# Effort Estimation: How Valuable is it for a Web Company to Use a Cross-company Data Set, Compared to Using Its Own Single-company Data Set?

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

Sergio Di Martino, Filomena Ferrucci, Carmine Gravino
Università di Salerno
Via Ponte Don Melillo, I-84084 Fisciano (SA)
Italy
0039 089963374
{sdimartino,fferrucci,gravino}@unisa.it

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

This paper therefore replicates a previous study by investigating how successful a cross-company effort 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 effort model. Our single-company data set had data on 15 Web projects from a single company and our cross-company data set had data on 68 Web projects from 25 different companies. The effort estimates used in our analysis were obtained by means of two effort estimation techniques, namely forward stepwise regression and case-based reasoning.

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.

## Categories and Subject Descriptors
D.2.9 [**Management**]: Cost estimation, Productivity, Time estimation; D.2.8 [**Metrics**]: Process metrics, Product metrics.

## General Terms
Management, Measurement, Economics, Experimentation.

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

## 1. INTRODUCTION
When planning a software project, early estimation of development effort/cost is a critical management activity. It aims at predicting an accurate effort estimate and using this information to allocate resources adequately. This activity is crucial for the competitiveness of a software company. In the context of Software- and Web-engineering, many techniques have been applied to estimate the effort necessary to develop a new project. These techniques use data from past projects, characterized by attributes that are related to effort, and the actual effort used to develop a project, to estimate effort for a new project under development (see e.g., [1],[2],[4],[7]). The techniques used and the characteristics of the data sets play a role in the accuracy of the predictions that are obtained [24].

Currently, one of the issues faced by Software and Web companies is if it is worthwhile to obtain estimates for their new projects based on cross-company data sets, i.e. data sets that contain project data volunteered by several companies. The use of a cross-company data set seems particularly useful for companies that do not have their own data on past projects from which to obtain their estimates, or that have data on projects developed in different application domains and/or technologies.

Previous studies in Software Engineering have suggested that a company needs its own data set (thus a single-company one) to produce more accurate effort estimates (e.g. [10],[14]). 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 data sets for effort estimation and productivity benchmarking. However, the use of cross-company data sets has problems of its own [2],[20]:

i) Care is necessary as data needs to be collected in a consistent manner by means of a uniform data collection control across different companies.
ii) Differences in processes and practices may result in trends that may differ significantly across companies.
iii) Projects should be partitioned (e.g. according to their completion dates) to identify those that used current development practices from those that did not.


*WWW 2007*, May 8-12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

iv) Project data should represent a random sample representative of a well-defined population (it is not sufficient that the data set is large).

To date, eleven studies (nine in Software engineering and two in Web engineering) have investigated if estimates obtained with cross-company data sets can be as accurate as the ones obtained with single-company data sets [1],[2],[8],[9],[13],[16],[17], [19],[20],[22],[25]. Here we only focus on the two studies (S1 [13] and S2 [20]) that used data on Web projects; however, a detailed comparison of all previous studies is provided in [15].

S1 and S2 found that using a single-company data set gave significantly better predictions than using a cross-company data set (see Table 1). Both studies employed data from the Tukutuku database [21]. S2 was an extended analysis (not a replication) of S1, and the cross-company data used in S2 combined the cross- and single-company data used in S1. However, both studies were independent because the single-company data used in S2 was not part of the Tukutuku data set used in S1. It was collected from a single company some time after the first set of data had been collected and analysed.

To our knowledge, the last published study that compared effort prediction accuracy between cross- and single-company data sets for Web projects was S2, published in 2004. Since then another 83 Web projects have been volunteered to Tukutuku database, which may have an impact on the results observed previously.

**Table 1. Comparison of previous studies using data on Web projects**

| | Study S1 [13] | Study S2 [20] |
|---|---|---|
| Database | Tukutuku | Tukutuku |
| Application domain(s) | Mainly corporate, Information, promotional, e-commerce | Mainly corporate, Information, promotional e-commerce |
| Type of application | Web-based | Web-based |
| Countries | worldwide | Worldwide |
| Total Dataset size | 53 | 67 |
| Single company | 13 | 14 |
| Cross-company model showed similar accuracy to Single-company model | No | No |

Since it is widely recognized that replicated studies are fundamental to establish the validity and generalisability of results [26], in the present paper we replicate S2 [20], using Web project data volunteered after that study was carried out.

