Available online at www.sciencedirect.com
ScienceDirect
The Journal of Systems and Software 81 (2008) 616–632
The Journal of Systems and Software
www.elsevier.com/locate/jss

Comparing cost prediction models by resampling techniques

Nikolaos Mittas *, Lefteris Angelis

Department of Informatics, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece

Received 26 February 2007; received in revised form 28 May 2007; accepted 27 July 2007
Available online 5 October 2007

Abstract

The accurate software cost prediction is a research topic that has attracted much of the interest of the software engineering community during the latest decades. A large part of the research efforts involves the development of statistical models based on historical data. Since there are a lot of models that can be fitted to certain data, a crucial issue is the selection of the most efficient prediction model. Most often this selection is based on comparisons of various accuracy measures that are functions of the model's relative errors. However, the usual practice is to consider as the most accurate prediction model the one providing the best accuracy measure without testing if this superiority is in fact statistically significant. This policy can lead to unstable and erroneous conclusions since a small change in the data is able to turn over the best model selection. On the other hand, the accuracy measures used in practice are statistics with unknown probability distributions, making the testing of any hypothesis, by the traditional parametric methods, problematic. In this paper, the use of statistical simulation tools is proposed in order to test the significance of the difference between the accuracy of two prediction methods: regression and estimation by analogy. The statistical simulation procedures involve permutation tests and bootstrap techniques for the construction of confidence intervals for the difference of measures. Four known datasets are used for experimentation in order to validate the results and make comparisons between the simulation methods and the traditional parametric and non-parametric procedures.

Keywords: Software cost estimation; Accuracy measure; Confidence interval; Bootstrap; Permutation test

1. Introduction

One of the most important phases in planning, scheduling and risk management of projects that has attracted the interest of many researchers during the recent decades, is software cost estimation. A lot of methods have been proposed in the literature for accurate predictions of the cost. A large part of the ongoing research concerns mathematical models that are developed based on historical data of complete projects.

Although there is a need to know which of the models is the best there seems to be no global answer for all kinds of data. Indeed, the model fitting and predictive accuracy are known to depend on the kind of data we use (types and number of project attributes, sample size, measurement accuracy, etc.). This fact can be concluded by the large number of comparative studies with controversial results appeared in the literature so far. The situation becomes much more complicated when we consider that each generic method can be applied under a large number of alternative variations. As an example we can consider regression analysis where a simple transformation of a variable can result to a significantly better model.

Having defined the dataset on which models will be based, it is crucial to decide what criterion will be used for the choice of the best model. A lot of accuracy measures have been proposed in the literature and used in practice so far and all of them are functions of the predictive error, measured by appropriate methods on the projects of the available dataset. However, a single measure is just a statistic, i.e. a value computed from a sample (usually mean, median or percentage of errors or relative errors) and as such contains significant

* Corresponding author.
E-mail addresses: nmittas@csd.auth.gr (N. Mittas), lef@csd.auth.gr (L. Angelis).

0164-1212/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jss.2007.07.039

variability. Thus, when we compare models based solely on a single value we take the risk to consider as a significant difference which in fact may be not so significant. A possible implication of such a choice is that when a small amount of data is changed, the ''best'' model may become no longer ''best''. Therefore, this policy of determining the ''most accurate'' model based on a single indicator can lead to unstable and erroneous decision-making.

From what we have already mentioned, we can see that there is a need for a formal comparison of models on the basis of an accuracy measure and this comparison should involve more inferential statistical techniques, such as hypothesis testing and confidence intervals. This formal testing is often overlooked in favor of simplicity, but its benefits are very important since a model that is declared the best through such a testing can gain the confidence for further use to other data.

In some of the most recent studies *parametric* and *non-parametric* procedures are carried out to test the validity of the most accurate prediction model. A parametric test is a procedure that requires an assumption regarding the underlying theoretical distribution of the sample and can be used in the case of the difference of means, whereas a non-parametric test does not depend on any distribution assumption. A well-known parametric test suitable for the comparisons we are interested in is the *paired sample t-test*, whereas an alternative non-parametric test is the *Wilcoxon signed rank test*.

