estimate effort for 53 Web projects from different companies. We analysed the data using two techniques, forward stepwise regression and case-based reasoning. We found estimates produced using stepwise regression models were better for the within company model while case-based reasoning predictions were better for the cross-company model.

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

1. Introduction

Several researchers have suggested that company-specific data sets are needed to produce accurate effort estimates (see for example [10] and [7]). However, three main problems can occur when relying on company-specific data sets [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 multi-company data sets (datasets containing data from several companies) for cost estimation and productivity benchmarking. However, the use of multi-company data sets also has problems of its own [2]

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

Furthermore, we believe there are additional problems:

- It is more difficult to ensure consistent data collection standards across many different companies than it is to ensure consistent standards within a specific company.
- It may also be difficult to be sure that the projects in a large data set used current practices, unless the data collection standards specify that submitted data must report the date of project completion, so that projects can be easily partitioned into new and old groups.
- We cannot be sure that the data set is a random sample from some defined population. In most cases companies are free to select the projects that they themselves wish to submit to the data base. This makes it difficult to be confident that models derived from cross-company data sets can generalise to other projects. The size of a dataset cannot compensate for the lack of any sampling methodology.

Five studies in Software engineering have investigated whether cross company models can be as accurate as

within company models [1],[2],[5],[6],[16]. These studies used data from two application domains: 'business' and 'space and military'. Their findings were as follows:

- Three studies found that a cross-company model gave similar prediction accuracy to that of a within-company model [1],[2],[16]. Here the data used was collected using rigorous quality-assurance procedures.
- Two studies found that a cross-company model did **not** give as accurate predications as a within-company model [5],[6]. Here the data used was collected without using rigorous quality-assurance procedures.

Recently we investigated the same issue using data on Web projects [9]. Our study employed data on 53 Web projects from the Tukutuku database [13]. In this case data was collected without a rigorous quality assurance mechanism. The data set had 13 projects from the same company (Company 1) and the remaining 40 projects from another 23 companies. This study employed four steps:

Step 1) A baseline cross-company model was fitted to the full data set. Goodness of fit statistics were calculated from the model predictions. This baseline model allowed us to identify appropriate independent variables from a large number of possible size and project metrics.

Step 2) We recalculated the baseline model omitting the Company 1 projects and used the resulting model to predict the Company 1 projects. The predictions were used to calculate accuracy statistics.

Step 3) We derived a within company model for Company 1 from scratch. We determined the prediction accuracy of estimates for this model using leave-one-out cross-validation. This allowed use to compare the prediction accuracy for the cross-company model with the prediction accuracy for the within company model.

Step 4) We also used the Company 1 data set to predict the values of the other 43 projects and constructed accuracy statistics from the predictions. This allowed us to assess how good the within-company model would be at predicting another company's projects.

This paper addresses the same issues as those discussed in Kitchenham and Mendes study [9] but uses an extended version of the Tukutuku data set containing 67 Web projects. The additional 14 projects all came from a single company (referred to as Company 2). These projects were volunteered to the Tukutuku database after the previous study on Web projects was carried out. This allowed us to replicate our previous analysis with another company.

However, we made a few modifications to our previous experimental procedure: For Step 1 we used the leave-one-out cross validation estimates rather than simple model estimates. For Step 2 we used two cross-validation models: CCM1 and CCM2. CCM1 was constructed after excluding the 14 projects from Company 2. (Note CCM1 corresponds to the baseline company model used in our previous study). Next we created a baseline model using all 67 projects and used the variables identified in the baseline model to construct CCM2 after excluding the 14 projects from Company 2.

Like [9], we used forward stepwise regression to build cost models and obtain effort estimates [8]. However, in this study, we also used case-based reasoning to construct our models to investigate if results would be consistent.

We measured prediction accuracy based on standard metrics such as MMRE and Pred(25), and also used Median MRE, Median and Mean of absolute residuals, and the Company estimates provided by some of the Web companies that volunteered data for Tukutuku. The Company estimates were based on an educated guess.

