A Comparison of Cross- versus Single-company Effort Prediction Models for Web Projects

Burak Turhan
Department of Information Processing Science
University of Oulu
90014, Oulu, Finland
burak.turhan@oulu.fi

Emilia Mendes
Software Engineering Research Laboratory
Blekinge Institute of Technology
Karlskrona, Sweden
emilia.mendes@bth.se

Abstract—**Background: In order to address the challenges in companies having no or limited effort datasets of their own, cross-company models have been a focus of interest for previous studies. Further, a particular domain of investigation has been Web projects. Aim: This study investigates to what extent effort predictions obtained using cross-company (CC) datasets are effective in relation to the predictions obtained using single-company (SC) datasets within the domain of web projects. Method: This study uses the Tukutuku database. We employed data on 125 projects from eight different companies and built cross and single-company models with stepwise linear regression (SWR) with and without relevancy filtering. We also benchmarked these models against mean and median based models. We report a case-by-case analysis per company as well as a meta-analysis of the findings. Results: Results showed that CC models provided poor predictions and performed significantly worse than SC models. However, relevancy filtered CC models yielded comparable results to that of SC models. These results corroborate with previous research. An interesting result was that the median-based models were consistently better than other models. Conclusions: We conclude that companies that carry out Web development may use a median-based model for prediction until it is possible for the company to build its own SC model, which can be used by itself or in combination with median-based estimations.**

I. Introduction

When planning a project, the estimation of development effort/cost is a critical management activity, also crucial for the competitiveness of a software company. It aims at predicting an accurate effort estimate and using this information to allocate resources adequately, such that projects are completed within time and on budget. Most research in this field has looked at improving the estimation process via the use of past data from finished projects to build estimation models in order to provide effort predictions for new projects; however, there are challenges that a company faces that are associated with building its own data set of past projects [1]: (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 enough 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 previous studies to investigate to what extent effort estimation models built using cross-company (CC) data sets, i.e., data sets that contain project data volunteered by several companies, can provide suitable effort estimates for projects belonging to another company, when compared to effort estimates obtained using that company's own data on their past projects (single-company data set (SC)). Hence, the previous studies on the topic pursued the two research questions:

1) How successful is a cross-company dataset at estimating effort for projects from a single company?
2) How successful is the use of a cross-company dataset, compared to a single-company dataset, for effort estimation?

The first research question investigates the feasibility of CC models being applied to a validation set of SC projects. The second one compares the accuracy between predictions obtained using CC models and SC models in estimating the effort for SC projects.

In this paper, we scope the analysis of cross- versus single company predictions within the domain of Web projects. This is motivated by earlier studies stating the importance of domain scoping in estimation studies. For instance, Zimmermann et al. discusses the relevance of application domain for the problem of defect prediction [2] and Bakir et al. demonstrates how domain scoping can be effective for effort estimation in their analysis of embedded systems applications [3]. Furthermore, Web projects domain require specific attention for there are several differences between Web and software development projects, and the results observed using datasets of non-Web projects can not be readily applicable within the context of Web development. A detailed discussion on the differences between Web and software development is provided in Mendes et al.'s work [4].

Five studies to date, detailed in the next Section, have used datasets of Web projects in order to investigate the abovementioned research questions [5], [6], [7], [8], [9]. In addition, all these studies used Stepwise regression (manual and automated) (SWR) and/ or case-based reasoning (CBR) as effort prediction techniques, yet only up to two single-company datasets each time. One additional study [10] employed a self-tuning relevancy filtering analogy-based effort estimation tool (TEAK) to compare cross- to single-company predictions, though their cross- datasets were merged with projects from the single company being investigated. Given the wide choice of techniques available, some of which providing greater flexibility than SWR, CBR, or TEAK, it is important to also investigate the effectiveness of these techniques in order to understand better to what extent the results that have been obtained to date in the topic of our investigation, using Web projects, were driven mostly by the dataset characteristics, or by the choice of techniques. Hence, we employ SWR and CBR (i.e. relevancy filtering) methods in our study. Moreover, in cross company estimation problems, where the data originate from different sources in varying proportions, data heterogeneity or source component/ covariate shifts cause accuracy problems or conclusion instability in predictions across projects [11]. One way of handling the issue of data heterogeneity is to use relevancy filtering; that is to construct models not using all available training data, but to filter out some training data that is not relevant to the test data, for which we are making predictions. The application of this method has yielded promising results for cross-company Web project effort estimation [5].

Based on the motivations stated above, in this study, we investigate the effectiveness of SWR models and CBR based relevancy filtering within the scope of cross-company Web project effort estimation, analyzing eight different company cases covering 125 projects, in comparison to single company estimations. The database (i.e. Tukutuku) organization employed in this study reflects a setting of data on Web projects from eight single companies that enables the comparison of patterns and results throughout a larger number of cross- vs. single-company projects than previously reported in earlier studies.