Most of the research works in the literature of software cost prediction models use validation measures based on the *magnitude of relative error* (MRE). In these studies, the results are quite contradictory and of course there is not a global conviction about the most accurate prediction method. Below we give a brief review of the related literature on comparison studies.

In Shepperd and Schofield (1997), nine datasets were used to compare estimation by analogy (EbA) with regression. The comparison was based on the mean of the MRE, called MMRE and the PREDmre (a measure counting the percentage of MREs in a predefined interval) accuracy measures with no statistical test and the authors concluded that EbA outperformed regression.

In Kitchenham (1998), a procedure called forward pass residual analysis for analyzing unbalanced datasets was presented and the results were compared with classification and regression trees (CART) in terms of MMRE and PREDmre without carrying out a significance test. The results showed the superiority of forward pass residual analysis.

In Myrtveit and Stensrud (1999), EbA was compared with regression models, whereas the procedure was based on four accuracy measures computed from statistics of MRE: mean (MMRE), median (MdMRE), standard deviation (SDmre) and maximum value (MAXmre). The statistical significance of the results was tested using parametric paired *t*-test for the case of MMRE and non-parametric Wilcoxon signed rank test for the case of MdMRE.

In Briand et al. (2000), several prediction methods were compared, such as ordinary least-squares (OLS) regression, stepwise analysis of variance (ANOVA) for unbalanced datasets, CART, EbA and also combinations of CART with OLS and EbA. In that study, the MdMRE accuracy measure was used and the comparisons were made by the Wilcoxon signed rank test. The results evidenced that OLS and ANOVA regression models outperformed CART, EbA and the combinations of CART with OLS and EbA.

In Mair et al. (2000), the authors presented a comparative study regarding the use of three machine learning methods: artificial neural networks (ANNs), case-based reasoning (CBR), rule induction (RI) and OLS. The comparison was based on the MMRE, MdMRE, MinMRE and MaxMRE, whereas no statistical test was carried out. The conclusion was that ANNs seem to be the most accurate technique, whereas RI methods were the least accurate.

In Shukla (2000), simulation experiments were conducted for comparative performance evaluation of a top–down strategy for the construction of a decision tree (CARTX) and neural network (NN) predictors, trained with back-propagation, quick-propagation and genetic algorithm. Student's *t*-test for the average prediction error was used to establish that NN predictors trained by a genetic algorithm were a significant improvement over both CARTX and quick-propagation NN.

In Jeffery et al. (2001), OLS regression, stepwise ANOVA, CART, EbA and robust regression were used to compare company specific models with models based on multi-company data. The comparison was carried out for MMRE, MdMRE and PREDmre validation measures by the Wilcoxon signed rank test. The conclusion was that robust regression and OLS performed most accurately in the case of multi-company data, whereas OLS, CART and EbA performed best when using company's own data.

In Myrtveit et al. (2005), a simulation study was presented in which regression models were compared with EbA models and several accuracy measures (MAR, MMRE, MMER, MBRE, MIBRE, RSD, LSD) were evaluated. No significance test was used for pairwise comparisons of the accuracy measures. The authors noted a lack of convergence about the most accurate prediction method and underlined three important factors about this.

Summarizing, most of the aforementioned studies use the MRE, whereas the tests for assessing statistically significant difference between the models' accuracy are the paired *t*-test and the Wilcoxon signed rank test for the MMRE and MdMRE measures, respectively.

The problem with the software projects cost data is that the samples are quite small and skewed, so it is not easy to make assumptions regarding the distribution of the prediction errors resulting from a certain process or a model and from which the accuracy measures are calculated. For such types of data it is known from the statistical literature that a certain class of simulation methods may be proved quite

beneficial. These methods are based on *resampling*, i.e. on drawing a large number of samples from the original sample in order to "reconstruct" the underlying theoretical distribution. It is obvious that these methods are computational techniques requiring large number of iterations in a computer. However, the rapid evolution of computer power has contributed in the wide spread of these methods which are nowadays used in all research areas tending to supersede the traditional statistical procedures. They are also implemented in all well-known statistical packages.