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

i) How successful is a cross-company data set at estimating effort for projects from a single company?
ii) How successful is the use of a cross-company data set, compared to a single-company data set?

Both research questions must be investigated in combination for the following reasons: i) obtaining results where the use of a cross-company data set provides good predictions accuracy for single-company projects is not enough to say that the use of a cross-company data set is successful; ii) the use of a cross-company data set also needs to provide prediction accuracy not significantly worse than that provided by the single-company data set in order to be considered successful.

To address these questions, we employ the new 83 Web projects from the Tukutuku database, where 15 come from a single company, and 68 come from other 25 companies. Like S2, we used forward stepwise regression and case-based reasoning to obtain effort estimates. These techniques have been widely and successfully employed for effort estimation both in Software- and Web-engineering (see e.g.: [1],[2],[7],[8],[11],[13],[20],[23],[24]).

The remainder of the paper is organised as follows: Section 2 describes the research method employed in this study. Results using forward stepwise regression 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 data coming from 83 Web projects of the Tukutuku database [21], 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 includes Web hypermedia systems and Web applications [3]. The former are 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. 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 and are characterized by functionality affecting the state of the underlying business logic. Web applications usually include tools suited to handle persistent data, such as local file system, (remote) databases, or Web Services. Typical developers are young programmers fresh from a Computer Science or Software Engineering degree, managed by more senior staff.

The Tukutuku database has data on 150 projects where:

- Projects come from 10 different countries, mainly New Zealand (56%), Brazil (12.7%), Italy (10%), Spain (8%), United States (4.7%), England (2.7%), and Canada (2%).
- Project types are new developments (56%) or enhancement projects (44%).
- The applications are mainly Legacy integration (27%), Intranet 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%).

Each Web project in the database was characterized by 25 variables, related to the application and its development process (see Table 2). These size measures and cost drivers have been obtained from the results of a survey investigation [21], 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.

**Table 2. Variables for the Tukutuku database**

| Variable Name | Scale | Description |
|---|---|---|
| COMPANY DATA | | |
| Country | Categorical | Country company belongs to. |
| Established | Ordinal | Year when company was established. |
| nPeopleWD | Ratio | Number of people who work on Web design and development. |
| PROJECT DATA | | |
| TypeProj | Categorical | Type of project (new or enhancement). |
| nLang | Ratio | Number of different development languages used |
| DocProc | Categorical | If project followed defined and documented process. |
| ProImpr | Categorical | If project team involved in a process improvement programme. |
| Metrics | Categorical | If project team part of a software metrics programme. |
| DevTeam | Ratio | Size of a project's development team. |
| TeamExp | Ratio | Average team experience with the development language(s) employed. |
| TotEffort | Ratio | Actual total effort in person hours used to develop an application. |
| EstEffort | Ratio | Estimated total effort in person hours to develop an application. |
| Accuracy | Categorical | Procedure used to record effort data. |
| WEB APPLICATION | | |
| TypeApp | Categorical | Type of Web application developed. |
| TotWP | Ratio | Total number of Web pages (new and reused). |
| NewWP | Ratio | Total number of new Web pages. |
| TotImg | Ratio | Total number of images (new and reused). |
| ImgNew | Ratio | Total number of new images created. |
| Fots | Ratio | Number of features reused without any 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 |
| FotsA | Ratio | Number of reused low-effort features adapted. |
| New | Ratio | Number of new low-effort features/functions. |
| TotNHigh | Ratio | Total number of low-effort features/functions |

Consequently it is our belief that the 25 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.

Within the context of the Tukutuku project, a new high-effort feature/function employs at least 15 hours to be developed by one experienced developer, and a high-effort adapted feature/function employs at least 4 hours to be adapted by one experienced developer. These values are based on collected data.

As for data quality, 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 |
|---|---|---|
| Hours worked per project task per day | 93 | 62 |
| Hours worked per project per day/week | 32 | 21.3 |
| Total hours worked each day or week | 13 | 8.7 |
| No timesheets (guesstimates) | 12 | 8 |

At least for 83% of Web projects in the Tukutuku database effort values were based on more than guesstimates. In relation to the 83 projects used in this study, 85% of the 68 cross-company projects and 100% of the 15 single-company projects are also more than guesstimates.