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. Many e-commerce applications fall into this category. 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 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. Finally, conclusions are given in Section 5.

2. Research Method

2.1 Data set Description

The analysis presented in this paper was based on Web projects from the Tukutuku database [13]. 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[1].

The analysis presented in this paper used data from 67 Web projects where 27 projects come from two companies (Company 1 with 13 and Company 2 with 14 projects respectively), and the remaining 40 come from another 23 companies. Each Web project in the database provided 43 variables to characterise a Web application and its development process (see Table 1).

Table 1 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	Effort considered high to develop a single function/feature[2] by one person
HEffAdpt	Ratio	Effort considered high to adapt a single function/feature[3] by one person.
hfots	Ratio	Number of reused High effort features/functions without adaptation
hfotsA	Ratio	Number of adapted High effort features/functions
hnew	Ratio	Number of new High effort features/functions
tothigh	Ratio	Total Number high effort features/functions
fots	Ratio	Low effort FOTS
fotsa	Ratio	Low effort FOTS-A
new	Ratio	Number new Low effort features/functions
totnhigh	Ratio	Total Number low effort features/functions
toteffor	Ratio	Total effort develop the Web application
accuracy	Nominal	Procedure used to record effort data

The size metrics and cost drivers employed represent early Web size metrics and cost drivers obtained from the results of a survey investigation [13], using data from 133 on-line Web forms aimed at giving quotes on Web development projects. In addition, these metrics 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 43 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.

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,

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

[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.

we asked companies how their effort data was collected (see Table 2).

Two companies used different data collection levels depending on the type of project (i.e. they used level 1 for some projects and levels 3 and 4 for other projects). Of the two companies that volunteered more than 10 projects, one used level 3 to record effort for all its 13 projects and the other used level 4 to record effort for all of its 14 projects. At least for 77.6% of Web projects in the Tukutuku database effort values were based on more than guesstimates. However, we are also aware that the use of timesheets does not guarantee 100% accuracy in the effort values recorded.

Table 2 How effort data was collected

Data Collection Method	Level	Number of Projects and Companies	
		# and % projects	# different companies
No timesheets	1	12 (17.9%)	8
Total hours worked each day or week	2	3 (4.5%)	3
Hours worked per project per day/week	3	24 (35.8%)	12
Hours worked per project task per day	4	28 (41.8%)	8

2.3 Modelling Techniques

For statistical model building we used forward stepwise analysis calculated with SPSS v.10.01.

The set of variables used for building the cost models is shown in Table 3. This is a subset of the Tukukuku data set since several variables were excluded based on the following criteria:

- Most instances of a variable were zero.
- The variable was categorical.
- The variable was related to another variable, in which case both could not be included in the same model. This was investigated using a Spearman's rank correlation analysis ($\alpha = 0.05$).

Categorical variables were excluded since we did not feel they would be valuable for the analysis and also because categorical variables with many categories (like ours) require a large number of dummy variables which rapidly reduce the degrees of freedom for analysis.

The data set we employed has 5 projects representing Web hypermedia applications and 62 projects representing Web software applications. Since 93% of projects belonged to the same type we did not include Web application type in our analyses.

Whenever variables were highly skewed they were transformed to a natural logarithmic scale to approximate a normal distribution [12]. 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.

The dependent variable *toteffor* was used as the dependent variable when fitting the best within-company model, however *lntoteff* was the one employed as dependent variable when fitting both cross-company models.