The goal of this paper is to further extend the research regarding the comparison of models proposing alternative tests for a wide range of accuracy measures. Two alternative methods for statistical inference, namely the bootstrap confidence intervals and the permutation tests are applied to test the difference between the most commonly used accuracy measures based on errors obtained from two prediction methods, the estimation by analogy and regression analysis. We have to emphasize that it is not our purpose to compare specific models and to find the best model, but rather to contribute in the systematic comparison of models. So, we can summarize the contribution of the paper in the following points:

- We use the traditional parametric and non-parametric procedures in order to evaluate the predictive performance of the two comparative models based on the most known accuracy measure, the magnitude of the relative error (MRE).
- We consider formal comparisons of alternative measures based on the magnitude of relative error to the estimate (MER), the squared error (SQE) and the *z*-ratio that have not been studied yet under statistical tests. The outcomes are interesting as they show that the conclusions of comparisons based on different criteria can be quite different.
- We perform statistical tests for the Pred measures that are essentially percentages and have not been considered yet in formal comparisons.
- We present two statistical simulation techniques, the bootstrap confidence intervals and the permutation test for comparisons of the accuracy measures. More precisely, we use three types of bootstrap confidence intervals; the *t confidence interval using bootstrap standard errors (t-CIbse)*, the *bootstrap percentile confidence interval (CIbp)*, and the *bias-corrected and accelerated confidence interval (CIbca)*. Furthermore, the permutation test is utilized for the comparison of accuracy measures that are based on the means, medians and percentages.

In order to present a comprehensive study of all the accuracy methods and the statistical tests, extensive experimentation is made with four known datasets from the literature.

In Section 2, we give the definition of all the accuracy measures that we use in our comparisons. In Section 3, we give an account of the traditional statistical tests that are suitable for the accuracy measures we consider in our study. In Section 4, we extensively present the resampling techniques (bootstrap confidence intervals and permutation tests) that are utilized in the formal comparisons of the two models. In Section 5, we give the results obtained from the application to real data. Finally, in Section 6, we present the conclusions and comments on future work.

2. Accuracy measures

The accuracy measures that are most frequently used for validating cost models are based on the y_A (actual) and the y_E (estimated from a model) cost values. Specifically, two measures of *local* absolute relative error have been extensively used so far: the *magnitude of relative error* (MRE) (Conte et al., 1986) and the *magnitude of relative error to the estimate* (MER) (Kitchenham et al., 2001) defined as

$$\mathrm{MRE} = \frac{|y_A - y_E|}{y_A} \quad \text{and} \quad \mathrm{MER} = \frac{|y_A - y_E|}{y_E}$$

Except from these, some other alternative measures have been proposed. The *squared error* (SQE) (Shan et al., 2002) and the *z-ratio* (Kitchenham et al., 2001) defined as

$$\mathrm{SQE} = (y_A - y_E)^2 \quad \text{and} \quad z = \frac{y_E}{y_A}$$

These local measures can yield a *global* predictive accuracy measure for a model by computing a statistic from them. The most commonly used measures are given in Table 1.

As we can see, the derivation of these accuracy measures is simple, either by computing the mean, the median or the

Table 1
Global accuracy measures

$\mathrm{MMRE} = \frac{1}{n}\sum_{i=1}^{n}\frac{\lvert y_{A_i} - y_{E_i}\rvert}{y_{A_i}}$	$\mathrm{MdMRE} = \mathrm{median}\left\{\frac{\lvert y_{A_i} - y_{E_i}\rvert}{y_{A_i}}\right\}$
$\mathrm{MMER} = \frac{1}{n}\sum_{i=1}^{n}\frac{\lvert y_{A_i} - y_{E_i}\rvert}{y_{E_i}}$	$\mathrm{MdMER} = \mathrm{median}\left\{\frac{\lvert y_{A_i} - y_{E_i}\rvert}{y_{E_i}}\right\}$
$\mathrm{PREDmre}(100p) = \frac{\#(\text{projects with MRE} \leqslant p)}{\#(\text{projects})}$	
$\mathrm{PREDmer}(100p) = \frac{\#(\text{projects with MER} \leqslant p)}{\#(\text{projects})}$	
$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_{A_i} - y_{E_i})^2$	
	$\mathrm{Median}(z) = \mathrm{median}\left\{\frac{y_{E_i}}{y_{A_i}}\right\}$
$\mathrm{Mean}(z) = \frac{1}{n}\sum_{i=1}^{n}\frac{y_{E_i}}{y_{A_i}}$	

proportion of relative errors that are less than p, where usually $p = 0.20$ or $p = 0.25$.