Similar to S2 [20], we excluded from our analysis some 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 [20] 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 final set of variables used in S2 [20] and by ourselves, while Table 5 contains the corresponding summary statistics.

**Table 4. Variables used in the studies**

| Variables | S2 (same for CC and SC) | Our study | |
|---|---|---|---|
| | | SC | CC |
| nLang | ✓ | ✓ | ✓ |
| DevTeam | ✓ | ✓ | ✓ |
| TeamExp | ✓ | ✓ | ✓ |
| TotWP | ✓ | ✓ | ✓ |
| ImgNew | ✓ | ✓ | ✓ |
| Fots | | ✓ | |
| HFotsA | ✓ | | |
| Hnew | | ✓ | |
| TotHigh | ✓ | ✓ | |
| New | | ✓ | |
| FotsA | ✓ | | |
| TotNHigh | ✓ | ✓ | ✓ |
| TotEffort | ✓ | ✓ | ✓ |
| EstEffort | ✓ | | |

CC= Cross-Company dataset; SC= Single-Company dataset

**Table 5. Summary statistics for variables**

| Single-company data set – 15 projects | | | | | |
|---|---|---|---|---|---|
| Variables | Mean | Median | Std. Dev. | Min. | Max. |
| nLang | 6.27 | 6 | 0.88 | 5 | 8 |
| DevTeam | 6.20 | 6 | 0.41 | 6 | 7 |
| TeamExp | 1.87 | 1 | 1.46 | 1 | 6 |
| TotWP | 84.13 | 74 | 40.56 | 31 | 161 |
| NewImg | 18.00 | 0 | 28.16 | 0 | 92 |
| Fots | 4.87 | 5 | 4.76 | 0 | 15 |
| Hnew | 15.60 | 14 | 6.09 | 7 | 27 |
| TotHigh | 15.73 | 16 | 6.08 | 7 | 27 |
| New | 6.60 | 5 | 3.40 | 3 | 13 |
| TotNHigh | 7.07 | 6 | 3.35 | 3 | 14 |
| TotEff | 2,677.87 | 2,792 | 827.11 | 1,176 | 3,712 |
| **Cross-company data set – 68 projects** | | | | | |
| nLang | 3.57 | 3 | 1.54 | 1 | 8 |
| DevTeam | 2.68 | 2 | 3.16 | 1 | 23 |
| TeamExp | 3.70 | 2 | 2.27 | 1 | 10 |
| TotWP | 37.44 | 19 | 47.14 | 1 | 200 |
| NewImg | 34.68 | 1.50 | 135.69 | 0 | 1,000 |
| TotNHhigh | 5 | 4 | 6.27 | 0 | 35 |
| TotEff | 321.33 | 30.15 | 800.42 | 1.10 | 3644 |
| EstEff | 268.33 | 21.50 | 1,337.31 | 1 | 10,020 |

Table 5 suggests that there are clear differences between the single- and cross-company projects. Single-company projects used twice the number of languages as the cross-company projects and three times the average number of people. However, cross-company developers presented, on average, twice the experience of the single-company developers.

The cross-company applications are smaller, in number of Web pages, than those of the single-company applications. And, therefore, the effort spent on cross-company projects is also smaller than that spent on single-company projects. These differences are not sufficient to suggest that the cross-company data cannot be useful to estimate effort for single company projects, as will be explained in Section 5. Both data sets however presented similar number of low-effort features/functions.

## 2.2 Modelling Techniques

Like S2 [20], the techniques used to obtain effort estimates 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 $\alpha = 0.05$.

### *2.2.1 Forward stepwise regression*

Stepwise regression [18] is a statistical technique whereby a prediction model (Equation) is built, and represents the relationship between independent (e.g. number of Web pages) and dependent variables (e.g. total Effort). 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).

In our study we employed the variables shown in Table 4. It is worth pointing out that 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 [18] (e.g. residuals should be independent and normally distributed; relationship between dependent and independent variables should be linear). The transformation employed was to take the natural log (ln), which makes larger values smaller and brings the data values closer to each other [18]. 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 contains the variables we considered in forward stepwise regression. All the cross-company variables needed to be transformed since they were not normally distributed; three of the single-company variables also needed to be transformed (*DevTeam*, *TeamExp*, and *TotEff*).