Table 3. Variables used in the stepwise regression

Variable	Meaning
lntoteff	Natural log. of the total effort to develop a Web application.
toteffor	Total effort to develop a Web application.
nlang	Number of different languages used on the project
devteam	The number of people who worked on the project
teamexp	Average team experience with the development language(s) employed
lnnewwp	Natural log. of (1+number of new Web pages)
lnimgnew	Natural log. of (1+number of new images in the applications)
lnimglib	Natural log. of (1+total number of images reused from a library)
lnimg3p	Natural log. of(1+total number of images developed by a third party)
hfotsa	Total number of adapted high effort functions.
lntoth	Natural log. of (1+total number of high effort functions).
fotsa	Total number of adapted low effort functions.
totnhigh	total number of low effort functions
Natural log. = Natural logarithm	

2.4 Analysis Methods

To verify the **stability** of each cost model built we used the following steps [9]:

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

S2. Calculate Cook's distance values [4] for all projects to identify influential data points. Any projects with distances higher than $4/n$, where n represents the total number of projects, were considered to have high influence on the results. When there were projects with high influence, the stability of the model was tested by removing these projects, and observing the effect their removal had on the model. If the model parameter variables remained stable, the high influence projects were retained in the data analysis.

The prediction **accuracy** of models was checked either by using the 'omit one project at a time' procedure (leave one out cross-validation), or by omitting a group of projects and predicting the effort for the group of omitted projects. In both situations, the rationale is 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 in [9], which are the MMRE, Median MRE, Pred(25), the

median and mean absolute residuals [11], and the Companies estimates.

Table 4- Summary of Companies estimates

First Study [9]		
Data set combination	Accuracy rates	Type
Cross-company data set (53 projects)	62.5%	AEG
Cross-company data set (without 13 projects from Company 1)	68.3%	AEG
13 Company 1 projects	10%	AEG
This paper's study		
Cross-company data set (67 projects)	61.1%	AEG/CE
Cross-company data set (without 14 projects from Company 2)	62.5%	AEG
14 Company 2 projects	47%	CE

A summary of the accuracy rates achieved by the Web companies is shown in Table 4. The type column identifies the basis of the estimate where AEG is an average educated guess obtained after the data was submitted to the Tukutuku data base, or and CE is a contemporary estimate provided by the company when the data was collected for submission to the Tukutuku database. All company estimates were underestimates.

3. Results

3.1 Model Construction

The first model to be built was a cross-company model using the full data set of 67 Web projects (see Table 5). Its adjusted R^2 was 0.67.

Table 5 Best Fitting Model to calculate lntoteff

Independent Variables	Coeff.	Std. Error	t	p>|t|	95% Confidence Interval
(constant)	2.154	0.260	8.281	0.000	1.634 – 2.674
lnnewwp	0.435	0.061	7.184	0.000	0.314 – 0.556
lntoth	0.671	0.160	4.198	0.000	0.352 – 0.991
devteam	0.239	0.083	2.876	0.005	0.073 – 0.406
Coeff. - Coefficient					

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

$$\ln(\text{toteffor}) = 2.154 + 0.435 \times \ln(\text{newWP}+1) + 0.671 \times \ln(\text{tothigh}+1) + 0.239 \times \text{devteam} \quad (1)$$

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

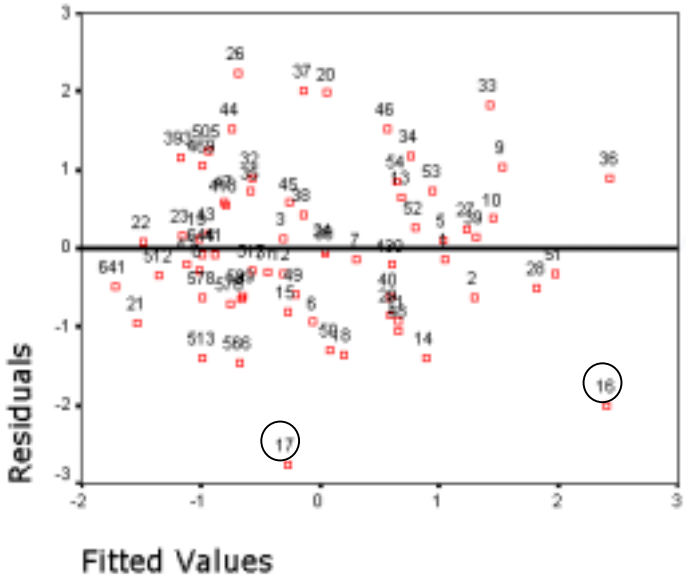
$$\text{toteffor} = 8.619 \times (\text{newWP}+1)^{0.435} \times (\text{tothigh}+1)^{0.671} \times e^{0.239 \times \text{devteam}} \quad (2)$$