3. Traditional statistical tests for the accuracy measures

Suppose that we wish to compare two cost prediction models (ModelA and ModelB) on the same dataset. Suppose also that we obtain predictions using the well-known method of *jackknife* (or *hold-one-out*), i.e. we estimate the cost of each one of the projects in the dataset using a model constructed by all the other projects. Since the cost of each project is predicted by two alternative models, it is reasonable to use statistical tests for two related or *paired* samples.

After applying two models on the same dataset, we obtain by the jackknife method two paired samples of errors (MRE_{ModelA} and MRE_{ModelB}, MER_{ModelA} and MER_{ModelB}, SQE_{ModelA} and SQE_{ModelB}, z_{ModelA} and z_{ModelB}), which are values of continuous variables. Based on these samples we have to draw conclusions concerning their differences. For example in the case of MMRE, MMER, MSE and Mean(z) we essentially have to perform a test for comparing two means. The statistical significance of the difference can be tested through parametric and non-parametric tests.

The *paired samples t-test* is a parametric procedure that compares the means of two samples. The test is performed by computing first a new sample of size n from the differences of all pairs of the original samples and then by calculating a *t statistic* which is used to test whether the mean difference is significantly different from zero. Equivalently, the same test can be performed by computing a *paired difference confidence interval (CI)*. For the difference of means $\mu_D = \mu_1 - \mu_2$ the CI is the following: From the new sample of differences, estimates of the mean $\bar{x}_D$, and the standard deviation s_D are calculated. Then, a $100(1-a)\%$ CI is obtained by the formula (Sheskin, 2004; Efron and Tibshirani, 1993):

$$\left[\bar{x}_D - t_{n-1;1-a/2}\frac{s_D}{\sqrt{n}},\ \bar{x}_D + t_{n-1;1-a/2}\frac{s_D}{\sqrt{n}}\right]$$

where $t_{n-1;1-a/2}$ represents the $1-a/2$ quantile of Student's t distribution with $n-1$ degrees of freedom. Note that the quantity $s_D/\sqrt{n}$ is the standard error of the mean, i.e. the standard deviation of $\bar{x}_D$. In general, the standard error of any statistic is very important for the computation of its CIs. The paired t-test and the corresponding CI can be used to compare MMRE, MMER, MSE and Mean(z) of two models.

A strong assumption of the t-test is that the difference of the samples (difference of MREs, MERs, SQEs and z-values) follows a normal or nearly normal distribution. However, there is no theoretical evidence that samples of errors obtained by any model are normally distributed. Furthermore, practice showed that very often in real data the distribution of errors is highly skewed, very often with outliers.

A non-parametric analogue of the paired t-test is the *Wilcoxon signed rank test* which tests whether there is a significant difference between the medians of two paired samples. The test is based on ranks of the sample values, so there is no need for assumptions regarding their distributions. The test is also robust in the sense that it is not affected by the presence of outliers. So, by definition the Wilcoxon test can be used to compare MdMRE, MdMER and Median(z) of two models.

Regarding the PREDmre($100p$) and PREDmer($100p$) accuracy measures, we can use the non-parametric McNemar procedure that tests for changes in responses using the chi-square distribution and can be applied for comparing two paired dichotomous samples. In our case, a dichotomous variable can be constructed for the MREs (or similarly for MERs) of a model, by the following rule:

$$\mathrm{dmre}_{\mathrm{Model}} = \begin{cases} 1, & \text{if } \mathrm{MRE}_{\mathrm{Model}} \leqslant p \\ 0, & \text{if } \mathrm{MRE}_{\mathrm{Model}} > p \end{cases}$$

Therefore, the test is suitable for Pred measures which are simply the proportion of units in each sample. The McNemar test is based on the construction of a 2×2 contingency table (Table 2) where the values of the two dichotomous variables (one for each model) are cross-tabulated.