Therefore, the main contributions of this paper are as follows: (i) We use 125 project data from eight different single-companies as part of the Tukutuku database, which provides the opportunity to conduct analysis across multiple companies that covers a wide spectrum of projects (to the best of our knowledge, this is the largest number of projects/companies used for analyzing cross- and single-company Web project effort estimation problem); (ii) We use a range of methods utilised in earlier studies and also employ relevancy filtering technique while reporting their effectiveness in a comparative way; (iii) We conduct a meta-analysis of our results based on the above-mentioned eight cases for Web project effort estimation for achieving a broader view on the subject and report new insights based on our meta-analysis; (iv) We advance the body of knowledge about Web project effort estimations by discussing our results in comparison to those of the previous studies on the topic, where we partially confirm and refute their findings.

II. Related Work

There have been five previous studies that compared cross-company to single-company predictions using Web project data, each detailed below. A comparison of those studies in terms of similarities and differences are summarized in Table I and Figure 1.

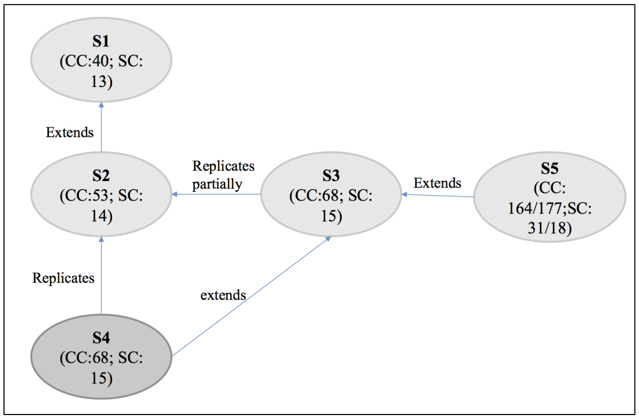

Fig. 1. Comparison of earlier studies on Tukutuku database of Web projects. The numbers associated with CC and SC in the figure correspond to the number of projects used in constructing CC and SC models.

S1: The first study (S1) was carried out in 2004 by Kitchenham and Mendes [6]. It investigated, using data on 53 Web projects from the Tukutuku database (40 cross-company and 13 from a single-company), to what extent a cross-company cost model could be successfully employed to estimate development effort for single-company Web projects. Their effort models were built using Forward Stepwise Regression (SWR) and they found that cross-company predictions were significantly worse than single-company predictions.

S2: The second study (S2) extended S1, also in 2004, by Mendes and Kitchenham [7], who used SWR and Case-based reasoning (CBR), and also data on 67 Web projects from the Tukutuku database (53 cross-company and 14 from a single-company). They built two cross-company and one single-company models and found that both SWR cross-company models provided predictions significantly worse than the single company predictions, and CBR cross-company data provided predictions significantly better than the single company predictions.

S3: By 2007 another 83 projects had been volunteered to the Tukutuku database (68 cross-company and 15 from a single-company), and were used by Mendes et al. [8] to carry out a third study (S3) partially replicating S2 (only one cross-company model was built), and using SWR and CBR. They corroborated some of S2's findings (SWR cross-company model provided predictions significantly worse than single-company predictions); however S2 found CBR cross-company predictions to be superior to CBR single-company predictions, which is the opposite of what was obtained in S3.

S4: Later, in 2008, Mendes et al. [9] conducted a fourth study (S4) that extended S3 to fully replicate S2. They used the same dataset used in S3, and their results corroborated most of those obtained in S2. The main difference between S2 and S4 was that one of S4's SWR cross-company models showed similar predictions to the single-company model, which contradicts the findings from S2.

TABLE I
COMPARISON OF PREVIOUS STUDIES USING DATA ON WEB PROJECTS

	Study S1 [6]	Study S2 [7]	Study S3 [8]	Study S4 [9]	Study S5 [5]
Year Published	2004	2004	2007	2008	2012
Database	Tukutuku	Tukutuku	Tukutuku	Tukutuku	Tukutuku
Application domain(s)	Mainly corporate, information, promotional, e-commerce	Mainly corporate, information, promotional, e-commerce	Mainly corporate, information, e-government, e-banking, Web portals, and intranet applications	Mainly corporate, information, e-government, e-banking, Web portals, and intranet applications	Mainly corporate, information, e-government, e-banking, Web portals, and intranet applications
Effort estimation techniques used	Manual Stepwise regression	Stepwise regression (SWR), Case-based reasoning (CBR)	Stepwise regression (SWR), Case-based reasoning (CBR)	Stepwise regression (SWR), Case-based reasoning (CBR)	Stepwise regression (SWR), Case-based reasoning (CBR)
Type of application	Web-based	Web-based	Web-based	Web-based	Web-based
Countries	Worldwide	Worldwide	Worldwide	Worldwide	Worldwide
Total Dataset size	53	67	83	83	195
Single company	13	14	15	15	31; 18
Underlying relationship bt. predictors and effort for Cross-company model	Non-linear	Non-linear	Non-linear	Non-linear	Non-linear
Underlying relationship bt. predictors and effort for Single-company model	Non-linear	Non-linear	Non-linear	Non-linear	Non-linear
Range of Effort values (converted to person hours)	Min:6 Max:5,000	Min:6 Max:5,000	Min: 1.10 Max: 3,712	Min: 1.10 Max: 3,712	
Size measure	23 different size measures	9 different size measures	11 different size measures	11 different size measures	11 different size measures
Was cross-company model built independently of the single company data?	No. Baseline model was built using whole dataset and variables selected by the model used to calibrate cross-company model (after removing single company projects)	Yes (CCM1) No (CCM2). CCM2: Baseline model was built using whole dataset and variables selected by the model used to calibrate cross-company model (after removing single company projects)	Yes	Yes (CCM1) No (CCM2). CCM2: Baseline model was built using whole dataset and variables selected by the model used to calibrate cross-company model (after removing single company projects)	Yes (CCM1) No (CCM2). CCM2: Baseline model was built using whole dataset and variables selected by the model used to calibrate cross-company model (after removing single company projects)
Single-company model showed significantly better accuracy than Cross-company model	Yes	Yes (SWR, for both CCM1 and CCM2) Yes (CBR)	Yes (SWR) Yes (CBR)	Yes (SWR for CCM1) No (SWR CCM2) Yes (CBR)	Yes (MSWR+LR) Yes (CBR) No (NN-Filtering+ MSWR+LR)