Equation 2 is very similar to that obtained in [9] (adjusted $R^2 = 0.597$), which was:

$$\text{toteffor} = 8.425 \times (\text{Webpages})^{0.433} \times (\text{tothigh}+1)^{0.632} \times e^{0.235 \times \text{devteam}} \quad (3)$$

Note. In the absence of any reuse Webpages=New Webpages.

Figure 1 – Residual plot for best fitting model (cases are indicated by the project id)



3.2 Model Evaluation

Step 1: Testing the Residuals

The residual plot (see Figure 1) shows that projects 16, and 17 have large residuals and their effort is overestimated. However, Figure 1 shows no other pattern in the residuals.

Step 2: Detecting Influential Observations

There were four projects with Cook's distances greater than 4/67, for which key variable values are presented in Table 6.

Table 6 - Projects Key Variable Values

Project	Residual	toteffor	newWP	tothigh	dv
16	-1.792	363	100	10	8
17	-2.455	6	12	2	1
20	1.763	30	7	0	1
32	1.626	105	16	0	2
Pro – project te – toteffor Res. – residual dv - devteam					

To check the model's stability, a new model was generated without the four projects that presented high Cook's distance, giving an adjusted R^2 of 0.732 (see Table 7). In the new model all the independent variables remain significant and the parameters have similar values to those in the previous model. Therefore, we used the model based on the full data set (i.e. we did not remove the high influence data points).

Table 7 New model to calculate lntoteff without high influence projects

Independent Variables	Coef.	Std. Error	t	p>|t|	[95% Confidence Interval]
(constant)	2.448	0.191	12.835	0.000	2.066 – 2.829
lnnewwp	0.410	0.054	7.606	0.000	0.302 – 0.518
lntoth	0.856	0.137	6.248	0.000	0.582 – 1.130
devteam	0.103	0.049	2.097	0.040	0.005 – 0.201
Coef. – coefficient					

3.3 Measuring Prediction Accuracy

To assess the accuracy of the predictions for the cross-company model a "leave-one-out" cross-validation was applied to the data set, using the raw scale. This means that for each of the 67 projects, one at a time was omitted from the data set, and an equation, similar to that shown by equation 1, was calculated using the remaining 66 projects. This equation was then transformed back to the raw scale, giving an equation similar to that shown by equation 2. Then 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 8, where we can see that the model's prediction accuracy was worse than the mean estimate accuracy provided using expert opinion, which was 61.1% (underestimate). Its accuracy was not significantly better than predictions based on the median of the data set (median = 90) using the Wilcoxon matched-paired signed rank test.

The median we obtained was smaller than the median obtained in [9], which was 103.5. This was caused by the insertion of 14 Web projects from Company 2 where 13 had effort values below 103.5, thus shifting the median to a smaller value. Company 2's projects are small, with minimum effort of 7 person hours, maximum effort of 178 person hours, average effort being 45 person hours, and median equal to 25.5 person hours.

The differences between values obtained for medians and means, for the MREs and absolute residuals suggests that the data set contains several outliers.

Table 8 Prediction accuracy statistics for the total data set

Prediction Accuracy	Estimates based on regression model
MMRE	99%
Median MRE	70%
Pred(25)	9%
Mean absolute residual	374.9
Median absolute residual	59.6
Prediction accuracy	Estimates made by company personnel
Average Underestimate	61.1%
Prediction Accuracy	Estimates based on median model
MMRE	194%

Our results are slightly different from those presented in [9], since our cost model did not show better accuracy than the median model. However, like [9], our model presented worse accuracy than the mean estimate accuracy based on expert opinion.

