Ranking and Clustering Software Cost Estimation Models through a Multiple Comparisons Algorithm

Nikolaos Mittas and Lefteris Angelis

Abstract—Software Cost Estimation can be described as the process of predicting the most realistic effort required to complete a software project. Due to the strong relationship of accurate effort estimations with many crucial project management activities, the research community has been focused on the development and application of a vast variety of methods and models trying to improve the estimation procedure. From the diversity of methods emerged the need for comparisons to determine the best model. However, the inconsistent results brought to light significant doubts and uncertainty about the appropriateness of the comparison process in experimental studies. Overall, there exist several potential sources of bias that have to be considered in order to reinforce the confidence of experiments. In this paper, we propose a statistical framework based on a multiple comparisons algorithm in order to rank several cost estimation models, identifying those which have significant differences in accuracy, and clustering them in nonoverlapping groups. The proposed framework is applied in a large-scale setup of comparing 11 prediction models over six datasets. The results illustrate the benefits and the significant information obtained through the systematic comparison of alternative methods.

Index Terms—Cost estimation, management, metrics/measurement, statistical methods

1 Introduction

The importance and the significant role of *Software Cost Estimation* (SCE) to the well-balanced management of a forthcoming project are clearly portrayed through the introduction and utilization of a large number of techniques during the past decades [1].

The rapidly increased need of large-scaled and complex software systems leads managers to settle SCE as one of the most vital activities that is closely related with the success or failure of the whole development process. Inaccurate estimates can be proved catastrophic to both developers and customers since they can cause the delay of the product deliverables or, even worse, the cancellation of a contract.

Due to the above-mentioned requirements, interest has been focused on the open research problem of the selection of the "*best*" estimation technique. According to an extended systematic review of studies [1], the most common research topic of SCE is the introduction and evaluation of estimation methods. On the other hand, the variety of prediction methods is also associated with contradictory and inconsistent findings regarding the superiority of one technique over others.

The most determining factor for these controversial results seems to be an inherent characteristic of prediction systems, i.e., their strong dependency on the kind of available data (types and number of project attributes and sample size) used in model fitting [2]. The complexity of building an accurate model swiftly increases if we consider the alternative variations of a generic estimation method (e.g., regression analysis). In several studies researchers have based their inferences on a small number of datasets, so generalization of findings may be quite misleading.

Furthermore, there is an ongoing discussion and lack of convergence regarding the appropriateness of the error measures used for the comparison of alternative models [3]. Although *Mean Magnitude of Relative Error* (MMRE) has been criticized as a problematic accuracy measure to select the "best" model [4], it continues to be considered as the main indicator for the performance of SCE methods.

A certain limitation of several past studies is comparison without using appropriate statistical hypothesis testing. This can lead to erroneous results and groundless generalizations regarding the predictive accuracy of estimation techniques [5]. Although comparison of methods without statistical tests may lead to unsound results [6], many recent papers still base their findings solely on single indicators [7].

Another source of bias can also be the statistical procedure that is used when comparing multiple prediction techniques. In the case of a simple comparison between two competitive models, the null hypothesis is examined via a classical statistical test (i.e., *paired t-test* or *Wilcoxon signed rank test*). With more than two comparative models, the meaning of "significant difference" becomes more complicated, and the problems associated with it are known in statistics as the "*multiple comparisons problems*." Due to the large number of proposed cost estimation methods, it is necessary for project managers to systematically base their

- *The authors are with the Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki 54124, Greece.*
E-mail: {nmittas, lef}@csd.auth.gr.

Manuscript received 21 July 2011; revised 24 Feb. 2012; accepted 26 June 2011; published online 5 July 2011.
Recommended for acceptance by R. Jeffery.
For information on obtaining reprints of this article, please send e-mail to: tse@computer.org, and reference IEEECS Log Number TSE-2011-07-0214.
Digital Object Identifier no. 10.1109/TSE.2012.45.

0098-5589/13/$31.00 © 2013 IEEE Published by the IEEE Computer Society

choice of the most accurate model on well-established statistical procedures in order to diminish the uncertainty threatening the estimation process [8]. However, to the best of our knowledge, the problem of simultaneous comparisons among multiple prediction models has not been studied yet in the sense that there is no statistical procedure which can identify the significant differences between a number of cost estimation methods and at the same time be able to rank and cluster them, designating the best ones.

All of the issues discussed above lead us to conclude that there is an imperative need to investigate what the state of the art in statistics is before trying to derive conclusions and unstable results concerning the superiority of a prediction model over others for a particular dataset. The answer to this problem cannot constitute a unique solution since the notion of "best" is quite subjective. In fact, a practitioner can always rank the prediction models according to a predefined accuracy measure, but the critical issue is to identify how many of them are evidently the best, in the sense that their difference from all the others is statistically significant. Hence, the research question of finding the "best" prediction technique can be restated as a problem of identifying a subset or a group of best techniques.

The aim of the paper is therefore to propose a statistical framework for comparative SCE experiments concerning multiple prediction models. It is worth mentioning that the setup of the current study was also inspired by an analogous attempt dealing with the problem of comparing classification models in *Software Defect Prediction*, a research area that is also closely related to the improvement of software quality [9].

The proposed methodology is based on the analysis of a *Design of Experiment* (DOE) or *Experimental Design*, a basic statistical tool in many applied research areas such as engineering, financial, and medical sciences. In the field of SCE it has not yet been used in a systematic manner. Generally, DOE refers to the process of planning, designing, and analyzing an experiment in order to derive valid and objective conclusions effectively and efficiently by taking into account, in a balanced and systematic manner, the sources of variation [10]. In the present study, DOE analysis is used to compare different cost prediction models by taking into account the blocking effect, i.e., the fact that they are applied repeatedly on the same training-test datasets.

The proposed statistical methodology is also based on an algorithmic procedure which is able to produce nonoverlapping clusters of prediction models, homogeneous with respect to their predictive performance. For this purpose, we utilize a specific test from the generic class of multiple comparisons procedures, namely, the *Scott-Knott* test [11], which ranks the models and partitions them into clusters.

The proposed statistical framework is applied on a relatively large-scale set of 11 methods over six public-domain datasets from the *PROMISE* repository [12] and the *International Software Benchmarking Standards Group* (ISBSG) [13]. Finally, in order to address the disagreement on the performance measures, we apply the whole analysis on three functions of error that measure different important aspects of prediction techniques: *accuracy*, *bias*, and *spread* of estimates.

The rest of the paper is organized as follows: In Section 2, we summarize related work and we specify the contribution of the current study. In Section 3, we present the limitations of current approaches for multiple comparisons of models and we analytically describe the proposed procedure based on the Scott-Knott test. In Section 4, we demonstrate the experimental setup of this study in a systematic manner, whereas in Section 5 we accumulate the results of the analysis. In Section 6, we perform some sensitivity analysis for two small datasets. Finally, certain limitations of the study are given in Section 7, whereas in Section 8 we conclude by discussing the novel findings of the proposed framework.

2 Related Work and Contribution

During the last decades there has been evolving research concerning the identification of the best SCE method [1]. Researchers strive to introduce prediction techniques including expert judgment, algorithmic, statistical, and machine learning methods. The usual practice of these studies is to compare the proposed estimation method with established models on a small number of datasets.

Earlier studies based their inference regarding the superiority of a prediction method against a comparative one on accuracy measures computed through a validation procedure. Despite the novelty and the promising results of each estimation technique, the researchers' interest has been rapidly focused on the problem of inconsistent findings regarding the determination of the best estimation method, while at the same time they started investigating the reasons behind unstable results.

Miyazaki et al. [14] claimed that the "de facto" MMRE accuracy measure tends to advance models that underestimate the actual effort ,while Kitchenham et al. [3] indicated the variation of accuracy measures as a primary source of inconclusive studies. Toward this direction, Foss et al. [4] investigated the basis of this criticism through a simulation study, proposing alternative accuracy indicators and concluding that there is need for applying well-established statistical procedures when conducting SCE experiments.

Myrtveit et al. [8] extended the above-mentioned findings and pointed out that inconsistent results are not caused only by accuracy measures but also by unreliable research procedures. Through a simulation study, they studied the consequences of three main ingredients of the comparison process: the single data sample, the accuracy indicator, and the cross-validation procedure. Furthermore, they provided possible explanations for the lack of convergence, such as the small sample size of many software studies and the splitting of training and test sets in the validation procedure that affects the comparison, even for samples drawn from the same populations. The researchers also inferred that the conclusions on "which model is best" to a large degree depend on the chosen accuracy indicator and that different indicators can lead to contradicting results.

Mittas and Angelis [5] showed that the usual practice of promoting a model against a competitive one just by reporting an indicator can lead to erroneous results since these indicators are single statistics of error distributions, usually highly skewed and nonnormal. In this regard, they proposed resampling procedures for hypothesis testing, such as permutation tests and bootstrap techniques for the construction of robust confidence intervals.

Recently, Menzies et al. [15] studied the problem of "conclusion instability" through the COSEEKMO toolkit that supports 15 parametric learners with row and column preprocessors based on two different sets of tuning parameters. In order to evaluate the predictive power of the alternative models, they used performance ranks, whereas the selection of the best method was based on a heuristic metric. Their experiments on COCOMO-style datasets concluded that there were four best methods and not just a single option.

The Scott-Knott test presented here was used in another context in [16], for combining classifiers applied to large databases. Specifically, the Scott-Knott test and other statistical tests were used for the selection of the best subgroup among different classification algorithms and the subsequent fusion of the models' decisions in this subgroup via simple methods, like weighted voting. In that study extensive experiments with very large datasets showed that the Scott-Knott test provided the highest accuracy in difficult classification problems. Hence, the choice of the test for the present paper was motivated by former results obtained by one of the authors.

In [17], Demšar discusses the issue of statistical tests for comparisons of several machine learning classifiers on multiple datasets reviewing several statistical methodologies. The method proposed as more suitable is the nonparametric analogue of ANOVA, i.e., the Friedman test, along with the corresponding Nemenyi post hoc test. The Friedman test ranks all the classifiers separately for each dataset and then uses the average ranks of algorithms to test whether all classifiers are equivalent. In case of differences, the Nemenyi test performs all the pairwise comparisons between classifiers to identify the significant differences. This method is used by Lessmann et al. [9] for the comparison of classifiers for prediction of defected modules. The methodology described in our paper, apart from the fact that is applied to a different problem, i.e., the SCE where cost and prediction errors are continuous variables, has fundamental differences regarding the goals, the way it is used, and the output.