S5: After S4 was published, another 45 projects (i.e., 31 coming from a single company and 14 from different companies) were volunteered to the Tukutuku dataset, therefore a fifth study - S5 [5], extended S3 using the entire set of 195 projects from the Tukutuku database, and two single-company datasets (31 and 18 projects respectively). In addition, they also investigated to what extent applying a filtering mechanism [10], [12], [13], [14] to cross-company datasets prior to building prediction models can affect the accuracy of the effort estimates they provide. Their results (without filtering) corroborated those from S3; however, the filtering mechanism significantly improved the prediction accuracy of cross-company models when estimating single-company projects, making their prediction accuracy similar.

The study reported in this paper is neither an *exact* replication nor an incremental study based on the previous one. In this paper, we extend the scope of previous analysis with all available relevant data in the Tukutuku database, i.e. cross company datasets from eight companies with a total of 125 projects, using SWR and relevancy filtering, and report the results per company as well as a meta-analysis of the results. Nevertheless, pursuing the same research questions would classify this study as a conceptual replication of earlier studies; whereas it could also be considered as a meta-analysis since we also investigate the overall status of the results for all companies in the Tuktutuku database. Please note that some descriptive parts of the paper (e.g. problem description, related work, description of dataset and methods) partially re-use relevant text from authors' previous publications on the topic.

III. RESEARCH METHODOLOGY

A. Research Questions

This study addresses the following research questions:

RQ1: How successful is a cross-company dataset at estimating effort for Web projects from a single company?

RQ2: How successful is the use of a cross-company dataset, compared to a single-company dataset, for Web effort estimation?

As mentioned earlier, the first research question investigates the feasibility of CC models being applied to a validation set of SC projects, whereas the second one compares the accuracy between predictions obtained using CC models and SC models in estimating the effort for SC projects.

B. Dataset Description

The analysis presented in this paper used the Web projects data from the Tukutuku database [7]. The data represents a wide range of Web projects, from static to dynamic applications mostly developed using content management systems. Each Web project in the Tukutuku database is characterized by process and product variables [4]. These size measures and cost drivers have been obtained from the results of a survey investigation [7], using data from on-line Web forms aimed at giving quotes on Web development projects. In addition, an established Web company as well as a second survey involving 33 Web companies in New Zealand, have also confirmed these measures and cost drivers. Furthermore, the identified variables are constructed from information their customers can provide at a very early stage in project development. For the purposes of our analysis, we used the available data in the Tukutuku database according to the following criteria. We included projects from the companies that have at least five projects in the database. This was necessary in order to be able to construct single-company regression models for answering RQ2. Further, we used the same set of projects for answering RQ1 in order to use consistent data across our research questions. This resulted in the inclusion of 125 projects from eight different companies (i.e. 70 projects were removed).

Table II summarizes the contribution of each identified company along with their contributions (in terms of projects) to the final dataset. We excluded categorical variables from the analysis as in [5], since they require creation of many dummy variables; a situation that we want to avoid given the small sample sizes per company. Table III provides the description of variables used in this analysis. Table IV provides the descriptive statistics for the dataset used in our analysis[1]. The distribution of all the variables was checked for normality, which led to their transformation in order to comply with the assumptions of the estimation technique being used. For each company, we have converted the variables to their standardized z-scores (using mean and standard deviation estimates of the samples) before constructing models. The real and estimated effort values are then converted back to their original forms during evaluation.

TABLE II
SUMMARY OF PROJECTS PER COMPANY

Company	# of Projects	% Projects
C1	14	11.20
C2	20	16.00
C3	15	12.00
C4	6	4.80
C5	13	10.40
C6	8	6.40
C7	31	24.80
C8	18	14.40
Total	125	100

TABLE III
DESCRIPTION OF VARIABLES (DEP:DEPENDENT, IND:INDEPENDENT)

Type	Variable	Description
IND	nLang	Number of different development languages used.
IND	DevTeam	Size of a project?s development team.
IND	TeamExp	Average team experience with the development language(s) employed.
IND	TotWP	Total number of Web pages (new and reused).
IND	NewWP	Total number of new Web pages.
IND	TotImg	Total number of images (new and reused).
IND	NewImg	Total number of new images created.
IND	Fots	Number of features reused without any adaptation.
IND	HFotsA	Number of reused high-effort features/functions adapted.
IND	Hnew	Number of new high-effort features/functions.
IND	TotHigh	Total number of high-effort features/functions
IND	FotsA	Number of reused low-effort features adapted.
IND	New	Number of new low-effort features/functions.
IND	TotNHigh	Total number of low-effort features/functions
DEP	TotEff	Actual total effort used to develop the Web application.