An interesting characteristic of this table is that there are two concordant cells in which the paired results are the same (both negative or both positive) and two discordant cells in which the paired results are different for the same project's MRE (positive–negative or negative–positive). The McNemar test utilizes only the information in the discordant cells in order to analyze whether the two models show equivalent results. The null hypothesis is that the proportions of positive results are the same for both prediction models and is retained if the discordant pairs are distributed uniformly in the two discordant cells.

Detailed description of all these tests can be found in any statistical textbook, see for example (Sheskin, 2004).

An important remark that we have to take into account for software cost datasets is that most often the samples are quite small. For small datasets from unknown underlying distributions, resampling techniques have been developed and these will be described next.

4. Resampling techniques

There is a large number of resampling techniques developed for quite different purposes. However, in our study, in order to compare the errors of two models we use two of the most popular resampling techniques: bootstrap

Table 2
McNemar's 2×2 table results

		$\mathrm{dmre}_{\mathrm{ModelB}}$	
		0	1
$\mathrm{dmre}_{\mathrm{ModelA}}$	0	#(negative–negative)	#(negative–positive)
	1	#(positive–negative)	#(positive–positive)

Table 3
Descriptive statistics for local measures of error

	MRE_{EbA} (%)	MRE_{OLS} (%)	MER_{EbA} (%)	MER_{OLS} (%)	SQE_{EbA}	SQE_{OLS}	z_{EbA}	z_{OLS}
Mean	39.86	24.68	32.95	27.09	9772.64	10958.16	1.077	1.050
Median	22.44	21.23	26.10	21.67	–	–	0.841	1.028
Hit rate ($100p \leqslant 25\%$)	57.14	57.14	47.62	61.90	–	–	–	–

Table 4
Significance of all paired samples tests for the global accuracy measures

	Paired *t*-test	Wilcoxon signed rank test	McNemar test
MMRE	0.135	–	–
MdMRE	–	0.203	–
PREDmre(25)	–	–	1
MMER	0.493	–	–
MdMER	–	0.191	–
PREDmer(25)	–	–	0.508
MSE	0.808	–	–
Mean(z)	0.817	–	–
Median(z)	–	0.759	–

confidence intervals and permutation tests. These will be used for testing the difference of means, medians and percentages of the errors between two models.

The main idea behind testing the significance of a difference is to test essentially whether the observed difference could reasonably occur "just by chance" due to the random sample used for developing the model. If this is not the case, we can infer that we have evidence for a significant difference (Sheskin, 2004).

Traditional methods use some statistics computed from the sample which is assumed to follow a theoretical distribution. In case the difference in the sample falls outside a range of critical values, the difference is considered significant. On the other hand, in the resampling procedures these critical values are computed by drawing repetitive samples from the original sample.

4.1. Bootstrap confidence intervals

Bootstrap is a computer-based simulation technique that can be used in order to extract and explore the sample distribution of a statistic (Efron and Tibshirani, 1993). We will only describe the non-parametric bootstrap method which makes no assumptions of theoretical distributions. The rationale behind non-parametric bootstrap procedure is the generation of a large number of independent samples drawn with replacement from the original sample.

In general, the problem is to use a random sample $\mathbf{x} = (x_1, \ldots, x_n)$ for statistical inference about an unknown population parameter θ (such as mean, median, percentage, variance, etc.). The sample statistic $\hat{\theta}$ is a point estimator of the parameter θ. Three basic steps are followed:

1. Obtain a large number of samples B, each one by the following procedure: from the set $\{1, 2, \ldots, n\}$ draw randomly with replacement a set of indices $\{j_1, \ldots, j_n\}$ and form the ith sample $\mathbf{x}^{*i} = (x_{j_1}, \ldots, x_{j_n})$ of size n, where $i = 1, 2, \ldots, B$.
2. From each bootstrap sample $\mathbf{x}^{*i}$, compute θ^{*i} ($i = 1, 2, \ldots, B$), the value of the statistic under consideration.
3. The B values of the bootstrap statistics θ^{*i} form an approximation of the sampling distribution of $\hat{\theta}$.