The variable *LTotEff* was used as the dependent variable when building the single- and cross-company models.

**Table 6. Variables used in the stepwise regression**

| Single-Company data set | Cross-Company data set |
|---|---|
| nLang | LnLang |
| LDevTeam | LDevTeam |
| LTeamExp | LTeamExp |
| TotWP | LTotWP |
| NewImg | LNewImg |
| Fots | |
| Hnew | |
| TotHigh | |
| New | |
| TotNHigh | LTotNHhigh |
| LTotEff | LTotEff |

*LVarname* = variable obtained by applying log transformation to the variable *Varname*

To verify the stability of each effort model built using forward stepwise regression, the following steps were employed [13]:

- Use of a residual plot showing residuals vs. fitted values to investigate if the residuals are randomly 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 [18]. 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.

### *2.2.2 Case Based Reasoning*

CBR is a branch of Artificial Intelligence where knowledge of similar past cases is used to solve new cases [24]. Within the context of our investigation, the idea behind the use of the CBR technique is to predict the effort of a new project by considering similar projects previously developed. In particular, completed

projects are characterized in terms of a set of *p* features (e.g. number of Web pages) and form the case base. The new project is also characterized in terms of the same *p* attributes and it is referred as the target case. Then, the similarity between the target case and the other cases in the *p*-dimensional feature space is measured, and the most similar cases are used, possibly with adaptations to obtain a prediction for the target case. To apply the method, we have to select: the relevant project features, the appropriate similarity function, the number of analogies to select the similar projects to consider for estimation, and the analogy adaptation strategy for generating the estimation. It is worth to point out that the selection of the similarity function and the number of analogies are crucial decisions. Like [20], 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).

### *2.2.3 Prediction accuracy*

The accuracy of the effort estimates obtained using stepwise regression and case-based reasoning was assessed by exploiting de facto standard accuracy measures, such as the mean Magnitude of Relative Error (MMRE), median MRE, and Prediction at 25% (Pred(25)) [4].

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:

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

where e represents actual effort and ê estimated effort. The difference between MMRE and MdMRE is that the former is sensitive to predictions containing extreme MRE values.

We also used the mean and median of absolute residuals, where residuals are calculated as actual effort – estimated effort.

## 2.3 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, by exploiting the data set, the modelling techniques, and the evaluation criteria described in the previous two Sections. These were the same steps used in [20]. Both questions are also presented to provide a context for each set of steps.

*Question 1: How successful is a cross-company data set 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 since when using CBR no explicit model is built.
2) If the model uses transformed variables, convert 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 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 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 company uses a cross-company data set to estimate effort for its new projects.

*Question 2: How successful is a cross-company data set, compared to a single-company data set, for effort estimation?*

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 since with CBR no explicit model is built.
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 *n* times, each time removing a different project.
3) The overall model accuracy is aggregated across the *n* validation sets. Same for CBR.
4) Compare the accuracy obtained in Step 3 to that obtained for the cross-company data set. Same for CBR.

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

# 3. OBTAINING EFFORT ESTIMATES USING FORWARD STEPWISE REGRESSION

## 3.1 Results Based on Cross-company Data

The best cross-company regression model is described in . Its adjusted $R^2$ is 0.788, thus explaining 78.8% of the variation in effort, which is quite remarkable for a cross-company model. This model is much better than that built in [20], which provided an adjusted $R^2$ of 0.63. Both models are exponential.

**Table 7. Best Fitting Model to calculate LTotEff**

| Independent Variables | Coefficient | Std. Error | t | p>\|t\| |
|---|---|---|---|---|
| (constant) | -2.461 | .438 | -5.624 | .000 |
| LTotWP | .576 | .127 | 4.539 | .000 |
| Lnlang | 1.450 | .245 | 5.915 | .000 |
| LNewImg | .365 | .100 | 3.663 | .001 |
| LTeamExp | 1.373 | .259 | 5.294 | .000 |
| LDevTeam | 1.027 | .279 | 3.678 | .001 |

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

$$LTotEff = -2.461 + 0.576LTotWP + 1.450Lnlang + 0.365LnewImg + 1.373LTeamExp + 1.027LDevTeam \quad (2)$$

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

$$TotEff = 0.085 \times TotWP^{0.576} \times nlang^{1.450} \times newImg^{0.365} \times TeamExp^{1.373} \times DevTeam^{1.027} \quad (3)$$

Finally, only one of the variables selected by this model was also one of those selected in [20] that included *LnewWP*, *DevTeam*, and *LTotHigh*.