TABLE IV
DESCRIPTIVE STATISTICS

Variable	Mean	Median	St.Dev.	Min	Max
nLang	3.89	4.00	1.45	1.00	8.00
DevTeam	2.58	2.00	2.38	1.00	23.00
TeamExp	3.83	4.00	2.03	1.00	10.00
TotWP	69.48	26.00	185.69	1.00	2000.00
NewWP	49.55	10.00	179.14	0.00	1980.00
TotImg	98.58	40.00	218.37	0.00	1820.00
NewImg	38.27	1.00	125.47	0.00	1000.00
Fots	3.19	1.00	6.24	0.00	63.00
HFotsA	11.96	0.00	59.85	0.00	611.00
Hnew	2.08	0.00	4.70	0.00	27.00
TotHigh	14.04	1.00	59.63	0.00	611.00
FotsA	2.24	0.00	4.53	0.00	38.00
New	4.24	1.00	9.65	0.00	99.00
TotNHigh	6.48	4.00	13.22	0.00	137.00
TotEff	468.11	88.00	938.51	1.10	5000.00

C. Method Description

The techniques used to obtain effort estimates were Stepwise Linear Regression (SWR) and Nearest Neighbor (NN) relevancy filtering [15]. All results presented were obtained using the scripts implemented in MATLAB 2009a with standard libraries and toolboxes.

1) Stepwise Linear Regression: Stepwise linear regression (SWR) is a statistical technique whereby a prediction model (Equation) that represents the relationship between independent (e.g. number of Web pages) and dependent variables (e.g. total Effort) is built. The independent variables used with SWR were selected in a stepwise way (using *stepwisefit* function provided by MATLAB Statistical Toolbox) on the training set. We also verified the stability of each model built using SWR by checking the Cook's distances for each data point within the regression model. We removed influential points that have greater Cook's distance than $4/n$, where n is the sample size.

2) Nearest-Neighbor (NN) Relevancy Filtering: The NN Filtering technique [15] uses the k-Nearest Neighbor (k-NN) method to measure the similarity among the projects in validation and training sets by computing the Euclidean

[1] Statistics per company and experimental scripts are available at http://cc.oulu.fi/~bturhan/ccsc.

distance between those projects' features. The aim is to reduce the size of the training set such that it only includes the most similar projects to those in the validation set. To measure the similarity between projects we used all dataset features except for the effort data (with $k = 10$), since this corresponds to a real life situation, where the development effort is unknown when estimation takes place.

D. Evaluation Criteria

The accuracy of the effort estimates was assessed using statistical tests together with accuracy measures based on absolute residuals (i.e., unsigned difference between actual and estimated effort). To check whether the differences in estimation accuracy between the cross-company and single-company models were legitimate or due to chance, we employed the Wilcoxon Signed Ranks tests (on absolute residuals) to check if there is a statistically significant difference between the medians of two samples. We set $\alpha = 0.05$ in all cases [16], [9], employed Hedges' g effect size [17], and used the Mean Magnitude of Relative Error (MMRE), the Median MRE (MdMRE), and Pred(25) as accuracy measures [18] .

E. Description of Experimental Setup

We used a leave-one-out cross-validation scheme in our experimental setup to evaluate the performance of SC models. For evaluation of CC models, the training sets are composed of combining seven cross company datasets, leaving one single company dataset as the validation set. We employed the mean and median-based predictions (i.e., effort estimate is respectively the mean or median effort for the training set) as benchmarks for our prediction models since performing similarly or worse than these benchmarks is a strong indicator of poor performance. Note that we have built mean and median-based models using both cross and single-company data. In addition to reporting case-by-case results for all eight companies in our analysis, we also pool the baseline statistics for detecting differences in prediction accuracy, that is absolute residuals, and conduct a meta-analysis of the eight cases to answer our research questions.

1) Steps to follow to answer RQ1:

- Apply SWR and SWR+NN filtering to build a cross-company cost model using seven cross-company datasets, leaving one dataset aside as single company dataset for validation. This step is repeated for each of the eight datasets.
- Use the models from previous step to estimate effort for each of the single-company projects. The single-company projects are the validation sets used to obtain effort estimates for each of the eight companies. The estimated efforts obtained for each project is also used to calculate accuracy statistics (i.e., MRE, MdMRE, Pred(25)) based on absolute residuals.

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

2) Steps to follow to answer RQ2:

- Apply SWR to build single-company cost models using the single-company data sets with leave-one-out cross validation.
- Obtain the prediction accuracy of estimates for the models obtained in the previous step.
- Compare the accuracy of models obtained in the second step to those obtained from CC models.

Steps 1 and 2 simulate the situation, where a company builds a model using its own data set and then uses this model to estimate effort for its new projects. Step 3 compares these models with CC models.