Specifically, the algorithm we propose ranks and clusters the cost prediction models based on the errors measured for a particular dataset. Therefore, each dataset has its own set of "best" models. This is more realistic in SCE practice since each software development organization has its own dataset and wants to find the models that best fit its data rather than trying to find a globally best model which is unfeasible. Furthermore, the clustering as an output is different from the output of pairwise comparisons tests, like the Nemenyi test. A pairwise test, for example, can possibly indicate that models A and B are equivalent, models B and C are also equivalent, but models A and C are different. The grouping of model B is therefore questionable. For larger numbers of models the overlapping homogeneous groups resulting from pairwise tests are ambiguous and problematic in interpretation. On the other hand, a ranking and clustering algorithm provides clear groupings of models, designating the group of best models for a particular dataset.

The goal of this paper is to further extend the research concerning the comparison and ranking of multiple alternative SCE models. We propose a framework for conducting comparative experiments and we present an evaluation of this analysis over different datasets and prediction models. The framework addresses problems caused by different sources of bias, inherent in comparative studies, toward the following directions:

- Alternative prediction models ranging from regression approaches to analogy-based and machine learning techniques are studied in order to cover a wide range of estimation methods proposed so far in the literature.
- Public-domain datasets with different characteristics are used in order to address the inherent problem of prediction systems, i.e., their high dependency on the types of data.
- Alternative error functions measuring different important aspects of error are studied.
- Comparison of models under a special type of DOE, the *Repeated Measures Design*, which is analyzed similarly to the *Randomized Complete Block Design* (RCBD). These designs account for possible heterogeneity in experimental circumstances that can be caused by the effect of splitting the training and the test sets during the validation procedure.
- An algorithm based on the multiple comparison Scott-Knott test ranks and clusters prediction models by indentifying statistically significant differences between them.

3 Limitations of Established Procedures and Description of the Proposed Methodology

In this section, we present the main aspects of the proposed methodology. First, we discuss problems of the procedures used to compare a set of candidate prediction methods. Second, we describe in detail the algorithm based on the Scott-Knott test, which addresses the limitations of the established techniques.

3.1 Problems with Comparison of Multiple Prediction Techniques

The important role of well-established statistical comparisons in SCE is highlighted in many recent studies, especially during the last decade (see, indicatively, [18], [19], and [20]), where the findings are derived through formal statistical hypothesis testing. Indeed, the researchers use parametric as well as nonparametric procedures, whereas there has also been increasing interest [7] for more robust statistical tests such as permutation tests and bootstrapping techniques for the construction of confidence intervals [5].

The problem we address in this paper belongs to a generic class in statistics known as "*multiple hypothesis testing*" and can be defined as the procedure of testing more than one hypothesis simultaneously [21].

Briefly describing the problem, the conclusions derived from a statistical hypothesis test are always subject to uncertainty. For this reason, we specify an acceptable maximum probability of rejecting the null hypothesis when

TABLE 1
Measures of Error

$AE_i = \left\| Y_{A_i} - Y_{E_i} \right\|$	$z_i = \frac{Y_{E_i}}{Y_{A_i}}$	$MRE_i = \frac{\left\| Y_{A_i} - Y_{E_i} \right\|}{Y_{A_i}}$

TABLE 2
Global Measures of Error

$MAE = \frac{1}{n}\sum_{i=1}^{n} AE_i$	$Meanz = \frac{1}{n}\sum_{i=1}^{n} z_i$	$MMRE = \frac{1}{n}\sum_{i=1}^{n} MRE_i$

it is true and this is referred to as a "*Type I error*" [22]. In the case of multiple comparison problems, when several hypotheses are carried out, the probability that at least a Type I error occurs increases dramatically with the number of hypotheses.

The problem can be easily described through the following example. Suppose, we wish to compare five hypothetical models, setting the significance level at $a = 0.05$. For each test of an overall set of 10 pairwise tests, the probability of making a Type I error equals $a = 0.05$ and therefore the probability of not making a Type I error equals $1 - 0.05 = 0.95$. In the case of 10 comparisons the probability of no Type I error is $0.95^{10} \approx 0.60$. So, with a level of $a = 0.05$ for each of the 10 tests, the probability of erroneously rejecting a null hypothesis is 0.40.

The problem of this error inflation can be treated by adjusting the overall (or family) error, but still the execution of a large number of pairwise comparisons is not so straightforwardly interpretable since the resulting groupings are overlapping. For this reason, various intelligent techniques have been proposed in the statistical literature in order to perform targeted multiple comparisons, i.e., comparisons that answer specific research questions. For a thorough treatment of the multiple comparisons problems we refer to the classic book by Hochberg and Tamhane [23]. The problem of ranking and clustering cost estimation models according to their accuracy can be handled by the algorithm, which is based on the Scott-Knott test, that we describe in Section 3.3.

Apart from the obvious advantage of the Scott-Knott test to produce nonoverlapping groups, researchers at the Universidade Federal de Lavras, Brazil, have evaluated the power of the test (i.e., the probability that the test will reject the null hypothesis when the null hypothesis is actually false) using Monte Carlo simulations. Their studies [24] and [25], reviewed also in [26], have shown that for the Scott-Knott test the comparison-wise and the experiment-wise error rates (i.e., the number of Type I errors divided by the number of comparisons and experiments, respectively) are in accordance with the predefined significance level while the power is high. Comparatively with other post hoc tests like the Tukey and the Newman-Keuls tests, the Scott-Knott test was found to be more powerful and robust under a wide variety of experimental situations and assumptions. In some cases the power of Scott-Knott test was eight times higher than the Tukey test. Furthermore, the test has been chosen for two recent studies [27] and [28] due to its robustness, low error rates, and discriminative power.

3.2 Experimental Design

The whole study is organized according to the formal principles of the *experimental design* or *design of experiment* (DOE). DOE refers to a systematic planning in order to maintain control over all factors that may affect the result of an experiment [10]. We have to point out that the term "experimental design" is often used informally in SCE by researchers to generally describe the conditions, the assumptions, the data, and the procedures of a study.

DOE constitutes an entire branch area in statistics involving fundamental concepts that have to be specified and controlled in advance. The basic element of a DOE is the *experimental unit*, which is the "object" on which the researcher wishes to measure a *response variable*. The purpose is to study the effect of one or more factors (categorical variables) on the response variable. The different categories of a factor are known as levels or *treatments*.

In the case of our experimental setup, the predictive performance of each competitive model is evaluated through a *k-fold cross-validation* approach in which the original dataset is randomly partitioned into k subsamples of equal size. During a repeated procedure, each one of the subsamples is considered as the validation sample (test set) and the remaining $k - 1$ subsamples as the training sets used for fitting the models. Then, the local measures of error (Table 1) are computed for each project of the test set. Each test set gives a global measure of error (Table 2). The error measures are discussed in detail in Section 4.3. The error measures from all the experimental runs are transformed so as to be normally distributed. These values are used as an input for the Scott-Knott algorithm. In our study, we used $k = 10$ as the number of folds.

Following the terminology of a DOE, the $k = 10$ different folds of the above-mentioned procedure can be considered as the experimental units of our context; the comparative prediction models represent the treatments and the response variable is the normalized expression of a measure in Table 2. The purpose of the experiment is to investigate the effect of different treatments (models) on a response variable (error measure), i.e., to test the differences of the predictive performance of different models.

Depending on the number of factors and the sources of variation that can be accounted for, on the response variable it is important to identify the most appropriate DOE technique that has to be adopted, i.e., the arrangement of the factors that will provide statistically valid results. The statistical literature offers a plethora of DOEs which can be utilized in different problems under certain conditions [10].

One problem that came up in SCE experiments is the known or suspected variation due to the split of the training and test sets. Myrtveit et al. [8] claim that the conclusions about the best prediction model are, to some extent, dependent on the validation procedure that is followed in a study, whereas Shepperd and Kadoda [2] also argue that the results frequently differ between the pairs of training sets that have been chosen from the underlying dataset.

TABLE 3
Randomized Complete Block Design of Experimentation

	Prediction Model			
Block	A	B	C	D
Test set 1	MeanE_{1A}	MeanE_{1B}	MeanE_{1C}	MeanE_{1D}
Test set 2	MeanE_{2A}	MeanE_{2B}	MeanE_{2C}	MeanE_{2D}
Test set 3	MeanE_{3A}	MeanE_{3B}	MeanE_{3C}	MeanE_{3D}
Test set 4	MeanE_{4A}	MeanE_{4B}	MeanE_{4C}	MeanE_{4D}
Test set 5	MeanE_{5A}	MeanE_{5B}	MeanE_{5C}	MeanE_{5D}
Test set 6	MeanE_{6A}	MeanE_{6B}	MeanE_{6C}	MeanE_{6D}
Test set 7	MeanE_{7A}	MeanE_{7B}	MeanE_{7C}	MeanE_{7D}
Test set 8	MeanE_{8A}	MeanE_{8B}	MeanE_{8C}	MeanE_{8D}
Test set 9	MeanE_{9A}	MeanE_{9B}	MeanE_{9C}	MeanE_{9D}
Test set 10	MeanE_{10A}	MeanE_{10B}	MeanE_{10C}	MeanE_{10D}

From what we have discussed above, it is essential for the researchers to take into account in their inferences for the superiority of a prediction technique the additional variability that can be caused from the heterogeneity of the validation sets. This is the reason why we decided to employ a specific type of DOE, namely, the *Repeated Measures Design*, which is equivalent to the *Randomized Complete Block Design* [10].