### *3.1.1 Checking the Model*

The residual plot for the 68 projects showed that 8 projects seemed to have large residuals. This trend was also confirmed using Cook's distance, where these eight projects presented a Cook's distance greater than 4/68. To check the model's stability, a new model was generated without those eight projects. In the new model the independent variables remained significant and the adjusted $R^2$ improved from 0.662 to 0.788; however the coefficients did not present similar values to those in the previous model, therefore, the eight high influence data points were removed from further analysis.

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.

### *3.1.2 Measuring Prediction Accuracy*

To assess the accuracy of the predictions for the cross-company model we used as validation set the 15 projects from the single-company data set.

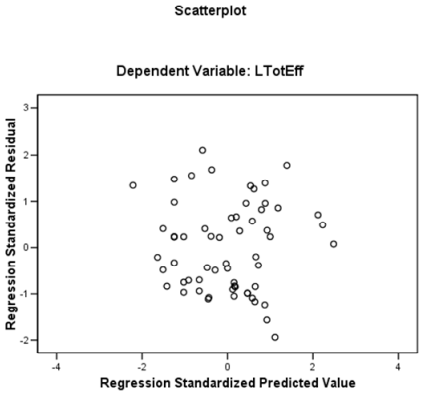
The prediction accuracy statistics are presented in Table 8, where we can see that the predictions using the stepwise regression model are very poor, assuming that a good model should present MMRE and MdMRE close to 25% and Pred(25) of at least 75% [4]. In addition, the cross-company model's prediction accuracy was significantly worse than predictions based on the median (2,792) and the mean (2,677.867) of the single-company data set. This pattern was confirmed by comparing the absolute residuals using Kendall's W Test, a non-parametric statistical test that compares the distributions of three or more related variables. If variables present very different distributions they are assumed to be significantly different from one another.

The estimates based on the mean and median models were very similar and not significantly different from one another. What these results suggest is that for a company using the mean or median effort of past projects is better than using a regression-based cross-company model.

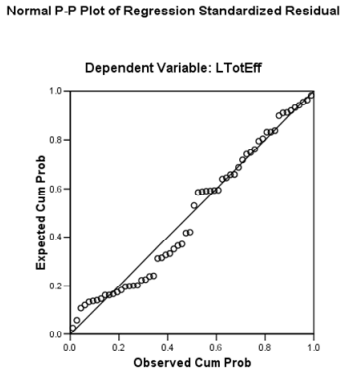
**Table 8. Prediction accuracy statistics for the cross-company data set**

| Prediction Accuracy | Estimates based on regression model (%) | Estimates based on mean model (%) | Estimates based on median model (%) |
|---|---|---|---|
| MMRE | 85.86 | 31.64 | 32.25 |
| MdMRE | 100 | 25.61 | 23.3 |
| Pred(25) | 6.67 | 46.67 | 66.67 |

These results do not corroborate those obtained in [20], where Mendes and Kitchenham found that the cross-company model presented similar predictions to those using a median model.




(a)




(b)

**Figure 1. Residual (a) and P-P plots (b) for best fitting cross-company model**

## 3.2 Results Based on Single-company Data

**The best single-company fitting model is described in**

Table 9. Its adjusted $R^2$ was 0.673, thus it explains 67.3% of the variation in *TotEff*. Note that this adjusted $R^2$ is smaller than that obtained for the cross-company model.

**Table 9. Best Fitting Model to calculate TotEff**

| Independent Variables | Coefficient | Std. Error | t | p>\|t\| |
|---|---|---|---|---|
| (constant) | 7.081 | .149 | 47.682 | .000 |
| TotHigh | .048 | .009 | 5.454 | .000 |

### *3.2.1 Checking the model*

The residual plot for the 15 projects showed two projects that seemed to have a large residual. This trend was also confirmed using Cook's distance, where these projects had their Cook's distances above the cut-off point (4/15).