The approximate distribution obtained by the bootstrap method can be used for computing CIs for the population parameter θ. In practice, the bootstrap CIs are particularly useful when the data are skewed and the samples are quite small, as the software cost data usually are.

In our case the unknown population parameter is considered to be the difference of means, medians or percentages between two paired distributions of errors obtained by two different cost models. The point therefore is to test whether a CI for the difference of means, medians or percentages contains the zero value. In such a case we cannot support the existence of significant difference. Next, we

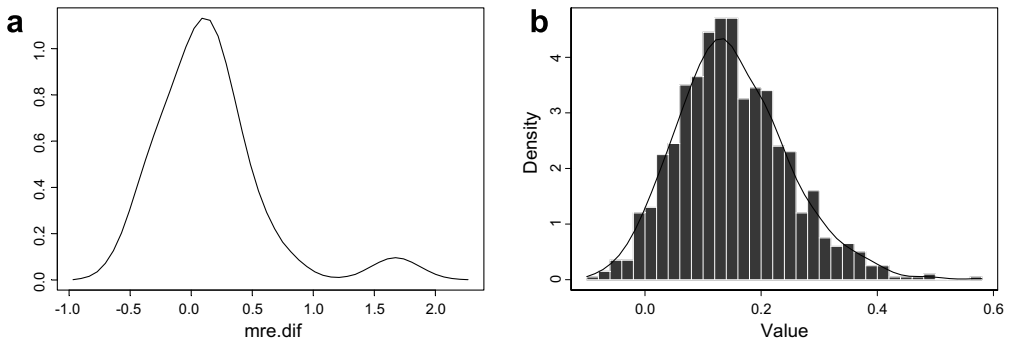

Fig. 1. (a) Density and (b) bootstrap estimation for the difference of the MRE_{dif}.

describe some methods for computing bootstrap confidence intervals (Venables and Ripley, 2002).

4.1.1. t CI using bootstrap standard errors (t-CIbse)

This approach can be applied when the bootstrap distribution of the $\hat{\theta}$ statistic shows a normal shape. First, the standard error of the statistic is estimated as the standard deviation of all the θ^{*i} values ($i = 1, 2, \ldots, B$) obtained from the bootstrap samples. This is denoted by $\hat{s}e_{\text{boot}}$ and measures how much the statistic varies under random resampling. Then, the bootstrap t-CI is obtained by the following expression:

$$\left[\hat{\theta} - t_{n-1;1-a/2} \times \hat{s}e_{\text{boot}},\ \hat{\theta} + t_{n-1;1-a/2} \times \hat{s}e_{\text{boot}}\right]$$

There is a limitation in the usage of this type of CI: these intervals are accurate only when the bootstrap distribution is approximately normal and has small bias.

4.1.2. Bootstrap percentile CI (CIbp)

This method is applied when the bootstrap distribution of the $\hat{\theta}$ statistic is non-normal, highly skewed and there are outliers. First, from the empirical distribution containing all the θ^{*i} values ($i = 1, 2, \ldots, B$) obtained from the bootstrap samples, we compute the values $\theta^*_{a/2}$ and $\theta^*_{1-a/2}$ corresponding to the $100(a/2)$th and the $100(1-a)$th percentiles. Then, the CIbp is simply given by

$$\left[\theta^*_{a/2},\ \theta^*_{1-a/2}\right]$$

A reasonable question arising from the two types of bootstrap CI is under which conditions it is safe to use the aforementioned methods. The answer is not so straightforward and the recommendation of the statisticians is to inspect whether these intervals are reasonably close by comparing them with each other (Moore et al., 2003). If the bootstrap distribution is close to normal, the t-CIbse and the CIbp bounds will be quite close. On the other hand, if there is a large divergence, neither type of the above intervals should be used.