The RCBD setup incorporates an additional factor, the so-called *block,* which takes into account the grouping of similar experimental units. The incorporation of this extra factor is considered advantageous in order to identify true differences between treatments or, equivalently, the true treatment effect. Indeed, when different treatments are applied to similar (or the same) experimental units which form, in any sense, a block, there is a source of variation between blocks which cannot be explained by the difference between treatments. This source of variation is represented by the block factor that is considered in the analysis. In our context, the splitting of data into different training-test pairs represents the blocking factor, i.e., each block is a specific pair of a training-test subsets, where all models are applied and validated.

Illustrating all the notions described above, suppose that one wishes to compare the prediction performance of four hypothetical models (A, B, C, and D) based on the evaluation of an error function E. The process of the 10-fold cross validation for each comparative model results in 10 values of global error (Mean$E$). Hence, the RCBD adopted in our experimental setup can be explicitly described in Table 3.

3.3 The Scott-Knott Test

The *Scott-Knott* test [11] is a multiple comparison procedure based on principles of cluster analysis. The clustering refers to the treatments (*methods* or in our case *models*) being compared and not to the individual cases, while the criterion for clustering together treatments is the statistical significance of differences between their mean values [21], [23]. Our preference for the Scott-Knott test relies on a specific desirable characteristic of the method, i.e., that it is able to separate the models into nonoverlapping groups. In our case, the values of the response variable that is affected by the models are translated to expressions of the prediction errors derived from the models being compared. The algorithm we describe next is therefore able to rank and cluster prediction models according to their accuracy.

Suppose that we want to compare the predictive accuracy of d alternative models on a specific dataset through the utilization of a functional expression of the error e. We also assume that, following a standard validation procedure, there are k pairs of training-test datasets, i.e., the original dataset is divided k times in training test subsets. All models are applied repeatedly to each one of these pairs, i.e., each model is trained and validated using the same pair. The predictions of the test dataset yield an overall measure of accuracy which is the value of the response variable for the unique combination of a model and a training-test dataset. Therefore, for a specific dataset we have $d \cdot k$ measurements.

The original methodology, as described in [11] and [23], does not take into account the repetitive character of the experiment, i.e., it just considers that we have, for each model k, different measurements. The Scott-Knott procedure follows and uses the *one-way analysis of variance* (one-way ANOVA) which tests the null hypothesis that the treatment means are all the same or, equivalently, that there is no statistical difference between the means of the accuracy measures obtained by the compared models [29]. However, the alternative hypothesis is that the models can be partitioned into two mutually exclusive and collectively exhaustive nonempty subsets.

Obviously, when ANOVA shows that there is no statistical evidence for rejecting the null hypothesis, the comparative prediction models form a homogeneous group and cannot be clustered anymore. However, when ANOVA finds at least one significant difference, the Scott-Knott algorithm uses the ratio of the so-called "*Sum of Squares due to Error*" (SSE) and the corresponding degrees of freedom from ANOVA to estimate the variance σ^2 of the random error, i.e., the part of the response variable that cannot be explained by the treatment effect.

This property of the procedure (Step 4 below) which essentially uses the ANOVA table gives the opportunity to include in our analysis another factor, in addition to the treatment factor. In cases when all treatments are applied to each one of the experimental units, the DOE is known as *repeated measures design*. Since the differences among responses from different experimental units receiving the same treatment may be large, it is better to consider in our analysis one more factor accounting for the differences between units. Otherwise, the detection of real differences between treatments becomes problematic. The ANOVA for repeated measures designs is equivalent to the analysis of *randomized complete block designs*, where the experimental units are considered as blocks [29].

In our case, the experimental setup is a design where the d models are the treatments and the experimental units are the k training-test pairs derived from the original dataset. The models are all trained and tested to each one of the pairs. In this setup, the ANOVA table accounts for the *blocking effect,* produces different SSE and degrees of freedom so that the estimation of σ^2 is different.

Therefore, the following algorithm can be used either without considering blocks, as a single-factor analysis of experiment, or by considering the blocking effect. We believe that the blocking effect i.e., the effect of the training-test dataset pair, should be taken into account since in all cases we tested it was found statistically significant.

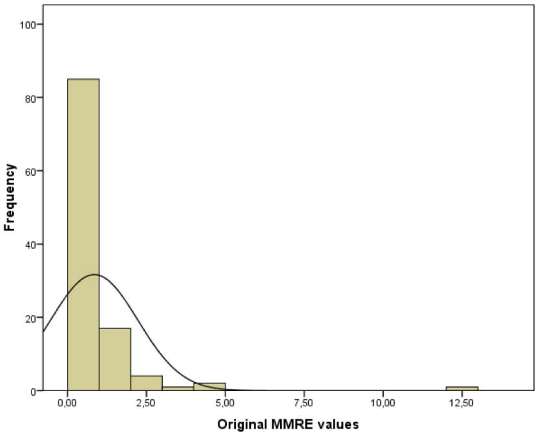

Fig. 1. Original MMRE values for the KEM dataset.

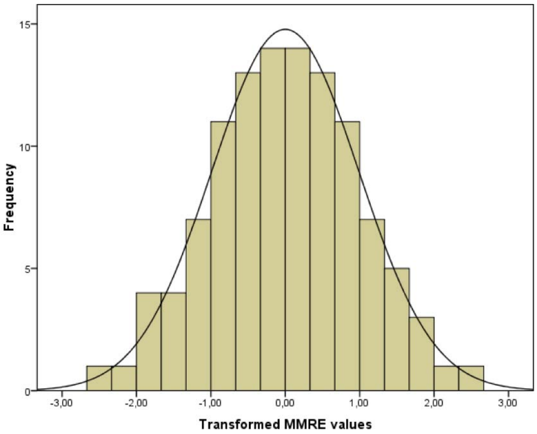

Fig. 2. Tranformed MMRE values for the KEM dataset.

The procedure we propose consists of consecutive steps aiming at a maximum differentiation between groups at each stage. Each group that is formed can be partitioned again if the new groups are significantly different. The whole process is fully described by the following steps:

1. Sort the means of the error measures $\bar{e}_j$, $j = 1, \ldots, d$, for each model in ascending order:

$$\bar{e}_{(1)} \leq \bar{e}_{(2)} \leq \cdots \leq \bar{e}_{(d)}. \tag{1}$$

2. For each $\bar{e}_{(j)}$, $j = 1, \ldots, d-1$, separate the group of all ordered means E into two subgroups $E_1 = \{\bar{e}_{(1)}, \ldots, \bar{e}_{(j)}\}$ and $E_2 = \{\bar{e}_{(j+1)}, \ldots, \bar{e}_{(d)}\}$ and compute the between-groups sum of squares:

$$G_j = k(|E_1|(\bar{e}_{E_1} - \bar{e}_E)^2 + |E_2|(\bar{e}_{E_2} - \bar{e}_E)^2), \tag{2}$$

where $|E_1|$, $|E_2|$ are the cardinalities of the two subgroups and $\bar{e}_E, \bar{e}_{E_1}, \bar{e}_{E_2}$ are the means of groups E, E_1 and E_2, respectively:

$$\begin{aligned} \bar{e}_E &= \frac{1}{d}\sum_{j=1}^{d} \bar{e}_{(j)}, \\ \bar{e}_{E_1} &= \frac{1}{|E_1|}\sum_{j \in E_1} \bar{e}_{(j)}, \bar{e}_{E_2} = \frac{1}{|E_2|}\sum_{j \in E_2} \bar{e}_{(j)}. \end{aligned} \tag{3}$$

3. Find the partition that maximizes the value of the above sum of squares:

$$G_{j^*} = \max\{G_j, j = 1, \ldots, d\} \tag{4}$$

4. Compute from the ANOVA table the s^2, the estimation of σ^2 (the variance that cannot be explained by the factors, i.e., the treatments and the blocks), by dividing the SSE by the corresponding degrees of freedom. Next, compute the statistic:

$$\lambda = \frac{\pi}{2(\pi - 2)} \frac{G_{j^*}}{s^2}, \tag{5}$$

which has approximately a χ^2_ν distribution where the degrees of freedom are given by $\nu = k/(\pi - 2)$ (rounded).

5. If $\lambda > \chi^2_{\nu;a}$ (where α is a predefined significance level), then the same test is applied to each group separately. If $\lambda < \chi^2_{\nu;a}$, then all means belong to the same homogeneous group. The procedure is continued by splitting each group into two subgroups if the λ-criterion is significant; otherwise, by identifying a homogeneous group until no groups can be split further.

The criterion for splitting the groups is based on the assumption that the distribution of λ in (5) is approximately a chi-square distribution. This is a theoretical result of another fundamental assumption, i.e., that the error measures yielded in the experiment are normally distributed. However, experience shows that the distributions of errors are usually skewed and certainly not normal. Therefore, it is necessary to apply a transformation on them in order to normalize the error values before we apply the algorithm described above. For this, we applied Blom's transformation [30] that utilizes the ranks r_i of the error values e_i and the inverse of the cumulative normal distribution function $\Phi^{-1}(x)$.

The formula for the Blom transformation is

$$\Phi^{-1}\left(\frac{r_i - 3/8}{n + 1/4}\right). \tag{6}$$

The transformation produces normally distributed numerical values, which are used instead of the original error measures as input for the algorithm. As an example, we present the original (Fig. 1) and the transformed (Fig. 2) values of MMRE for the KEM dataset. It is important to note that the Blom transformation is monotonous and therefore the order of the values is kept intact. The output of the algorithm is a ranking of the models according to their transformed error measures and, moreover, a clustering scheme where each cluster consists of the sorted models that do not have significant difference in their error measures.

It is also important to note that α, i.e., the predefined significance level, is not the nominal error rate of the Type I error of the whole algorithm (see [26]). If we set $\alpha = 0.05$, then each decision of the Scott-Knott procedure to divide or not a subgroup has a Type I error rate of 5 percent, but the overall Type I error rate of the algorithm depends on the number of tests performed.