To check the model's stability, a new model was built without these two projects, giving an adjusted $R^2$ of 0.721, which is greater than that for the previous model. In the new model the independent variables remained significant and the coefficients had very similar values to those in the previous model, indicating that the high influence data point did not need to be permanently removed from further analysis.

This model is not as good as the one described in S2 [20] (adjusted $R^2$ of 0.95), and also has not selected any of the variables selected in [20] that included *HFotsA* and *FotsA*. However, both models are exponential.

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

$$LTotEff = 7.081 + 0.048TotHigh \quad (4)$$

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

$$TotEff = 1189.157 \times e^{0.048TotHigh} \quad (5)$$

The P-P plot and the residual plot for the final single-company model are presented in Figure 2(a) and Figure 2(b) respectively. Figure 2(a) suggests that the residuals are normally distributed.

### *3.2.2 Measuring Prediction Accuracy*

To assess the accuracy of the predictions for the single-company model we employed a 15-fold cross-validation to the data set, where 14 projects at a time were in the training set and one project in the validation set. This means that for 15 times, a project was omitted from the data set, and an Equation, similar to that shown by Equation 4, was calculated using the remaining 14 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.

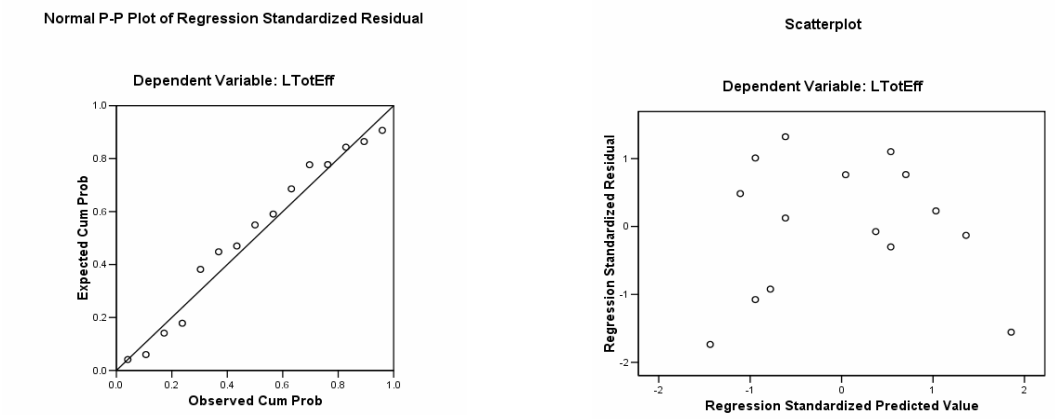
The prediction accuracy statistics are presented in Table 10, where we can see that the single-company model's prediction accuracy was not significantly different from predictions based on the mean and median of the data set. This result was confirmed using Kendall's W Test for related samples on absolute residuals.

**Table 10. Prediction accuracy statistics for the single-company data set**

| Prediction Accuracy | Estimates based on regression model (%) | Estimates based on mean model (%) | Estimates based on median model (%) |
|---|---|---|---|
| MMRE | 19.51 | 31.64 | 32.25 |
| MdMRE | 15.44 | 25.61 | 23.3 |
| Pred(25) | 73.33 | 46.67 | 66.67 |

What these results suggest is that effort estimates for the single-company projects based on the single-company data will be similar (when using a regression-based effort model) to the mean or the median effort for past projects.

These results do not corroborate those obtained in [20], where Mendes and Kitchenham found that the single-company model presented significantly better prediction than estimates based on the median effort.




**(a)** **(b)**

**Figure 2. P-P plot (a) and Residual (b) for best fitting single-company model**

## 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 15 single-company projects employed to validate the regression-based cross-company model (see Section 3.1) and the absolute residuals for each of the 15 single-company validation sets used to validate the regression-based single-company model (see Section 3.2). Their box plots are presented in Figure 3, 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 most 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.

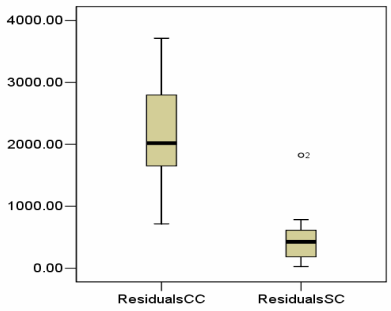



**Figure 3. Box plots for absolute residuals**

Apart from using box plots, we also applied the Paired T-test and the Wilcoxon Signed Ranks test for two related 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 S2 [20].