F. Threats to Validity

In terms of construct validity, as discussed in Section 2, the size measures and cost drivers used in the Tukutuku database, and therefore in our study, have been obtained from the results of a survey investigation and have also been confirmed by an established Web company and a second survey [19]. Consequently, it is our belief that the 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 the project development. As for data quality, it was found that at least for 93.8% of Web projects in the Tukutuku database effort values were based on recorded data [19]. With respect to the conclusion validity we carefully applied the statistical tests, verifying all the required assumptions. Moreover, we also employed effect size to assess the relevance of the obtained results. As for external validity, let us observe that the Tukutuku dataset comprises data on projects volunteered by individual companies, and therefore it does not represent a random sample of projects from a defined population. This means that we cannot conclude that the results of this study apply to other companies different from the ones that volunteered the data used here. However, we believe that Web companies that develop projects with similar characteristics to those used in this paper may be able to apply our results to their Web projects. On a final note, we used data from companies that has at least five projects in the TukuTuku dataset in order to be able to construct single company models for comparison with cross company models.

IV. Results

A. Addressing RQ1

RQ1: How successful is a cross-company dataset at estimating effort for projects from a single company?

Table V shows the prediction accuracies, obtained by applying a leave-one-out cross-validation procedure, for cross-company models built without any filtering mechanism (Cross Company Regression: CCR), and cross-company models built using the NN filtering mechanism (Cross Company Filtered Regression: CCFR), based on MMRE, MdMRE and Pred(25). The statistical significance of the results, based on absolute residuals, was checked using the Wilcoxon test ($\alpha = 0.05$). In Table V, values in italics point to statistically significantly results comparing the CCR model with the mean-based model;

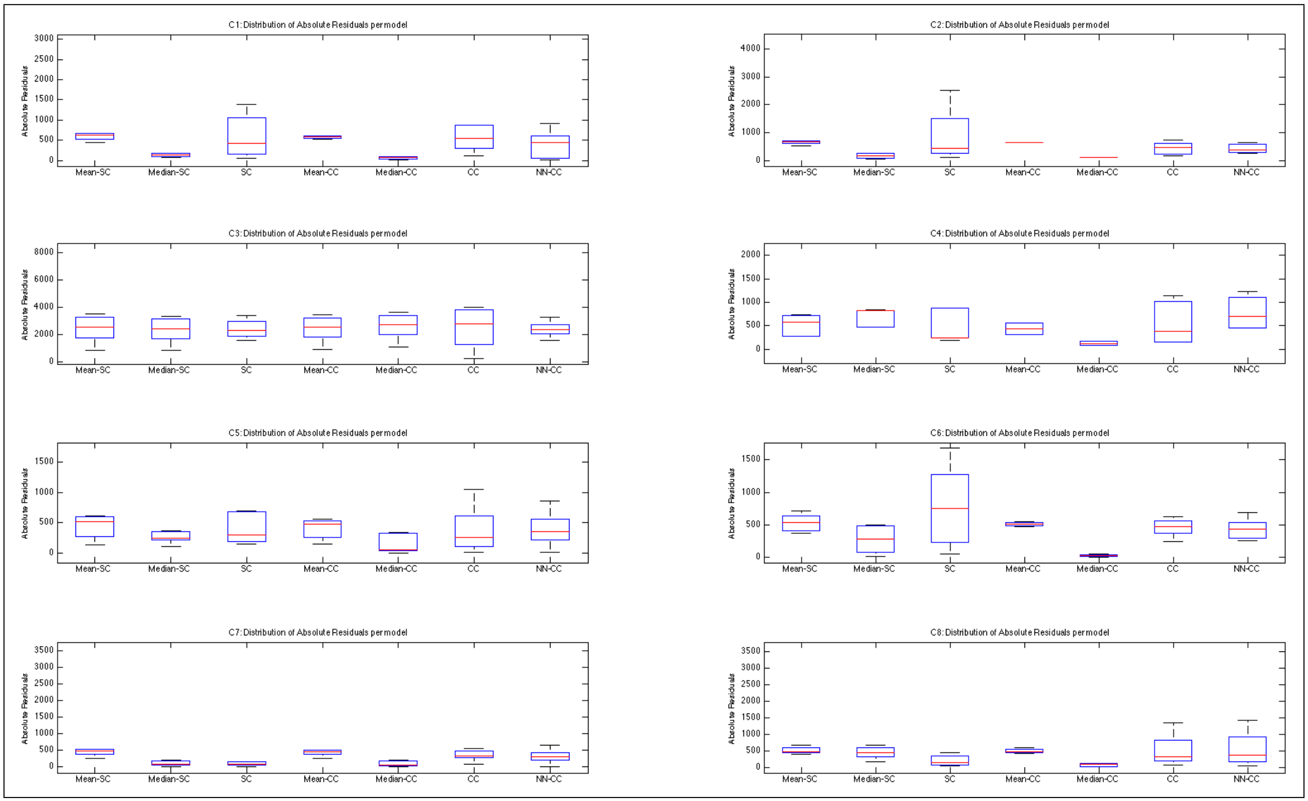

Fig. 2. Boxplots of Absolute Residuals per Company

and similarly, values in bold face represents statistically significantly results comparing the CCR model with the median-based model.

Overall, the accuracy, either based on stepwise regression, or on the mean/median models, is very poor. In addition, except for companies 3 and 5 (in CCR), and companies 1, 3 and 5 (CCFR), in 33 out of the 48 cases (67%) median-based model provided significantly superior results than both the mean- or both regression-based models. These results are also visible in the boxplots of residuals for each company (see Figure 2).