TABLE 4
Prediction Models of the Comparative Experimentation

Prediction Method	General Idea of Method
Regression-based	The relationship between dependent and independent variables is expressed as a known explicit function with unknown parameters (the regression coefficients). Then, the parameters are estimated in such a way so as to minimize a predefined criterion.
Ordinary Least Squares Regression (OLS)	Minimize the sum of squared residuals [31].
Least Median of Squares Regression (LMS)	Minimize the median of squared residuals [3].
Least Trimmed Squares Regression (LTS)	Minimize the sum of squared residuals over a subset of points [32].
Robust M-estimator Regression (RobMM)	Minimize an M-estimator [33].
Analogy-based	Mimics the human instinctive decision-making by comparing with similar cases. A type of non-parametric regression procedure (i.e. the relationship is not expressed in the form of an explicit function), where the unknown values of the dependent variable are estimated by the known values, of the same variable, corresponding to neighbours (analogies) of the estimated case.
Estimation by Analogy (EbA)	The classical approach without adjustment [34].
Estimation by Analogy with first adjustment (EbA_Adj1)	Performs a linear size adjustment to the estimated effort [35].
Estimation by Analogy with first adjustment (EbA_Adj2)	Performs a similarity adjustment solution function to the estimated effort [36].
Machine Learning	A branch of artificial intelligence techniques that are concerned with the construction of a relationship between dependent and independent variables based on empirical data and past experience.
Bagging (Bag)	Generates multiple versions of a prediction function based on EbA using these to get an aggregated function [37].
Neural Network (NN)	Inspired by the structure and functional aspects of human neural networks. The study utilizes a multilayer perceptron which represents a relationship between dependent and independent variables with a network structure consisting of a layer of nodes in a directed graph [38].
Naïve Bayes Classifier (NB)	Applies Bayes' theorem to construct a probabilistic classifier based on the independence assumption between variables [39].
Classification and Regression Tree (CART)	Constructs a tree-structured decision process for predicting the cost of a project by recursively partitioning according to splitting rules [40].

4 Experimental Setup

This section provides details concerning the setup of the framework and the experimental design of the study. The basic idea of the experimental setup was to take into account: 1) different cost prediction methods covering a major part of the variety of the proposed methodologies that have appeared so far in the SCE literature and which are governed by a diversity of principles, 2) different datasets, and 3) different measures of error. Moreover, the experiment was designed to take into account the effect of training-test splitting of each dataset.

4.1 Comparative Prediction Models

The 11 selected methods can be grouped into three main categories that are *regression-based* models, *analogy-based* techniques, and *machine learning* methods.

All these models are well-established methods, there is a vast literature on them, and they have been already applied in SCE. In Table 4, we present a list of the methods being compared along with a brief description in order to give the basic principles of each method.

The choice of the alternative prediction techniques was also based on the conclusions of a systematic review on SCE studies. Jorgensen and Shepperd [1] pointed out that the regression-based models dominate since half of all studies deal with the problem of fitting, improvement, or comparison of a regression model. Furthermore, they claim that the researchers' interest for the analogy-based techniques is steadily increased during the last decade. Finally, the distribution of estimation methods also reveals that the proportion of machine learning techniques (i.e., Classification and Regression Trees and Neural Networks) presents an increasing trend.

It is obvious that the prediction techniques used in our experimentation presuppose the tuning of certain parameters in order to build meaningful models. For example, the ratio-scaled variables of regression-based models are checked in order to investigate whether the normality assumption is satisfied, whereas the nominal and ordinal variables are replaced with new dummy variables and then a stepwise procedure is adopted to extract the most significant independent variables. In the case of analogy-based methods, we use the Kaufman-Rousseeuw dissimilarity measure [41], taking into account various types of variables, whereas the selection of the best number of the "*neighbor*" projects is determined through the *leave-one-out cross-validation* procedure [37]. The exclusion of irrelevant variables is based on published studies ([34], [42], [43]) that utilize a brute-force algorithm.

Regarding neural network models, we specify the number of nodes for hidden layers [38]. In particular, we adopted the RMiner for NN [44], a library for R language that facilitates the use of data mining techniques in classification and regression tasks. The network includes one hidden layer of H neurons and the output neuron uses a linear function. In RMiner, the NN hyperparameter H is optimized using a grid search with a backward selection algorithm, whereas to avoid overfitting, an internal k-fold process is used. After selecting the best parameter, the model is retrained with all training data.

TABLE 5
Datasets Characteristics

Name	#projects	#categorical	#scale
Kemerer (KEM)	15	-	4
Telecom (TEL)	18	-	2
Albrecht (ALB)	24	-	7
Miyazaki (MIY)	48	-	7
Desharnais (DES)	77	1	7
ISBSG10 (ISBSG)	506	-	10

As far as the CART model is concerned, we utilize the Recursive Partitioning [45] algorithm as implemented in S-PLUS [46] in which the model is fitted using binary recursive partitioning whereby the data are successively split along coordinate axes of the predictor variables so that at any node, the split which maximally distinguishes the response variable in the left and the right branches is selected. Splitting continues until nodes are pure or data are too sparse, according to the recommendations of S-PLUS manual [46].

Finally, for the case of the Naive Bayes classifier, the methodology computes the conditional a-posterior probabilities of the dependent variable given the independent predictors using the Bayes rule [39], whereas a stepwise procedure for the selection of the best subset of attributes is also used.

4.2 Datasets

The datasets for the experimentation are derived from two sources, namely, the *PROMISE* repository [12] and the *International Software Benchmarking Standards Group* (ISBSG, release 10) [13]. The main reason for this selection was that these datasets have been extensively used to empirically validate or justify a large amount of research results, whereas they are also publically available. Each dataset contains a different number of projects and a set of independent variables with mixed-type characteristics (Table 5), whereas the dependent variable that has to be predicted is the *actual effort*. Another criterion for the selection of the datasets was the ability to apply all the competitive prediction methods on them. Therefore, we did not consider datasets with too many categorical variables which cause problems to certain methods like regression and Neural Networks.

The ISBSG repository contains 4,106 software projects from more than 25 countries, but most of the variables have a large amount of missing values. Keeping in mind the guidelines of ISBSG suggesting filtering of the data projects, we decided to discard the projects with missing values. Moreover, an important issue in SCE is the utilization of datasets with high quality in the process of evaluation and comparison of prediction models [47]. Due to this fact and the instructions of the ISBSG organization that point out not taking into account projects with low quality, we based our analysis on projects marked with "A" in *Data Quality Rating* and *UFP Rating*. Finally, the independent characteristics utilized in the construction of the alternative models are the same as in [48] in order to retain the compatibility with other studies.

4.3 Accuracy Measures

In Section 2, we have already mentioned that the overwhelming majority of SCE studies base their inferences on MMRE. However, MMRE has been criticized as an inappropriate measure since it tends to favor models that underestimate the actual cost of projects [14].

During the last decade there has been a thorough discussion concerning the determination of the most appropriate error function without essential convergence. Having in mind that all the accuracy indicators exhibit certain advantages, while at the same time they suffer from either a flaw or a limitation [4], the basic key for resolving the problem is to realize what is really measured by each error function [3].

The lesson learned from the above-mentioned discussion is that there is a need for utilization of three different error functions measuring three aspects of the prediction performance of comparative models. More precisely, *Absolute Error* (AE) is used in order to evaluate the *accuracy* of models, whereas error ratio z has been adopted as a measure of *bias* accounting for underestimations ($z < 1$) or overestimations ($z > 1$) with an optimum value of 1 [3]. The most widely known MRE indicator was also used since, according to Kitchenham et al. [3], it provides a measure of the *spread* of the error ratio z.

The local measures of error (Table 1) that are computed through the actual (Y_A) and the estimated (Y_E) values of each single project i constitute the basis for the evaluation of the overall prediction performance of the comparative models by computing a statistic (i.e., mean) for a set of n test cases (Table 2).

5 Experimental Results

In this section, we present the results of the experiments conducted on six datasets. Tables 6 and 7 summarize the results of all experimental runs. Each dataset corresponds to a different DOE where each model was applied to 10 test sets and then the values of the error measures were transformed and used as an input for the Scott-Knott algorithm. So, the ranking was based on the means of the transformed values and not the means of the original values of the global measures.

Table 6 presents the results of the ANOVA procedure on which the clustering algorithm was based. We can see that for each dataset and error measure we have two columns, labeled randomized complete block design and complete randomized design (CRD). The first one takes into account the effect of the models (treatment effect) and the blocking factor (block effect), while the second only the model effect. For each one of the ANOVA tests, we report the *significance* (p-value) and, in parentheses, the *eta-squared* statistic, which provides an estimate of the effect size in ANOVA. The effect size is a measure of the importance of a result in the population and for the eta-squared the benchmark values are: 0.01 (small), 0.06 (medium), and 0.14 (large) [49].

The columns of Table 7 contain the output of the algorithm for each different dataset under a specific global measure of error. The models are ranked starting from the best to the worst, i.e., lower rank means better model. Models that do not have significant difference are in consecutive cells of the table, shaded with the same color. So the shading of the cells represents the clustering of the models. The ranking of models is the same for both designs

TABLE 6
Significance Values and Effect Size (Eta-Squared) of ANOVA

		MMRE		MAE		Meanz	
Dataset	*Effect size and p-value*	*RCBD*	*CRD*	*RCBD*	*CRD*	*RCBD*	*CRD*
KEM	Treatment Effect: *p*-value (eta-sq)	0.438 (0.101)	0.804 (0.058)	0.117 (0.152)	0.787 (0.060)	0.415 (0.104)	0.977 (0.030)
	Block Effect: *p*-value (eta-sq)	<0.001 (0.455)	-	<0.001 (0.645)	-	<0.001 (0.0.731)	-
TEL	Treatment Effect: *p*-value (eta-sq)	0.048 (0.172)	0.549 (0.082)	0.009 (0.222)	0.583 (0.079)	0.966 (0.037)	0.999 (0.014)
	Block Effect: *p*-value (eta-sq)	<0.001 (0.568)	-	<0.001 (0.699)	-	<0.001 (0.624)	-
ALB	Treatment Effect: *p*-value (eta-sq)	<0.001 (0.310)	0.002 (0.240)	<0.001 (0.329)	0.030 (0.176)	0.275 (0.121)	0.817 (0.056)
	Block Effect: *p*-value (eta-sq)	<0.001 (0.327)	-	<0.001 (0.565)	-	<0.001 (0.567)	-
MIY	Treatment Effect: *p*-value (eta-sq)	0.149 (0.144)	0.627 (0.075)	0.197 (0.134)	0.959 (0.035)	0.052 (0.176)	0.422 (0.094)
	Block Effect: *p*-value (eta-sq)	<0.001 (0.517)	-	<0.001 (0.762)	-	<0.001 (0.511)	-
DES	Treatment Effect: *p*-value (eta-sq)	<0.001 (0.296)	0.047 (0.148)	0.002 (0.254)	0.829 (0.055)	0159. (0.141)	0.727 (0.066)
	Block Effect: *p*-value (eta-sq)	<0.001 (0.585)	-	<0.001 (0.829)	-	<0.001 (0.574)	-
ISBSG	Treatment Effect: *p*-value (eta-sq)	<0.001 (0.764)	<0.001 (0.387)	<0.001 (0.320)	0.577 (0.080)	<0.001 (0.840)	<0.001 (0.501)
	Block Effect: *p*-value (eta-sq)	<0.001 (0.805)	-	<0.001 (0.816)	-	<0.001 (0.808)	-

(RCBD and CRD), but the clustering is not necessarily the same. The ticks in the CRD columns indicate exactly this fact, i.e., that the ranking is exactly the same, so there is no need to write the names of the models again.