3.4 Comparison of Cross-company and Within-company Models

We calculated and compared the prediction accuracy for the 14 projects from Company 2 derived from three different estimation models

- A cross-company model (CCM2) based on the 53 other projects but using the variables identified in equation 1).
- A cross-company model (CCM1) built without the 14 projects from Company 2. CCM1 was also used as the baseline model in our previous study.
- A within-company model (WCM) built from scratch using projects from Company 2 with a leave-one-out validation process.

To determine the cross-company model CCM2 we rcalculated the model presented in Section 3.1, using the same variables shown in Table 5, after excluding all 14 projects from Company 2. The model is reported in Table 9. Its adjusted R^2 was 0.63.

Table 9 Best Fitting Model to calculate lntoteff after excluding 14 projects from Company 2

Independent Variables	Coeff.	Std. Error	t	p>|t|	[95% Confidence Interval]
(constant)	2.300	0.340	6.755	0.000	1.615 – 2.984
lnnewwp	0.411	0.084	4.894	0.000	0.242 – 0.580
lntoth	0.699	0.171	4.083	0.000	0.355 – 1.043
devteam	0.221	0.092	2.412	0.020	0.037 – 0.405
Coeff. – Coefficient					

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

$$\ln(\text{toteffor}) = 2.3 + 0.411 \times \ln(\text{newWP}+1) + 0.699 \times \ln(\text{tothigh}+1) + 0.221 \times \text{devteam} \quad (4)$$

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

$$\text{toteffor} = 9.974 \times (\text{newWP}+1)^{0.411} \times (\text{tothigh}+1)^{0.699} \times e^{0.221 \times \text{devteam}} \quad (5)$$

Using equation (5) we estimated the effort for the 14 projects from Company 2 projects and calculated the MRE and absolute residuals. The prediction accuracy (see Table 10) was significantly, better than the general cross-company model presented in Section 3.1 based on the Mann-Whitney test of the absolute residuals ($p<0.05$).

In addition, the predictions for the 14 projects were compared with a prediction based on the median of the total effort for the remaining 53 projects, which is 41.5 person hours (see Table 10). Their paired absolute residuals were compared using the Wilcoxon signed rank test and no significant differences were found, meaning that the predictions based on the cross-company model were not significantly better than those based on a median model. The cross-company model gave worse predictions than the estimates provided by the Company 2 estimators (i.e. an underestimate of 47%).

Table 10 Prediction accuracy statistics for new cross-company model (CCM2) and median model

Prediction statistics	Predictions based on regression model	Predictions based median effort model
MMRE	93%	143%
Median MRE	61%	70%
Pred(25)	7.1%	7.1%
Mean absolute residual	25.33	33.1
Median absolute residual	21.95	26.9

Using CCM1 cross-company model (equation 6), we calculated MRE and absolute residuals for each of the 14 Company 2 projects.

The predictions based on the cross-company model CCM1 not significantly different than those based on a median model (see Table 11) using the Wilcoxon signed rank test. The cross-company model gave worse predictions than the Company 2 estimates (underestimate of 47%). In addition, apart from Pred(25) it has worse accuracy statistics than the CCM2.

$$\text{toteffor} = 8.68 \times (\text{Webpages})^{0.456} \times (\text{tothigh}+1)^{0.501} \times e^{0.241 \times \text{devteam}} \quad (6)$$

Table 11 Prediction accuracy statistics for cross-company model CCM1 and median model

Prediction statistics	Predictions based on regression model	Predictions based on median effort
MMRE	230%	428%
Median MRE	151%	304%
Pred(25)	14.3%	7.1%
Mean absolute residual	55.4	69.4
Median absolute residual	54.4	77.9

Neither CCM2 nor CCM1 cross-company gave better prediction accuracy for Company 2 project than their corresponding median models, using Company 2 projects.

Most variables selected for CCM2 and CCM1 were the same, except for Webpages and newWP, which will only be equivalent when there is no page reuse.

The best fitting model for the 14 projects from Company 2 (WCM) is presented in Table 12. Its adjusted R^2 is 0.95.