4.1.3. Bias-corrected and accelerated confidence interval (CIbca)

This is a method attempting to shift and scale the percentile intervals for cases where there is a large divergence between t-CIbse and CIbp. This happens in cases of skewness (Moore et al., 2003). The method is also based on the percentiles of the bootstrap distribution but their computation depends on two numbers $\hat{a}$ and $\hat{z}_0$, which make the appropriate adjustments in order to correct bias and skewness. The value $\hat{a}$ is called *acceleration* because it refers to the rate of change of the standard error of $\hat{\theta}$ with respect to the true parameter value θ and the value $\hat{z}_0$ is the correction of bias. These values are computed from the original sample and the bootstrap samples by the following expressions:

Table 5
Confidence intervals for the difference of means

CI	MMRE		MMER		MSE		Mean(z)	
	90%	95%	90%	95%	90%	95%	90%	95%
Student's t-CI	(−0.016, 0.320)	(−0.051, 0.355)	(−0.086, 0.204)	(−0.117, 0.234)	(−9485.5, 7114.5)	(−11224.0, 8852.9)	(−0.120, 0.252)	(−0.144, 0.302)
t-CIbse	(−0.003, 0.307)	(−0.036, 0.339)	(−0.087, 0.204)	(−0.119, 0.235)	(−9155.7, 6784.6)	(−10828.1, 8457.0)	(−0.158, 0.211)	(−0.197, 0.250)
CIbp	(0.015, 0.309)	(−0.016, 0.346)	(−0.080, 0.199)	(−0.113, 0.220)	(−9616.7, 5752.1)	(−10634.0, 6853.5)	(−0.139, 0.210)	(−0.165, 0.256)
CIbca	(0.033, 0.345)	(0.013, 0.402)	(−0.119, 0.176)	(−0.151, 0.195)	(−9992.9, 5409.8)	(−10999.0, 6392.7)	(−0.121, 0.252)	(−0.141, 0.303)

Table 6
Confidence intervals for the difference of medians

CI	MdMRE		MdMER		Median(z)	
	90%	95%	90%	95%	90%	95%
CIbp	(−0.113, 0.174)	(−0.131, 0.175)	(−0.046, 0.207)	(−0.066, 0.229)	(−0.282, −0.065)	(−0.300, −0.013)

Table 7
Confidence intervals for the difference of percentages

CI	PREDmre(25)		PREDmer(25)	
	90%	95%	90%	95%
Student's t-CI	(−0.244, 0.244)	(−0.289, 0.289)	(−0.365, 0.088)	(−0.406, 0.132)
t-CIbse	(−0.258, 0.259)	(−0.313, 0.313)	(−0.375, 0.089)	(−0.424, 0.138)
CIbp	(−0.238, 0.238)	(−0.286, 0.286)	(−0.381, 0.048)	(−0.429, 0.095)
CIbca	(−0.238, 0.238)	(−0.286, 0.286)	(−0.381, 0.062)	(−0.381, 0.143)

$$\hat{z}_0 = \Phi^{-1}\left(\frac{\#(\theta^{*i} < \hat{\theta})}{B}\right) \quad \text{and}$$

$$\hat{a} = \frac{\sum_{i=1}^{n}\left(\overline{\hat{\theta}^{(J)}} - \hat{\theta}^{(-i)}\right)^3}{6\left[\sum_{i=1}^{n}\left(\overline{\hat{\theta}^{(J)}} - \hat{\theta}^{(-i)}\right)^2\right]^{3/2}}$$

where # means "number of", $\Phi^{-1}(\cdot)$ denotes the inverse of the standard normal cumulative distribution function, $\hat{\theta}^{(-i)}$ is the value of the statistic using the sample with the ith data point removed (the ith jackknife sample) and

$$\overline{\hat{\theta}^{(J)}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}^{(-i)}$$

The $100(1-a)\%$ CIbca confidence interval is then given by

$$\left[\hat{\theta}^{*}_{a_1},\ \hat{\theta}^{*}_{a_2}\right]$$

where

$$\alpha_1 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{a/2}}{1 - \hat{a}(\hat{z}_0 + z_{a/2})}\right) \quad \text{and}$$

$$\alpha_2 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{1-a/2}}{1 - \hat{a}(\hat{z}_0 + z_{1-a/2})}\right)$$

In the above expressions, $\Phi(\cdot)$ denotes the standard normal cumulative distribution function and $z_{a/2}$ is the $100a/2$th percentile of a standard normal distribution. Setting z_0 and a equal to zero, the CIbca is the same as the percentile interval. For more details on the CIbca see (Efron and Tibshirani, 1993). The CIbca method is recommended for general use, especially for non-parametric problems or when high accuracy is required.