The last column of Table 7 presents the mean rank (MR) for each model over all datasets. Moreover, the models are ranked according to their overall performance in all datasets.

As we can observe from Table 7, it is clear that the common belief that different dataset characteristics may favor different prediction models is confirmed in our experiments. Generally, the prediction performance of each model varies from dataset to dataset and it is impossible to extract a global conclusion. Overall, OLS and CART seem to be the best-rated and worst-rated estimation techniques since they have the smallest and the highest values of MRs, respectively.

An interesting conclusion is also derived from the three indicators of error within each dataset. Table 7 suggests that there is a ranking instability of different models across the same dataset. For example, based on the results of ISBSG dataset, NB presents the best performance in terms of MMRE and Mean*z*, but the second worst for the

TABLE 7
Experimental Results

	KEM		TEL		ALB		MIY		DES		ISBSG		*MR*
	RCBD	***CRD***	***RCBD***	***CRD***	***RCBD***	***CRD***	***RCBD***	***CRD***	***RCBD***	***CRD***	***RCBD***	***CRD***	
MMRE													
Models	OLS	✓	EbA_Adj1	✓	LTS	✓	OLS	✓	LMS	✓	NB	✓	**OLS (2.00)**
	NN	✓	OLS	✓	OLS	✓	EbA_Adj1	✓	LTS	✓	OLS	✓	RobMM (3.33)
	RobMM	✓	RobMM	✓	LMS	✓	RobMM	✓	RobMM	✓	LMS	✓	LMS (4.83)
	EbA_Adj1	✓	EbA	✓	RobMM	✓	NB	✓	OLS	✓	RobMM	✓	LTS (5.17)
	Bag	✓	EbA_Adj2	✓	NN	✓	EbA	✓	NN	✓	LTS	✓	EbA_Adj1 (5.17)
	LMS	✓	LMS	✓	NB	✓	NN	✓	Bag	✓	NN	✓	NN (5.67)
	EbA	✓	LTS	✓	Bag	✓	Bag	✓	EbA_Adj2	✓	EbA_Adj1	✓	EbA (7.00)
	LTS	✓	Bag	✓	EbA_Adj1	✓	LTS	✓	EbA	✓	EbA	✓	Bag (7.00)
	EbA_Adj2	✓	CART	✓	EbA_Adj2	✓	CART	✓	EbA_Adj1	✓	Bag	✓	NB (7.17)
	CART	✓	NN	✓	EbA	✓	LMS	✓	NB	✓	EbA_Adj2	✓	EbA_Adj2 (8.50)
	NB	✓	NB	✓	CART	✓	EbA_Adj2	✓	CART	✓	CART	✓	CART (10.17)
MAE													
Models	OLS	✓	EbA_Adj1	✓	OLS	✓	OLS	✓	OLS	✓	OLS	✓	**OLS (1.17)**
	NN	✓	OLS	✓	LTS	✓	NB	✓	RobMM	✓	RobMM	✓	RobMM (2.83)
	RobMM	✓	RobMM	✓	LMS	✓	RobMM	✓	LMS	✓	LMS	✓	EbA_Adj1 (5.50)
	EbA_Adj1	✓	EbA	✓	RobMM	✓	EbA	✓	LTS	✓	EbA	✓	LMS (5.50)
	Bag	✓	EbA_Adj2	✓	NB	✓	EbA_Adj1	✓	NN	✓	EbA_Adj2	✓	LTS (5.83)
	LMS	✓	LTS	✓	EbA_Adj1	✓	Bag	✓	Bag	✓	Bag	✓	EbA (6.00)
	EbA	✓	LMS	✓	EbA_Adj2	✓	LTS	✓	EbA_Adj2	✓	NN	✓	EbA_Adj2 (6.83)
	LTS	✓	Bag	✓	EbA	✓	EbA_Adj2	✓	EbA_Adj1	✓	LTS	✓	Bag (6.83)
	EbA_Adj2	✓	CART	✓	NN	✓	CART	✓	EbA	✓	EbA_Adj1	✓	NN (7.33)
	CART	✓	NB	✓	Bag	✓	NN	✓	NB	✓	NB	✓	NB (8.00)
	NB	✓	NN	✓	CART	✓	LMS	✓	CART	✓	CART	✓	CART (10.17)
Meanz													
Models	RobMM	✓	NB	✓	NB	✓	NB	✓	RobMM	✓	NB	✓	**OLS (2.67)**
	OLS	✓	OLS	✓	RobMM	✓	OLS	✓	LTS	✓	OLS	✓	NB (3.00)
	EbA_Adj2	✓	EbA	✓	LMS	✓	EbA	✓	OLS	✓	LMS	✓	RobMM (3.67)
	NN	✓	LTS	✓	LTS	✓	EbA_Adj1	✓	NB	✓	RobMM	✓	LMS (5.67)
	EbA_Adj1	✓	LMS	✓	OLS	✓	Bag	✓	LMS	✓	LTS	✓	LTS (5.83)
	Bag	✓	EbA_Adj1	✓	EbA_Adj2	✓	EbA_Adj2	✓	NN	✓	EbA_Adj1	✓	EbA (5.83)
	EbA	✓	RobMM	✓	EbA	✓	RobMM	✓	EbA_Adj1	✓	EbA	✓	EbA_Adj1 (6.00)
	LMS	✓	EbA_Adj2	✓	EbA_Adj1	✓	CART	✓	EbA	✓	NN	✓	EbA_Adj2 (7.00)
	LTS	✓	CART	✓	Bag	✓	NN	✓	EbA_Adj2	✓	Bag	✓	NN (7.83)
	NB	✓	NN	✓	NN	✓	LMS	✓	Bag	✓	EbA_Adj2	✓	Bag (8.33)
	CART	✓	Bag	✓	CART	✓	LTS	✓	CART	✓	CART	✓	CART (10.17)

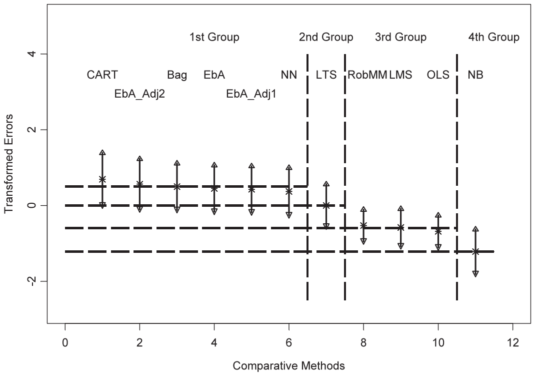

Fig. 3. Plot of the Scott-Knott algorithm (MMRE-ISBSG).

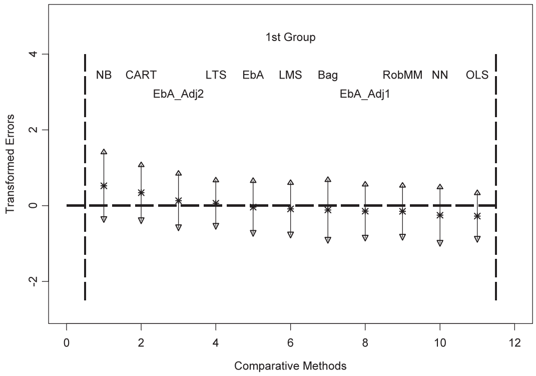

Fig. 4. Plot of the Scott-Knott algorithm (MMRE-KEM).

case of MAE. Hence, these findings reinforce the conclusions of other researchers (i.e., [8]), signifying the important role of the selection of accuracy indicator in comparative studies. Moreover, we also have to recall that different error functions quantify different aspects of prediction performance [3]. Subsequently, although there is a ranking instability for NB in the above-mentioned example, it is crucial to acquire the significant information that is hidden in this lack of convergence and to interpret the remarkable findings with caution. So, NB presents the least biased predictions (Mean$z$) with the smallest spread (MMRE) but at the same time NB suffers from inaccuracy (MAE) problems.

Although the rankings of models clearly portray an overview of their predictive performance, it is essential to statistically test whether some models are superior to others since these observed differences could reasonably occur "just by chance." Next, we present the results of the formal comparison of models through the inferential statistical procedures of RCBD and Scott-Knott test.

The first remarkable issue from Table 6 (RCBD columns) is that the blocking factor (choice of training-test datasets) shows a statistically significant effect ($p < 0.001$ and large effect size) for all the experiments. Thus, the choice to consider the variability caused by the splitting of data into training and test sets is justified. Otherwise, it is possible to derive erroneous results.

In order to illustrate this issue, except for the results of RCBD (Tables 6 and 7), we also present for each dataset the ANOVA results (Table 6) and the groups of homogeneous models (Table 7) in terms of prediction performance derived from the Scott-Knott algorithm without taking into account the blocking factor (columns labeled as CRD). We can clearly see that there are significant differences in the ANOVA results and the formation of groups between RCBD and CRD. For example, the Scott-Knott algorithm resulted in three groups of statistically different models in the case of the ISBSG dataset when MAE was examined and when the blocking factor was taken into consideration (RCBD). On the other hand, for CRD no significant differences were found, meaning that all the models presented similar prediction accuracy.