Although not directly related to RQ1, Table VII (description provided in the next section) also shows that there are no significant differences between accuracy values obtained using a simple stepwise regression and NN-filtering with stepwise regression. The same observation can be made from Figure 3. Overall, these results suggest that a median-based cross company estimation could be used for prediction 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 median-based estimations [6].

B. Addressing RQ2

RQ2: How successful is the use of a cross-company dataset, compared to a single-company dataset, for effort estimation?

Table VI reports the prediction accuracies obtained by applying a leave-one-out cross-validation procedure to the eight single-company datasets employed herein. Italic and bold face entries follow the same notation as in Table V.

To address our second research question we used the Wilcoxon test to compare: i) the absolute residuals obtained from using the different single-company models (i.e. Table VI) to those obtained using the cross-company models (i.e. Table V). Boxplots of all absolute residuals per company are given in Figure 2. A summary of the results, in addition to overall results from the meta-analysis, are also presented in Table VII and Figure 3.

In Table VII, each cell shows eight characters side by side, which can either be a plus sign ('+'), a zero ('0'), or a minus sign ('-'); underneath these characters, there is also a single character that is displayed, which also either can be a plus sign, a zero, or a minus sign. Each of the eight characters represents the outcome of a statistical significance test: zero meaning no statistical difference; minus sign meaning 'row' model significantly worse than 'column' model; and plus sign meaning 'row' model significantly better than 'column' model. For example, if we look at Table VII focusing on the cell in the first row and the third column, we have the following configuration: '000000–'. This means that: i) there were no statistically significant differences in absolute residuals between the SC-Mean and the SCR models, based on the data from the first six single company datasets; ii) the SC-Mean models built using data from the single company datasets 7 and 8 were significantly inferior to the SCR models also built using the same two datasets. In addition, the single minus

TABLE V
PREDICTION ACCURACY FOR CC MODELS

Comp.	Metric	CCR	CCFR	CC-Mean	CC-Median
C1	MMRE	36.90	16.20	30.07	**4.11**
	MdMRE	16.64	8.57	23.03	**2.90**
	Pred(25)	0.00	0.00	0.00	**7.14**
C2	MMRE	*101.97*	125.42	192.03	**34.33**
	MdMRE	*107.48*	97.80	140.43	**24.88**
	Pred(25)	*0.00*	0.00	0.00	**0.00**
C3	MMRE	1.23	0.99	0.89	0.97
	MdMRE	1.33	0.78	0.91	0.97
	Pred(25)	20.00	0.00	0.00	0.00
C4	MMRE	59.50	47.08	*53.27*	**7.81**
	MdMRE	41.89	35.42	*31.82*	**4.65**
	Pred(25)	0.00	0.00	*14.29*	**0.00**
C5	MMRE	3.69	4.55	7.03	0.97
	MdMRE	1.83	1.50	3.42	0.80
	Pred(25)	7.69	15.38	0.00	15.38
C6	MMRE	7.62	5.94	8.25	**0.54**
	MdMRE	5.97	6.05	7.31	**0.41**
	Pred(25)	0.00	0.00	0.00	**37.50**
C7	MMRE	*4.71*	4.20	6.63	**0.79**
	MdMRE	*3.15*	1.98	4.46	**0.71**
	Pred(25)	*0.00*	6.45	0.00	**22.58**
C8	MMRE	5.40	5.79	5.21	**0.57**
	MdMRE	2.83	3.26	2.87	**0.58**
	Pred(25)	0.00	5.56	0.00	**22.22**

TABLE VI
PREDICTION ACCURACY FOR SC MODELS

Comp.	Metric	SCR	SC-Mean	SC-Median
C1	MMRE	16.54	32.32	**8.02**
	MdMRE	13.06	24.37	**5.69**
	Pred(25)	0.00	0.00	**7.14**
C2	MMRE	163.54	202.59	**68.52**
	MdMRE	109.09	146.01	**39.44**
	Pred(25)	0.00	0.00	**0.00**
C3	MMRE	0.95	0.88	0.85
	MdMRE	1.00	0.91	0.88
	Pred(25)	0.00	0.00	0.00
C4	MMRE	24.40	68.27	77.27
	MdMRE	17.77	40.44	46.13
	Pred(25)	0.00	0.00	0.00
C5	MMRE	3.09	7.77	3.31
	MdMRE	2.54	3.74	0.97
	Pred(25)	0.00	0.00	0.00
C6	MMRE	10.12	9.19	5.76
	MdMRE	9.31	7.80	4.36
	Pred(25)	0.00	0.00	12.50
C7	MMRE	*1.22*	6.86	1.00
	MdMRE	*0.55*	4.62	0.76
	Pred(25)	*16.13*	0.00	22.58
C8	MMRE	***3.33***	5.59	5.44
	MdMRE	***1.56***	2.89	2.72
	Pred(25)	***11.11***	5.56	5.56

sign underneath the eight-characters represents the combined results from the meta-analysis. Therefore, in this example, our meta-analysis shows that the SC-Mean model provides significantly inferior predictions when compared to the SCR model.