# 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 a given 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 [24]. 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. Since the Tukutuku data set presents some level of collinearity and outliers, like [20], we also investigated the use of CBR to obtain effort estimates.

Like S2 [20], 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 11, as follows:

- CBR using cross-company data set provided predictions not significantly different from those for the regression-based cross-company model (p<0.05). Our results corroborate those found in S2 [20].
- CBR using single-company data set provided predictions not significantly different from those for the regression-based single-company model (p<0.05). Our results contradict those in S2, where the CBR using single-company data set provided predictions significantly worse than those for the regression-based single-company model.
- CBR using cross-company presented significantly worse predictions than the CBR using single-company data set (p<0.05). Mendes and Kitchenham [20] found the opposite.

# 5. DISCUSSION

The research questions addressed in this study are as follows:

1. How successful is a cross-company data set when estimating effort for projects from a single company, when the estimate is obtained from a data set that does not include that company.
2. How successful is the use of a cross-company data set, compared to a single-company data set.

Our first research question is addressed by the results from Sections 3.1 and 4. The accuracy of estimates obtained for the 15 single-company projects using the regression-based cross-company model (see Equation 3) does not indicate good prediction accuracy. MMRE is 85.86%, which is poor (25% is considered "good" [4]), and Pred(25) is extremely poor (6.67%, when 75% indicates a good prediction model [4]). The same pattern is present for predictions obtained using CBR: MMRE is 92.54% and Pred(25) is 0%, both poor. Here our results corroborate those by Mendes and Kitchenham [20].

The absolute residuals obtained using the CBR and regression-based cross-company estimates were significantly worse than residuals obtained using both the mean and median effort. This suggests that, at least for the data set employed, there is no advantage to a company that does have past projects from which to develop their own models, to use a cross-company model to obtain effort estimates.

**Table 11. 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)** |
| MMRE | 92.54 | 15 | 85.86 | 19.51 |
| Median MRE | 93.13 | 15 | 100 | 15.44 |
| Pred(25) | 0 | 80 | 6.67 | 73.33 |

It can rely on the mean or median-based estimates. Our results did not corroborate those by Mendes and Kitchenham [20]. To address our second research question we compared the absolute residuals for the 15 single-company projects with the single-company model (see Sections 3.2 and 4) to those obtained using 15 single-company projects with the cross-company model (see Sections 3.3 and 4). The comparison was done using the paired T-test and the Wilcoxon Signed ranks test for two related samples. Results for the SWR and CBR indicated that absolute residuals for the single-company projects using the estimates obtained with the single-company data set were significantly lower (better) than absolute residuals obtained for the single-company projects using the estimates obtained with the cross-company data set. Our results using regression models corroborate those in S2, however our results for CBR did not. Similar to the trends observed for Regression-based models, our CBR predictions using the cross-

company data were significantly worse than both mean and median-based effort models; however, when using the single-company data, predictions were similar to those from the mean and median-based models. This comparison was not carried out in S2.

These results suggest that a mean or median-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 mean or median-based estimations. This is even more appropriate for Web companies that develop Web applications of the same type, using the same technologies and staff [13].

In comparison to the results obtained in S2, the patterns observed are as follows:

- Both regression- and CBR-based predictions for single-company projects using the cross-company data set were poor. These results corroborate those from S2.
- Both regression- and CBR-based predictions for single-company projects using the single-company data set were significantly better than regression- and CBR-based predictions for single-company projects using the cross-company data set. Our regression-based results corroborate those from S2, but our CBR-based results contradict those obtained in S2.
- The absolute residuals obtained using CBR and SWR, and employing the cross-company data set, were significantly worse than residuals obtained using both the mean and median effort.
- The absolute residuals obtained using CBR and SWR, and employing the single-company data set, were not significantly different to the residuals obtained using both the mean and median effort. Our results for regression did not corroborate those in S2.