As we have already mentioned, in Table 7 significant differences in performances across the alternative models are indicated through differently shaded cells. As the shading of rows becomes darker, the performance of the group of the corresponding estimation models with the same color becomes worse. For example, in the case of the TEL dataset, the Scott-Knott test results in two homogeneous groups of estimation models for MMRE and MAE. The best group (white cells) includes all except for the last three models (CART, NB, and NN) that belong to the second group. An initial interesting conclusion is the case of the KEM dataset, in which the Scott-Knott test does not find any statistically significant difference among the competitive techniques.

The results of the Scott-Knott procedure are also presented in a graphical manner for two cases (Figs. 3 and 4). The diagram plots comparative models (x-axis) against the transformed mean errors (y-axis), whereby all methods are sorted according to their ranks. The vertical dashed lines indicate which models give statistically different results and thus are clustered into homogeneous group. For example, in Fig. 3, we present the results of the ISBSG dataset for the experiment with MMRE. The Scott-Knott algorithm resulted in four homogeneous groups of models with similar performances. Each small vertical solid line represents the prediction performance for each one of the competitive models and, more precisely, the asterisk depicts the mean value of the transformed error function, whereas the bounds of a 95 percent confidence interval for the mean derived from the 10 folds are shown through the lower and upper ends of the lines, respectively. Furthermore, the horizontal lines represent the mean transformed errors for each one of the four different groups. As we can also observe from Fig. 4 (referring to the KEM dataset and the MMRE), the whole analysis does not provide statistical evidence for different prediction performances among the comparative models and for this reason all models are clustered into one group.

Of course, the results of statistical tests should be interpreted using technical details of models and methods. For example, NB seems to produce inaccurate estimates for the case of the TEL dataset. This may be due to the philosophy of the NB technique since this specific model classifies the effort of a forthcoming project into predefined categories. Keeping in mind the small size of the TEL dataset (Table 5), we can easily observe that each category contains a small number of projects, which in turn may result in low predictive power of the method. Furthermore, the small size of projects in the KEM dataset can also be the most important factor leading the Scott-Knott test to not separate the means of error functions into nonoverlapping groups.

Generally, the application of Scott-Knott tests for all datasets reveals one of the most appealing findings of this study: Despite the large divergences of error functions among alternative prediction models, there is no statistical evidence that some methods differ significantly. Hence, the notion of the "best" estimation technique should be revised, whereas at the same time it is probably more proper to refer to the "best group of estimation techniques."

Moreover, the size of the samples in which the alternative prediction models are built plays a critical role in SCE experiments. Indeed, Table 7 shows that there is no evidence of statistical differences among alternative prediction models in small datasets (i.e., KEM, TEL, etc.).

This pattern in the results also brings to light a conviction concerning the vanity of using many estimation methods on small datasets since they cannot be more informative despite their promising and sophisticated principles.

Concerning the predictive power of the alternative models, there are also some interesting findings that are derived from this study.

First, it seems that OLS performs generally well for all datasets since this specific form of regression-based techniques gives the best ranking for seven out of 18 experiments (Table 7), whereas the MRs for the three global measures of error are the lowest. Furthermore, Scott-Knott tests suggest that OLS can be categorized into the best group of estimation techniques in all except two experiments (MMRE and Meanz of ISBSG dataset). This result is in accordance with the literature if we consider that the relation between software effort and size has been modeled as exponential in the sense that the natural logarithm of effort is expressed as a linear function of the logarithm of size (there is extensive literature on the form of the relation between size and effort; see, for example, [50]). Regarding the other regression-based models, although these forms (LTS, LMS, and RobMM) provide robust estimations when there are outliers in the data, in our experiments they present similar or even worse predictive performance than OLS. A possible explanation for this may be the lack of bad quality data points that can influence the accuracy of OLS.

Second, the predictive accuracy of analogy-based methods seems to be similar despite the adjustments that may be applied to the final estimation. This is clearly inferred from the results of Scott-Knott tests, where the analogy-based techniques are clustered together in the same group of methods for all experiments.

Finally, we can also notice that machine learning methods (Bag, NN, NB) provide accurate estimations for a few experiments, constituting an alternative choice to regression and analogy-based methods. On the other hand, it seems that CART is not able to catch the strong linear dependency between effort and size of projects and this may be the main reason for the poor predictive performance. This result is also consistent with the findings of Kitchenham [51], suggesting that only in the cases of a large number of projects contained in a dataset should a practitioner consider using CART.

6 Sensitivity Analysis

We have already mentioned in Section 3 that the validation procedure can play a significant role in the conclusions derived due to the added variability of splitting a dataset into a training-test pair [8], whereas the results can frequently differ. More recently, Menzies and Shepperd [52] also pointed out that the validation procedure may be a potential source of noise in the experiments since minor changes to the training data can lead to large conclusion instabilities in the internal parameters of prediction models.

TABLE 8
Sensitivity Analysis for Small Datasets (MMRE)

TEL			ALB		
k=10	*k*=3	*LOOCV*	*k*=10	*k*=3	*LOOCV*
EbA_Adj1	LTS	LMS	LTS	RobMM	LTS
OLS	OLS	EbA_Adj1	OLS	OLS	OLS
RobMM	RobMM	RobMM	LMS	LTS	LMS
EbA	EbA_Adj1	LTS	RobMM	LMS	EbA_Adj1
EbA_Adj2	NB	OLS	NN	NN	RobMM
LMS	LMS	EbA	NB	NB	EbA
LTS	EbA_Adj2	Bag	Bag	EbA_Adj1	EbA_Adj2
Bag	CART	EbA_Adj2	EbA_Adj1	EbA	NN
CART	EbA	CART	EbA_Adj2	Bag	Bag
NN	Bag	NN	EbA	EbA_Adj2	NB
NB	NN	NB	CART	CART	CART

Kohavi [53] suggested that if the prediction algorithm is stable for a given dataset, the variance of the cross-validation estimates should be approximately the same, independent of the number of folds. He also remarked that as the sample sizes get smaller and the number of folds decreases, the instability of the training sets leads to an increase in variance of the whole procedure recommending the 10-fold cross-validation scheme as it tends to provide less biased estimation of the accuracy.

Efron [54] conducted five sampling experiments and compared leave-one-out cross validation and several other methods in order to investigate some related estimators which seem to offer considerably improved estimation in small samples. The results indicate that leave-one-out cross validation gives nearly unbiased estimates of the accuracy, but often with unacceptably high variability, particularly for small samples.

Hence, the choice of the validation procedure is a critical topic affecting the derived conclusions of an experiment, whereas the choice of the number of folds can also contribute to the instability of the results, especially for small datasets. For this reason, we decided to perform some sensitivity analysis for two small datasets (ALB and TEL) in which the Scott-Knott procedure results in different homogeneous groups of estimation models for MMRE for the initial experiments. More precisely, we conduct the whole analysis for $k = 3$ and $k = n$. The latter approach is the special case of k-fold cross validation, where the number of folds equals the number of instances in the dataset, and is essentially equivalent to the *leave-one-out cross validation* (LOOCV), which takes advantage of the raw measurements of errors. It is important to emphasize here that the proposed algorithm is applicable to any validation schema.

The results of the Scott-Knott test for the three different numbers of folds ($k = 3, 10, n$) are given in Table 8. We also have to note that the ANOVA tests under the RCBD setup indicate that the effect of the blocking factor is statistically significant for all the experiments. Moreover, the findings of Table 8 show that changes to the splitting of training and test datasets lead to quite different results.

TABLE 9
Experimental Results with the TUKEY Post-Hoc Tests

	KEM	TEL	ALB		MIY	DES	ISBSG				
MMRE											
Models	Groups	Groups	Groups		Groups	Groups	Groups				
	OLS	EbA_Adj1	LTS		OLS	LMS	NB				
	NN	OLS	OLS		EbA_Adj1	LTS	OLS	OLS			
	RobMM	RobMM	LMS		RobMM	RobMM	LMS	LMS	LMS		
	EbA_Adj1	EbA	RobMM		NB	OLS	RobMM	RobMM	RobMM		
	Bag	EbA_Adj2	NN	NN	EbA	NN		LTS	LTS	LTS	
	LMS	LMS	NB	NB	NN	Bag		NN	NN	NN	
	EbA	LTS	Bag	Bag	Bag	EbA_Adj2		EbA_Adj1	EbA_Adj1	EbA_Adj1	
	LTS	Bag	EbA_Adj1	EbA_Adj1	LTS	EbA		EbA	EbA	EbA	
	EbA_Adj2	CART	EbA_Adj2	EbA_Adj2	CART	EbA_Adj1		Bag	Bag	Bag	
	CART	NN	EbA	EbA	LMS	NB			EbA_Adj2	EbA_Adj2	
	NB	NB		CART	EbA_Adj2	CART				CART	
MAE											
Models	Groups	Groups	Groups		Groups	Groups	Groups				
	OLS	EbA_Adj1	OLS		OLS	OLS	OLS				
	NN	OLS	LTS		NB	RobMM	RobMM				
	RobMM	RobMM	LMS	LMS	RobMM	LMS	LMS				
	EbA_Adj1	EbA	RobMM	RobMM	EbA	LTS	EbA				
	Bag	EbA_Adj2	NB	NB	EbA_Adj1	NN	EbA_Adj2				
	LMS	LTS	EbA_Adj1	EbA_Adj1	Bag	Bag	Bag				
	EbA	LMS	EbA_Adj2	EbA_Adj2	LTS	EbA_Adj2	NN				
	LTS	Bag	EbA	EbA	EbA_Adj2	EbA_Adj1	LTS				
	EbA_Adj2	CART	NN	NN	CART	EbA	EbA_Adj1				
	CART	NB	Bag	Bag	NN	NB	NB				
	NB	NN		CART	LMS	CART	CART				
Meanz											
Models	Groups	Groups	Groups		Groups	Groups	Groups				
	RobMM	NB	NB		NB	RobMM	NB				
	OLS	OLS	RobMM		OLS	LTS	OLS	OLS			
	EbA_Adj2	EbA	LMS		EbA	OLS	LMS	LMS	LMS		
	NN	LTS	LTS		EbA_Adj1	NB		RobMM	RobMM	RobMM	
	EbA_Adj1	LMS	OLS		Bag	LMS		LTS	LTS	LTS	LTS
	Bag	EbA_Adj1	EbA_Adj2		EbA_Adj2	NN			EbA_Adj1	EbA_Adj1	EbA_Adj1
	EbA	RobMM	EbA		RobMM	EbA_Adj1				EbA	EbA
	LMS	EbA_Adj2	EbA_Adj1		CART	EbA				NN	NN
	LTS	CART	Bag		NN	EbA_Adj2				Bag	Bag
	NB	NN	NN		LMS	Bag				EbA_Adj2	EbA_Adj2
	CART	Bag	CART		LTS	CART					CART