The shaded cells are of particular interest within the context of RQ2, as they highlight the comparisons between SC and CC models. The meta-analysis results show that SC models presents significantly superior predictions to CC models (with small effect size); however, when compared to CC models built in combination with NN filtering, SC models shows similar accuracy. This suggests that CC models that are built in combination with a NN filtering mechanism may be competing models to SC models. The results between the SC and CC (with and without NN filtering) corroborate those from previous studies that compared cross- and single-company models within the context of Web development projects. Note that the meta-analysis showed no significant differences between CCR and CCFR models, as can be observed in Figure 3.

Finally, unlike previous studies, this is the first time in which median-based models (CC-Median) show significantly superior predictions overall, when compared to CC models (with and without NN filtering) (with large effect size). Given that the SC-Median models showed comparable accuracy to SCR models, our recommendation remains unchanged: that companies could use a median-based model for prediction 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 median-based estimations.

V. CONCLUSIONS

In this study, we have used data from eight different single company datasets in the Tukutuku database to compare the accuracy between estimates obtained using cross-company and single-company models built using respectively the nearest neighbor filtering with stepwise regression and regular stepwise regression. In addition, we also carried out a meta-analysis based on the combined absolute residuals obtained from all individual analyses of the company cases.

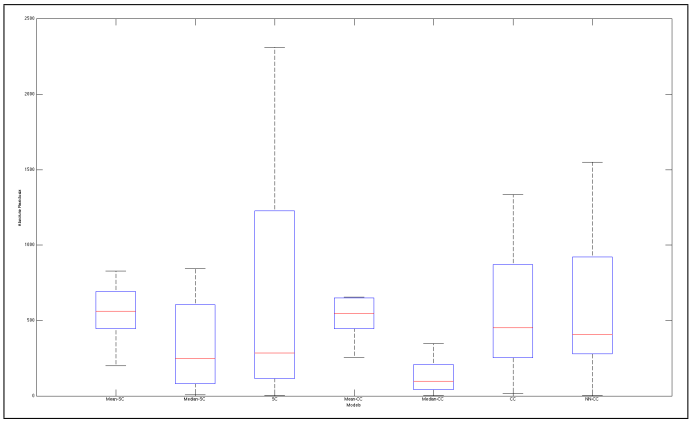

Fig. 3. Boxplot of Absolute Residuals Combined for Meta-Analysis

The cross-company models in general provided poor predictions for the single company projects; however, when compared to the single-company predictions, results were mixed: cross-company models built using regular (i.e. no filtering) regression models provided predictions significantly worse than the predictions obtained from single-company models; however, when built using the nearest neighbor filtering with stepwise regression, cross-company models presented competing accuracy to single-company models. This finding corroborate with previous work in this area, and suggest that using filtering techniques, which may contribute to creating more homogenous training sets, may provide the means to

TABLE VII
SUMMARY TABLE OF RESULTS. IN EACH CELL, THE EIGHT MARKERS ('0': NO DIFFERENCE, '-': SIGNIFICANTLY WORSE, '+': SIGNIFICANTLY BETTER) CORRESPOND TO EIGHT COMPANIES IN THE DATASET, AND THE SINGLE MARKER UNDERNEATH THEM SUMMARISES THE RESULT OF THE META-ANALYSIS OVER EIGHT COMPANIES.

	SC-Mean	SC-Med	SCR	CC-Mean	CC-Med	CCR	CCFR
SC-Mean			0 0 0 0 0 0 - - -				
SC-Med			+ + 0 0 0 0 0 - 0				
SCR	0 0 0 0 0 0 + + +	- - 0 0 0 0 0 + 0				0 0 0 0 0 0 + + +	0 0 0 0 0 0 + 0 0
CC-Mean						0 - 0 + 0 0 - 0 -	0 + 0 0 0 0 + 0 -
CC-Med						+ + 0 + 0 + + + +	0 + 0 + 0 + + + +
CCR			0 0 0 0 0 0 - - -	0 + 0 - 0 0 + 0 +	- - 0 - 0 - - - -		0 0 0 0 0 0 0 0 0
CCFR			0 0 0 0 0 0 - 0 0	0 - 0 0 0 0 - 0 +	0 - 0 - 0 - - - -	0 0 0 0 0 0 0 0 0	

improve the effectiveness of cross-company models, when compared to single-company models. These results corroborate those reported in the literature on traditional software projects for software effort estimation and defect prediction. Unlike in previous studies, the median-based cross-company model presented significantly better predictions that any of the cross-company models, thus suggesting that companies that carry out Web development may use a median-based model for prediction until it is possible for the company to build its own single-company model, which can be used by itself or in combination with median-based estimations.

In future work we will investigate other filtering methods and additional techniques such as the Weighted Least Squares Regression to take the relative importance of the projects into account while building estimation models. We will also examine the effect of the cross-company dataset size in relation to the prediction accuracies obtained from single-company datasets.

REFERENCES

[1] B. Kitchenham, E. Mendes, and G. H. Travassos, "Cross versus within-company cost estimation studies: A systematic review," *Software Engineering, IEEE Transactions on*, vol. 33, no. 5, pp. 316–329, 2007.