Some of the variables selected by this model are different from those selected by either CCM1 or CCM2 cross-company models, however all have in common the selection of high effort features (either total number of high features or number of high effort features adapted).

Table 12 Best fitting model for calculating toteffor using 14 projects from Company 2

Independent Variables	Coeff.	Std. Error	t	p>|t|	[95% Confidence Interval]
(constant)	11.621	3.973	2.925	0.014	2.876 – 20.366
hfotsa	37.389	3.016	12.397	0.000	30.751 – 44.027
fotsa	3.189	1.103	2.891	0.015	0.761 – 5.617
Coeff. – Coefficient					

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

$$\text{toteffor} = 11.621 + 37.389 \times \text{hfotsa} + 3.189 \times \text{fotsa} \quad (7)$$

We used a "leave-one-out" cross-validation to assess the predictive accuracy of the within company model

model. In addition, the predictions for the 14 projects were compared with a prediction based on the median of effort for the same 14 projects, which is 25.5 person hours. The accuracy statistics are shown in Table 13.

The predictions based on the within-company model were significantly better than those based on the simple median model ($\alpha<0.05$) using the Wilcoxon signed rank test on the absolute residuals. The within company model also gave better prediction accuracy than the Company estimate, which was 47% (underestimate).

Table 13 Prediction accuracy statistics for within-company model and median model

Prediction statistics	Predictions based on regression model	Predictions based on median effort
MMRE	38%	82%
Median MRE	38%	61%
Pred(25)	28.6%	14.3%
Mean absolute residual	11.2	30.3
Median absolute residual	8.36	15.8

Finally, we compared the predictive accuracy of the within company model with the two cross-company modelsCCM1 and CCM2 using the Wilcoxon signed rank test for the paired absolute residuals. Results confirm that the absolute residuals for the within-company model are significantly better (smaller) than the absolute residuals for CCM1 and CCM2 ($\alpha<0.05$).

This is a similar result to that obtained in [9], and also corroborates findings previously published [5],[6], where similarly to the Tukutuku data set, the data was collected without using rigorous quality-assurance procedures.

3.5 Applying the Within-company Model to the 53 Web projects

To assess whether a within-company model can be useful to predict effort for projects from other companies, we calculated estimated effort for each of the projects in the Tukutuku data set (excluding the 14 projects from Company 2) using the within-company model from equation 7.

Prediction accuracy statistics and absolute residuals were obtained (see Table 14), and all suggest that estimations for the 53 projects based on the within-company model are poor. The prediction accuracy statistics are slightly worse than those obtained for the cross company model based on the full data set (see Table 8)

Table 14 Prediction accuracy statistics for 53 projects based on within-company model

Prediction Statistics	Prediction for other companies based on Company 1 model
MMRE	94%
Median MRE	89%
Pred(25)	3.8%
Mean absolute deviation	395.1
Median absolute deviation	88.4

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 Kadoga suggested that data set characteristics should have a strong influence on the choice of techniques to employ to obtain effort estimates [15]. The less “messy” the data set, i.e., small number of outliers, small amount of collinearity, strong relationship between predictors and response 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 study presented here has used forward stepwise regression since this was the technique employed in [9]. However, the Tukutuku data set presents some level of collinearity, outliers, and a non-linear relationship between predictors and response for the cross-company models we obtained. Therefore, as there is some level of “messiness” in our data set, we also investigated the use of case-based reasoning to obtain effort estimates.

We used a commercial case-based reasoning tool (CBR-works) to obtain our effort estimates. Estimates were based on the average effort of the two closest analogues identified on the basis of Euclidean distance, with no weights or adaptation. This choice was based on previous work where this was the combination that provided the best effort prediction accuracy [14]. The set of variables employed was the same one presented in Table 3.