For the case of the TEL dataset, the Scott-Knott procedure results in two homogeneous groups of prediction models, but it is also clear and reasonable that there is a ranking instability for the different numbers of folds, whereas the homogeneous groups do not contain the same prediction models. On the other hand, it is worth noting that a subset of models (OLS, LTS, RobMM, and EbA_Adj1) is always included in the best group, despite the number of folds of the validation procedure. The results for the ALB dataset are even more representative of the prediction capability of the alternative models since three methods (OLS, LTS, and LMS) constitute a set of best estimation techniques. In conclusion, our results from sensitivity analysis are in accordance with former studies which recommend the use of different validation schemes when comparing models, especially when we use small datasets.

Another question that deserves some further investigation is why one should follow the analysis based on the Scott-Knott procedure and not perform one of the more known and, in a sense, more traditional post-hoc tests. Indeed, the literature on multiple comparisons [22], [23] contains a variety of post-hoc techniques, like the Bonferroni, Scheffe, Tukey, and Duncan tests, among many others.

The most significant drawback of the traditional post-hoc procedures is the fact that they typically produce overlapping groups. So, with a large number of treatments (prediction models in our case), the interpretation of the derived results is often difficult and ambiguous.

In order to illustrate the difficulties in the interpretation of the results in the multiple comparison problem of alternative prediction models, we conducted the whole experimentation using the traditional Tukey's post-hoc test. The outcomes given in Table 9 reveal that Tukey's test also results in homogeneous groups of prediction models for a few datasets and accuracy measures. On the other hand, it is also obvious that the prediction models are separated into overlapping groups, causing interpretation and decision problems regarding the superiority of a method against some other.

For example, the analysis of the ALB dataset for the case of MMRE indicates that there are two homogeneous groups of prediction models. The first "best" group includes a set of 10 models that are LTS, OLS, LMS, RobMM, NN, NB, Bag, EbA_Adj1, EbA_Adj2, and EbA. In the second group, there are many prediction models that are common with the first group, such as NN, NB, Bag, EbA_Adj1, EbA_Adj2, and EbA, whereas CART is the only model that is not included in the first group. On the other hand, the analysis conducted through the Scott-Knott algorithm results in a set of three homogeneous groups in which models are separated without ambiguity. This problem is even worse in the case of the ISBSG dataset for MMRE, where there are four homogeneous groups with many models common to two (OLS and EbA_Adj2) or three groups (RobMM, LTS, NN, EbA_Adj1, EbA, and Bag).

7 Threats to Validity

Regarding the experiments conducted in this study, we have to discuss some validity issues. A primary source of bias can stem from the selection of the datasets that are used.

In this regard, we state that these repositories contain data from a wide range of projects and, more importantly, they are in the public domain. Hence, all the experiments can be verified by a replicate study and compared with other findings from previous studies.

The second validity issue is related to the selection of the competitive models. Keeping in mind that there is a plethora of proposed SCE techniques, it is infeasible to include the whole range of possible candidates of methods. However, we chose to evaluate techniques which are representative of the three main SCE categories. Furthermore, our goal was to select techniques that are considered well-established since researchers utilize and study their performances in various experimental setups within the SCE domain. It is also important to note that the performance of all models, even the simplest ones, depends on choices of certain parameters. In the present study, we cannot claim that we compared the best possible model from each class. Instead, we chose representative models that are fitted reasonably well to our datasets.

Of course, we recognize that the study does not include techniques belonging to the general class of expert-judgment methods. Expert-based methods presuppose the participation of a group of capable and skillful practitioners in the development process in order to derive accurate estimations. In this paper, we decided to explore the predictive power of algorithmic-based models based on training datasets, but it is a challenge for future work to replicate the present study by taking into account expert judgment.

Regarding the bias of the validation procedure, we applied a k-fold cross validation instead of a hold-out method that splits the available data into two nonoverlapping parts, one for training and the other for testing. The basic advantage of the former method is that all observations are used for both training and validation, and each observation is used for validation exactly once. On the contrary, the holdout strategy does not use all the available data and the results are highly dependent on the choice of the training-test partition. Although the k-fold cross validation may also have certain limitations due to the splitting of training and test sets, the risk is minimized through the repeated drawn samples. More importantly, we also consider an experimental design and the corresponding analysis that takes into account the extra variability due to validation procedure, trying to smooth the consequences of the random sampling of data.

The ANOVA procedure which is the basis of our approach involves threats when basic assumptions are violated. These assumptions involve normally distributed response variables and homogeneity of variances. In our study, we used a transformation for deriving a normally distributed response variable, reducing the effect of skewness. However, the presence of heterogeneous variances and their effect on the algorithm is an open research question.

Finally, the lack of agreement concerning the measures of error in SCE experiments is also taken into account through the examination of different error functions. The results of the comparisons we made are not generalized, in the sense that for a specific dataset, a prediction candidate can present the best performance in terms of a certain aspect of accuracy, but at the same time, this model can also suffer from inaccuracy problems based on an alternative aspect of error.

8 Conclusions

In this paper, we deal with a critical research issue in software cost estimation concerning the simultaneous comparison of alternative prediction models, their ranking and clustering in groups of similar performance. We examined the predictive power of 11 models over six public domain datasets.

The whole procedure is settled on well-established statistical methodologies taking into consideration the multiple comparison problems. Keeping in mind the critical role of the adoption of reliable practices in the development process for both project managers and customers, we proposed a formal framework and structured guidelines in order to reinforce the knowledge acquisition and diminish the inherent uncertainty in SCE.

In this regard, we proposed certain directions concerning the utilization of alternative prediction models from different classes of estimation techniques, different datasets from public domain repositories, alternative error functions measuring different aspects of predictive performance, an experimental design in order to overcome the problem of splitting the datasets in the validation process, and more importantly, an efficient hypothesis testing procedure that signifies whether a set of prediction models gives statistically better results than another set of comparative models.

We have to emphasize that the experimentation section is used as a means for illustrating how the whole framework can be evaluated on the comparison setup, contributing to the systematic research of the performances of any kind of prediction techniques. Thus, it is not our purpose to determine the superiority of any prediction method and, even more, it is not wise to generalize the derived findings for the population of software projects.

On the other hand, the derived results of our experimentation either bring to the surface a few significant conclusions or confirm other essential results from past studies. Although the indicators from alternative models designate generally different predictive performances for the datasets, the proposed statistical hypothesis testing through the Scott-Knott test verifies that the predictive accuracy of a set of methods does not confirm a statistically significant difference among them. This suggests that the notion of the "best" estimation method may not have been so well defined thus far in the previous SCE research. Hence, there is a need to relocate the basis of the whole research. In other words, when a practitioner wishes to perform a comparative study, she or he ought to seek *a set of best estimation models* and not just a single one. Moreover, it seems that there is not a global solution since alternative methods can exhibit a few advantages in terms of certain aspects of prediction performance, but at the same time these candidates may suffer in terms of another aspect of accuracy. In order to overcome these inconsistencies, managers should utilize their experience and lead the whole process through the necessities that arise in each case. Therefore, it is important to emphasize here that the proposed method is an aid to the process of decision making by ranking and clustering the candidate models. However, it does not make decisions itself. The final decision is left to the expert and depends on several issues, even on personal criteria like experience, preference of statistical software, etc.

Another interesting finding concerns the utilization of complicated and more sophisticated models. It seems that very often a linear model is adequate enough to catch the trend between effort and other cost drivers of projects. Therefore, in certain cases it may be useless to strive to introduce new, highly complicated algorithms which, in practice, just cannot provide any further improvement. Finally, it is our strong belief that new estimation techniques should be tested and compared using appropriate statistical procedures.

Acknowledgments

The authors would like to thank the reviewers for their constructive comments which helped them to improve the paper.

References

[1] M. Jorgensen and M. Shepperd, "A Systematic Review of Software Development Cost Estimation Studies," *IEEE Trans. Software Eng.*, vol. 33, no. 1, pp. 33-53, Jan. 2007.

[2] M. Shepperd and G. Kadoda, "Comparing Software Prediction Techniques Using Simulation," *IEEE Trans. Software Eng.*, vol. 27, no. 11, pp. 1014-1022, Nov. 2001.

[3] B. Kitchenham, S. MacDonell, L. Pickard, and M. Shepperd, "What Accuracy Statistics Really Measure," *IEE Proc. Software Eng.*, vol. 148, pp. 81-85, June 2001.

[4] T. Foss, E. Stensrud, B. Kitchenham, and I. Myrtveit, "A Simulation Study of the Model Evaluation Criterion MMRE," *IEEE Trans. Software Eng.*, vol. 29, no. 11, pp. 985-995, Nov. 2003.

[5] N. Mittas and L. Angelis, "Comparing Cost Prediction Models by Resampling Techniques," *J. Systems and Software*, vol. 81, no. 5, pp. 616-632, May 2008.

[6] E. Stensrud and I. Myrtveit, "Human Performance Estimating with Analogy and Regression Models: An Empirical Validation," *Proc. IEEE Fifth Int'l Software Metrics Symp.*, pp. 205-213, Nov. 1998.

[7] B. Kitchenham and E. Mendes, "Why Comparative Effort Prediction Studies May Be Invalid," *Proc. ACM Fifth Int'l Conf. Predictor Models in Software Eng.*, pp. 1-5, May 2009.