[2] T. Zimmermann, N. Nagappan, H. Gall, E. Giger, and B. Murphy, "Cross-project defect prediction: A large scale experiment on data vs. domain vs. process," in *Proceedings of the*, ser. ESEC/FSE '09. New York, NY, USA: ACM, 2009, pp. 91–100. [Online]. Available: http://doi.acm.org/10.1145/1595696.1595713

[3] A. Bakır, B. Turhan, and A. B. Bener, "A new perspective on data homogeneity in software cost estimation: a study in the embedded systems domain," *Software Quality Journal*, vol. 18, no. 1, pp. 57–80, 2010. [Online]. Available: http://dx.doi.org/10.1007/s11219-009-9081-z

[4] E. Mendes, N. Mosley, and S. Counsell, "The need for web engineering: An introduction," in *Web Engineering*, E. Mendes and N. Mosley, Eds. Springer Berlin Heidelberg, 2006, pp. 1–27. [Online]. Available: http://dx.doi.org/10.1007/3-540-28218-1_1

[5] F. Ferrucci, E. Mendes, and F. Sarro, "Web effort estimation: The value of cross-company data set compared to single-company data set," in *Proceedings of the 8th International Conference on Predictive Models in Software Engineering*, ser. PROMISE '12. New York, NY, USA: ACM, 2012, pp. 29–38. [Online]. Available: http://doi.acm.org/10.1145/2365324.2365330

[6] B. A. Kitchenham and E. Mendes, "A comparison of cross-company and single-company effort estimation models for web applications," in *Proceedings EASE?04*, 2004, pp. 47–55.

[7] E. Mendes and B. Kitchenham, "Further comparison of cross-company and within-company effort estimation models for web applications," in *Software Metrics, 2004. Proceedings. 10th International Symposium on*, 2004, pp. 348–357.

[8] E. Mendes, S. D. Martino, F. Ferrucci, and C. Gravino, "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?" in *WWW*, C. L. Williamson, M. E. Zurko, P. F. Patel-Schneider, and P. J. Shenoy, Eds. ACM, 2007, pp. 963–972. [Online]. Available: http://dblp.uni-trier.de/db/conf/www/www2007.html#MendesMFG07

[9] ——, "Cross-company vs. single-company web effort models using the tukutuku database: An extended study," *Journal of Systems and Software*, vol. 81, no. 5, pp. 673 – 690, 2008, ¡ce:title¿Software Process and Product Measurement¡/ce:title¿. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0164121207002385

[10] E. Kocaguneli and T. Menzies, "How to find relevant data for effort estimation?" in *Empirical Software Engineering and Measurement (ESEM), 2011 International Symposium on*, 2011, pp. 255–264.

[11] B. Turhan, "On the dataset shift problem in software engineering prediction models," *Empirical Softw. Engg.*, vol. 17, no. 1-2, pp. 62–74, Feb. 2012. [Online]. Available: http://dx.doi.org/10.1007/s10664-011-9182-8

[12] E. Kocaguneli, T. Menzies, A. Bener, and J. Keung, "Exploiting the essential assumptions of analogy-based effort estimation," *Software Engineering, IEEE Transactions on*, vol. 38, no. 2, pp. 425–438, 2012.

[13] T. K. Le-Do, K.-A. Yoon, Y.-S. Seo, and D.-H. Bae, "Filtering of inconsistent software project data for analogy-based effort estimation," in *Computer Software and Applications Conference (COMPSAC), 2010 IEEE 34th Annual*, 2010, pp. 503–508.

[14] Y. F. Li, M. Xie, and T. N. Goh, "A study of project selection and feature weighting for analogy based software cost estimation," *J. Syst. Softw.*, vol. 82, no. 2, pp. 241–252, Feb. 2009. [Online]. Available: http://dx.doi.org/10.1016/j.jss.2008.06.001

[15] B. Turhan, T. Menzies, A. B. Bener, and J. Di Stefano, "On the relative value of cross-company and within-company data for defect prediction," *Empirical Softw. Engg.*, vol. 14, no. 5, pp. 540–578, Oct. 2009. [Online]. Available: http://dx.doi.org/10.1007/s10664-008-9103-7

[16] B. Kitchenham, L. Pickard, S. MacDonell, and M. Shepperd, "What accuracy statistics really measure [software estimation]," *Software, IEE Proceedings -*, vol. 148, no. 3, pp. 81–85, 2001.

[17] V. B. Kampenes, T. Dybå, J. E. Hannay, and D. I. K. Sjøberg, "Systematic review: A systematic review of effect size in software engineering experiments," *Inf. Softw. Technol.*, vol. 49, no. 11-12, pp. 1073–1086, Nov. 2007. [Online]. Available: http://dx.doi.org/10.1016/j.infsof.2007.02.015

[18] D. Azhar, E. Mendes, and P. Riddle, "A systematic review of web resource estimation," in *Proceedings of the 8th International Conference on Predictive Models in Software Engineering*, ser. PROMISE '12. New York, NY, USA: ACM, 2012, pp. 49–58. [Online]. Available: http://doi.acm.org/10.1145/2365324.2365332

[19] E. Mendes, N. Mosley, and S. Counsell, "A comparison of cross-company and single-company effort estimation models for web applications," in *Proceedings EASE?03*, 2003, pp. 1–22.