We have observed in Section 2.3 (see Table 5) that the cross-company projects were overall smaller in size and effort than the single-company projects, however, unless their productivities vary widely, this should not be a reason to justify the results we have obtained. In order to compare their productivity we employed the productivity method proposed by Kitchenham and Mendes [12], where productivity is measured using the following Equation:

$$Productivity = AdjustedSize/Effort \quad (6)$$

The *AdjustedSize* measure contains only size measures that *together* are strongly associated with effort. In addition, the relationship between these size measures and effort does not need to be linear.

The benefits of using this method for measuring productivity are as follows [12]:

- The standard value of productivity is one, since it is obtained using the ratio of estimated to actual effort.
- A productivity value greater than one suggests above-average productivity.
- A productivity value smaller than one suggests below-average productivity.
- The stepwise regression technique used to build a regression model that represents the *AdjustedSize* measure can also be employed to construct upper and lower bounds on the productivity measure. These bounds can be used to assess whether the productivity achieved by a specific project is significantly better or worse than expected.
- The productivity measure automatically allows for diseconomies (or economies) of scale before being used in a productivity analysis. This means that an investigation of factors that affect productivity will only select factors that affect the productivity of all projects. If we ignore the impact of diseconomies (or economies) or scale, we run the risk of detecting factors that differ between large and small projects rather than factors that affect the productivity of all projects.

We compared both sets of productivity values (cross-company vs. single-company) using both parametric (the independent samples T-test) and non-parametric (the Mann-Whitney Test) tests. Both confirmed that productivity values all came from the same distribution, thus this confirms that the differences in size between cross- and single-company projects would not have affected the results we have observed.

# 6. CONCLUSIONS

Our results show that the cross-company data set provided poor predictions for the single-company projects and much worse predictions than the single-company data set. These results suggest that the cross-company data set was not successful either at estimating effort for projects from a single company, or in comparison to a single-company data set. Mendes and Kitchenham [20] obtained the same results for stepwise regression, using a different Tukutuku data set to ours. However, despite providing poor predictions for the single-company projects, the use of Case-Based Reasoning with cross-company data set by Mendes and Kitchenham did not give worse predictions than the single-company data set.

One possible reason for the better performance of the single-company data set, compared to the cross-company one, may be related to the size of the single-company data sets. Recently, Kitchenham et al. [15], by means of a systematic review, found that all studies where single-company predictions were significantly better than cross-company predictions employed smaller single-company data sets, smaller number of projects in the cross-company models, and databases where maximum effort was also smaller. They speculate that as single-company data sets grow, they incorporate less similar projects so that differences between single- and cross-company data sets cease to be significant.

As part of our future work we aim 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] G. Costagliola, S. Di Martino, F. Ferrucci, C. Gravino, G. Vitiello, G. Tortora, "A COSMIC-FFP Approach to Predict Web Application Development Effort", *Journal of Web Engineering* (Rinton Press), 5(2), 2006, 93-120.

[7] G. Costagliola, S. Di Martino, F. Ferrucci, C. Gravino, G. Tortora, G. Vitiello, "Effort estimation modeling techniques: a case study for web applications", in *Proceedings of International Conference on Web Engineering* (ICWE'06), Palo Alto, California, USA, 2006, 9-16.

[8] 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.

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

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

[11] B. A. Kitchenham, "A Procedure for Analyzing Unbalanced Datasets", *IEEE Transactions on Software Engineering*, 24(4), 1998, 278-301.

[12] Kitchenham, B.A., and E. Mendes. Software Productivity Measurement Using Multiple Size Measures*, IEEE Transactions on Software Engineering*, 30(12), 1023-1035.

[13] 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.

[14] Kitchenham, B.A. and N.R. Taylor. Software Cost Models. *ICL Technical Journal*, May 1984, 73-102.

[15] Kitchenham, B.A., E. Mendes, and G. Travassos, A Systematic Review of Cross- and Within-company Cost Estimation Studies, 2006, *Proceedings of Empirical Assessment in Software Engineering*, 89-98.

[16] 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.

[17] Lokan, C.J., and E. Mendes, Cross-company and Single-company Effort Models Using the ISBSG Database: a Further Replicated Study, *Proceedings of the ISESE'06*, 2006, 75-84.

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

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] I. Myrtveit, E. Stensrud, "A Controlled Experiment to Assess the Benefits of Estimating with Analogy and Regression Models", *IEEE Transactions on Software Engineering*, 25(4), 1999, 510-525.

[24] 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.

[25] 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.

[26] 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.