[8] I. Myrtveit, E. Stensrud, and M. Shepperd, "Reliability and Validity in Comparative Studies of Software Prediction Models," *IEEE Trans. Software Eng.*, vol. 31, no. 5, pp. 380-391, May 2005.

[9] S. Lessmann, B. Baesens, C. Mues, and S. Pietsch, "Benchmarking Classification Models for Software Defect Prediction: A Proposed Framework and Novel Findings," *IEEE Trans. Software Eng.*, vol. 34, no. 4, pp. 485-496, July/Aug. 2008.

[10] J. Antony, *Design of Experiments for Engineers and Scientists.* Butterworth-Heinenmann, 2003.

[11] A. Scott and M. Knott, "A Cluster Analysis Method for Grouping Means in the Analysis of Variance," *Biometrics*, vol. 30, no. 3, pp. 507-512, Sept. 1974.

[12] J. Sayyad Shirabad and T. Menzies, "The PROMISE Repository of Software Engineering Databases," School of Information Technology and Eng., Univ. of Ottawa, http://promise.site.uottawa.ca/SERepository. 2005.

[13] ISBSG Data Set 10, http://www.isbsg.org. 2007.

[14] Y. Miyazaki, M. Terakado, K. Ozaki, and H. Nozaki, "Robust Regression for Developing Software Estimation Models," *J. Systems and Software*, vol. 27, pp. 3-16, 1994.

[15] T. Menzies, O. Jalali, J. Hihn, D. Baker, and K. Lum, "Stable Rankings for Different Effort Models," *Automated Software Eng.*, vol. 17, no. 4, pp. 409-437, Dec. 2010.

[16] G. Tsoumakas, L. Angelis, and I. Vlahavas, "Selective Fusion of Heterogeneous Classifiers," *Intelligent Data Analysis*, vol. 9, no. 6, pp. 511-525, Dec. 2005.

[17] J. Demšar, "Statistical Comparisons of Classifiers over Multiple Data Sets," *J. Machine Learning Research*, vol. 7, pp. 1-30, 2006.

[18] L. Briand, T. Langley, and I. Wieczorek, "A Replicated Assessment and Comparison of Common Software Cost Modeling Techniques," *Proc 22nd. IEEE Int'l Conf. Software Eng.*, pp. 377-386, 2000.

[19] B. Kitchenham, S. Pfleeger, B. McColl, and S. Eagan, "An Empirical Study of Maintenance and Development Accuracy," *J. Systems and Software*, vol. 64, no. 1, pp. 57-77, Oct. 2002.

[20] B. Kitchenham, E. Mendes, and H. Travassos, "Cross versus Within-Company Cost Estimation Studies: A Systematic Review," *IEEE Trans. Software Eng.*, vol. 33, no. 5, pp. 316-329, May 2007.

[21] R. Miller, *Simultaneous Statistical Inference*, second ed. McGraw-Hill, 1981.

[22] D. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, third ed. Chapman & Hall/CRC, 2004.

[23] Y. Hochberg and A. Tamhane, *Multiple Comparison Procedures.* Wiley & Sons, 1987.

[24] E. Da Silva, D. Ferreira, and E. Bearzoti, "Evaluation of Power and Type I Error Rates of Scott-Knott's Test by the Method of Monte Carlo," *Cincias Agrotécnicas*, vol. 23, pp. 687-696, 1999.

[25] L. Borges and D. Ferreira, "Power and Type I Errors Rate of Scott-Knott, Tukey and Newman-Keuls Tests under Normal and No-Normal Distributions of the Residues," *Revista de Matemática e Estatística*, vol. 21, no. 1, pp. 67-83, 2003.

[26] E. Jelihovschi and J. Faria, "ScottKnott: A Package for Performing the Scott-Knott Clustering Algorithm in R," *The R J.*, article in press.

[27] E. Ferreira, A. Dusi, J. Costa, G. Xavier, and N. Rumjanek, "Assessing Insecticide and Fungicide Effects on the Culturable Soil Bacterial Community by Analyses of Variance of Their DGGE Fingerprinting Data," *European J. Soil Biology*, vol. 45, nos. 5-6, pp. 466-472, Sept.-Dec. 2009.

[28] M. Ferreira, F. De Araujo, D. Costa, P. Rosa, H. Figueiredo, and L. Murgas, "Influence of Dietary Oil Sources on Muscle Composition and Plasma Lipoprotein Concentrations in Nile Tilapia, Oreochromis Niloticus," *J. World Aquaculture Soc.*, vol. 42, no. 1, pp. 24-33, Feb. 2011.

[29] D. Montgomery, *Design and Analysis of Experiments.* John Wiley & Sons, 1991.

[30] G. Blom, *Statistical Estimates and Transformed Beta.* Wiley, 1958.

[31] A. Abran and P. Robillard, "Function Points Analysis: An Empirical Study of Its Measurement Processes," *IEEE Trans. Software Eng.*, vol. 22, no. 12, pp. 895-910, Dec. 1996.

[32] A. Gray and S. MacDonell, "Software Metrics Data Analysis—Exploring the Relative Performance of Some Commonly Used Modeling Techniques," *Empirical Software Eng.*, vol. 4, no. 4, pp. 297-316, Dec. 1999.

[33] R. Jeffery, M. Ruhe, and I. Wieczorek, "Using Public Domain Metrics to Estimate Software Development Effort," *Proc. Seventh Int'l Software Metrics Symp.*, pp. 16-27, Apr. 2001.

[34] M. Shepperd and C. Schofield, "Estimating Software Project Effort Using Analogies," *IEEE Trans. Software Eng.*, vol. 23, no. 11, pp. 736-743, Nov. 1997.

[35] E. Mendes, I. Watson, C. Triggs, N. Mosley, and S. Counsell, "A Comparative Study of Cost Estimation Models for Web Hypermedia Applications," *Empirical Software Eng.*, vol. 8, no. 2, pp. 163-196, June 2003.

[36] Z. Li, G. Ruhe, A. Al-Emran, and M. Richter, "A Flexible Method for Software Effort Estimation by Analogy," *Empirical Software Eng.*, vol. 12, no. 1, pp. 65-106, Feb. 2007.

[37] N. Mittas, M. Athanasiades, and L. Angelis, "Improving Analogy-Based Software Cost Estimation by a Resampling Method," *Information and Software Technology,* vol. 50, no. 3, pp. 221-230, Feb. 2008.

[38] G. Finnie, G. Wittig, and J. Desharnais, "A Comparison of Software Effort Estimation Techniques: Using Function Points with Neural Networks, Case-Based Reasoning and Regression Models," *J. Systems and Software,* vol. 39, no. 3, pp. 281-289, Dec. 1997.

[39] P. Pendharkar, G. Subramanian, and J. Rodger, "A Probabilistic Model for Predicting Software Development Effort," *IEEE Trans. Software Eng.*, vol. 31, no. 7, pp. 615-624, July 2005.

[40] L. Briand, V. Basili, and W. Thomas, "A Pattern Recognition Approach for Software Engineering Data Analysis," *IEEE Trans. Software Eng.*, vol. 18, no. 11, pp. 931-942, Nov. 1992.

[41] L. Kaufman and P. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis.* Wiley, 1990.

[42] M. Shepperd, C. Schofield, and B. Kitchenham, "Effort Estimation Using Analogy," *Proc. 18th IEEE Int'l Conf. Software Eng,* pp. 170-178, 1996.

[43] J. Keung, B. Kitchenham, and R. Jeffery, "Analogy-X: Providing Statistical Inference to Analogy-Based Software Cost Estimation," *IEEE Trans. Software Eng.*, vol. 34, no. 4, pp. 471-484, July/Aug. 2008.

[44] P. Cortez, "Data Mining with Neural Networks and Support Vector Machines Using the R/rminer Tool," *Advances in Data Mining Application and Theoretical Aspects,* vol. 6171, pp. 572-583, 2010.

[45] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees,* Wadsworth Int'l Group, 1984.

[46] SPLUS 6 for Windows, *Guide to Statistics,* 2. Insightful Corp., 2001.

[47] G. Liebchen and M. Shepperd, "Data Sets and Data Quality in Software Engineering," *Proc. Fourth Int'l Workshop Predictor Models in Software Eng.*, pp. 39-44, May 2008.

[48] J. Keung, "Empirical Evaluation of Analogy-X for Software Cost Estimation," *Proc. Second ACM-IEEE Int'l Symp. Empirical Software Eng. and Measurement*, pp. 294-296, Oct. 2008.

[49] P. Ellis, *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results,* Cambridge Univ. Press, 2010.

[50] B. Kitchenham, "The Question of Scale Economies in Software—Why Cannot Researchers Agree?" *Information and Software Technology,* vol. 44, no. 1, pp. 13-24, Jan. 2002.

[51] B. Kitchenham, "A Procedure for Analyzing Unbalanced Data Sets," *IEEE Trans. Software Eng.*, vol. 24, no. 4, pp. 278-301, Apr. 1998.

[52] T. Menzies and M. Shepperd, "Special Issue on Repeatable Results in Software Engineering Prediction," *Empirical Software Eng.*, vol. 17, nos. 1/2, pp. 1-17, 2012.

[53] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," *Proc. 14th Int'l Joint Conf. Artificial Intelligence,* pp. 1137-1145, 1995.

[54] B. Efron, "Estimating the Error Rate of a Prediction Rule Improvement on Cross-Validation," *J. Am. Statistical Assoc.*, vol. 78, pp. 316-330, 1983.

Nikolaos Mittas received the BSc degree in mathematics from the University of Crete and the MSc and PhD degrees in informatics from the Aristotle University of Thessaloniki (A.U.Th). His research interests include application of statistics, especially computational statistics, to cost estimation of software projects, and generally to data from software projects.

Lefteris Angelis received the BSc degree in mathematics and the PhD degree in statistics from the Aristotle University of Thessaloniki (A.U.Th.). Currently, he is an associate professor in the Department of Informatics of A.U.Th. His research interests include statistical methods with applications to information systems and software engineering, computational methods in mathematics and statistics, planning of experiments, and simulation techniques.