Table 15 Summary Results for CBR and Regression models

	Predictions based on CBR				Predictions based on regression			
Prediction statistics	Whole data set	Company 2 using other project data	Company 2 using Company 2 data	Other projects using Company 2 data	Whole dataset	Company 2 using other project data (CCM1)	Company 2 using Company 2 data	Other projects using Company 2 data
Number of predictions	67	14	14	53	67	14	14	53
MMRE	100%	176%	236%	113%	99%	230%	38%	94%
Median MRE	45%	93%	136%	82%	70%	151%	38%	89%
Pred(25)	25.4%	14.3%	7.1%	5.7%	9%	14.3%	28.6%	3.8%
Mean absolute residual	156.2	35.4	46.9	372.01	374.9	55.4	11.2	395.1
Median absolute residual	41.5	31.3	45.9	59.5	61.1	54.4	8.4	88.4

Our results for CBR are summarised in Table 15. We found that:

- CBR predictions for the whole data set were significantly better than the regression model predictions ($p<0.05$)
- CBR predictions for Company 2 projects using the other company data were not significantly different from the regression model predictions.
- CBR predictions for Company 2 using Company 2 data were significantly worse than the regression model predictions ($p<0.05$)
- CBR predictions for the other projects using Company 2 data were not significantly different from the regression model predictions.

Thus, CBR was better for predictions across the large heterogeneous data set, but regression was better for within-company predictions.

5. Conclusions

For our data set, we found that cross-company prediction models gave much worse predictions for a specific company than within company predictions for two different within company datasets. In this study, predictions based on a model that included no information about Company 2 (i.e.CCM1) were worse than the existing accuracy of Company 2 estimates. However, predictions based on a within company model were significantly better than the cross-company model and slightly better also than the existing accuracy of Company 2 estimates (MMRE=38% compared with an underestimate of 47%).

Although some studies report cross-company models having comparable accuracy to within-company models ([1][2][16]), our study and others have reported contradictory results ([5][6]). It is important therefore to determine under what circumstances a company can place reliance on a cross-company model.

One systematic difference between the studies appears to be the quality controls applied to data collection. Another factor that could have influenced the results obtained in the different studies is the process used to construct the various models. We used two cross company models. One model (CCM1) was completely independent of Company 2 data – i.e. was based solely on the data from the other 53 projects. The other model (CCM2) was based on an analysis of the full data set including Company 2 data, where we used the variables selected by analysing the full data set and recalibrated the parameters after removing the company 2 data. CCM2 gave much better estimates than CCM1. This means the way in which the models are constructed can affect the results. Furthermore we estimated our within-company model from scratch rather than simply using the variables selected in the cross-company model and recalibrating the parameters based on the within company data. Other researchers could have made a different choice. Unfortunately only one study [16] makes clear its methodology with respect to construction of the cross-company and within company models so we cannot assess the impact of the model construction process.

Given the results of the research to date, we cannot recommend the use of cross-company models, unless model users are sure that the data has been collected using stringent quality control procedures and the users of the model have already contributed some project data to the data set used to construct the cross-company model. Furthermore, our results strongly support early studies (e.g. [10] and [7]) that suggested models built on a specific data set could not be used on other projects without calibration i.e. within company models do not travel. With respect to model construction, our results suggest that CBR may be useful when analysing cross-company datasets, but in our case it did not work well on the small within company dataset.

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

[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 7th International Software Metrics Symposium, London, IEEE Computer Society Press, 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. A procedure for analysing unbalanced data sets. IEEE Trans. Software Engineering. 24(4), 1998, pp 278-301.

[9] Kitchenham, B.A., and E. Mendes. A Comparison of Cross-company and Within-company Effort Estimation Models for Web Applications, Proceedings 8th International Conference on Empirical Assessment in Software Engineering EASE 2004, Computer Society Press, 2004, pp 47-55.

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

[11] 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.

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

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

[14] 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.

[15] 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.

[16] Wieczorek, I. and M. Ruhe. How valuable is company-specific data compared to multi-company data for software cost estimation? . Proceedings 8th International Software Metrics Symposium, Ottawa, IEEE Computer Society Press, June 2002, pp 237-246.

[17] 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.