Table 8
Significance of all paired samples permutation tests for the global accuracy measures

	Permutation test
MMRE	0.098
MdMRE	0.856
PREDmre(25)	1
MMER	0.534
MdMER	0.364
PREDmer(25)	0.504
MSE	0.842
Mean(z)	0.824
Median(z)	0.066

Table 9
General results for the Abran and Robillard dataset

Statistical test	Rejection of null hypothesis			
	MRE	MER	SQE	z
Paired t-test for the difference of means	No	No	No	No
90% t-CIbse for the difference of means	No	No	No	No
90% CIbp for the difference of means	Yes	No	No	No
90% CIbca for the difference of means	Yes	No	No	No
Permutation test for the difference of means	Yes	No	No	No
Wilcoxon test for the difference of medians	No	No	–	No
90% CIbp for the difference of medians	No	No	–	Yes
Permutation test for the difference of medians	No	No	–	Yes
McNemar test for the difference of percentages	No	No	–	–
90% t-CIbse for the difference of percentages	No	No	–	–
90% CIbp for the difference of percentages	No	No	–	–
90% CIbca for the difference of percentages	No	No	–	–
Permutation test for the difference of percentages	No	No	–	–

4.2. Permutation tests

Permutation tests (Efron and Tibshirani, 1993), are resampling techniques based on rearrangements of the data. For the case of paired-samples comparisons, the samples are in the form $(x_1, y_1), \ldots, (x_n, y_n)$. The method generates a large number of paired samples where each pair (x_i, y_i) is permuted randomly. The statistic under consideration is computed for each generated paired sample and its sampling distribution is used for testing any hypothesis. The main difference between permutation tests and bootstrap is the way the resampling procedure is performed,

Table 10
Descriptive statistics for local measures of error

	MRE_{EbA} (%)	MRE_{OLS} (%)	MER_{EbA} (%)	MER_{OLS} (%)	SQE_{EbA}	SQE_{OLS}	z_{EbA}	z_{OLS}
Mean	82.07	27.61	49.06	31.86	75.693	195.326	1.503	1.043
Median	29.49	25.03	31.99	20.19	–	–	0.976	1.028
Hit rate ($100p \leqslant 25\%$)	41.67	50.00	45.83	54.17	–	–	–	–

Table 11
Significance of all paired samples tests for the global accuracy measures

	Paired *t*-test	Wilcoxon signed rank test	McNemar test
MMRE	0.196	–	–
MdMRE	–	0.136	–
PREDmre(25)	–	–	0.754
MMER	0.234	–	–
MdMER	–	0.128	–
PREDmer(25)	–	–	0.754
MSE	0.475	–	–
Mean(z)	0.284	–	–
Median(z)	–	0.684	–

i.e. permutation tests draw samples without replacement in contrast to what the bootstrap does.

In our comparisons the paired samples consist of the local measures of error (MRE, MER, SQE or z) obtained by two cost models. Permutation tests are applied to test the significance of difference between means, medians or percentages. A typical permutation test involves the following steps:

1. The paired data are randomly permuted as we already described.
2. The difference of the statistic under consideration (for example the difference of means or medians between MRE_{ModelA} and MRE_{ModelB}) is computed.
3. Steps 1 and 2 are repeated a large number of times (say B).
4. The statistic from the original sample (for example $MMRE_{ModelA} - MMRE_{ModelB}$) is computed and is located in the sampling distribution of all values obtained from Steps 1–3 in order to estimate the significance of the hypothesis (p-value). Note that the null hypothesis in our case is that any difference is equal to zero.

Permutation tests are used when the t-test fails to give accurate results in situations where the normality assumption is not valid.

4.3. The comparative prediction models

Two cost prediction methods were compared: EbA and OLS. Each model's prediction accuracy was evaluated through the jackknife procedure which estimates the cost of each project from all the others.

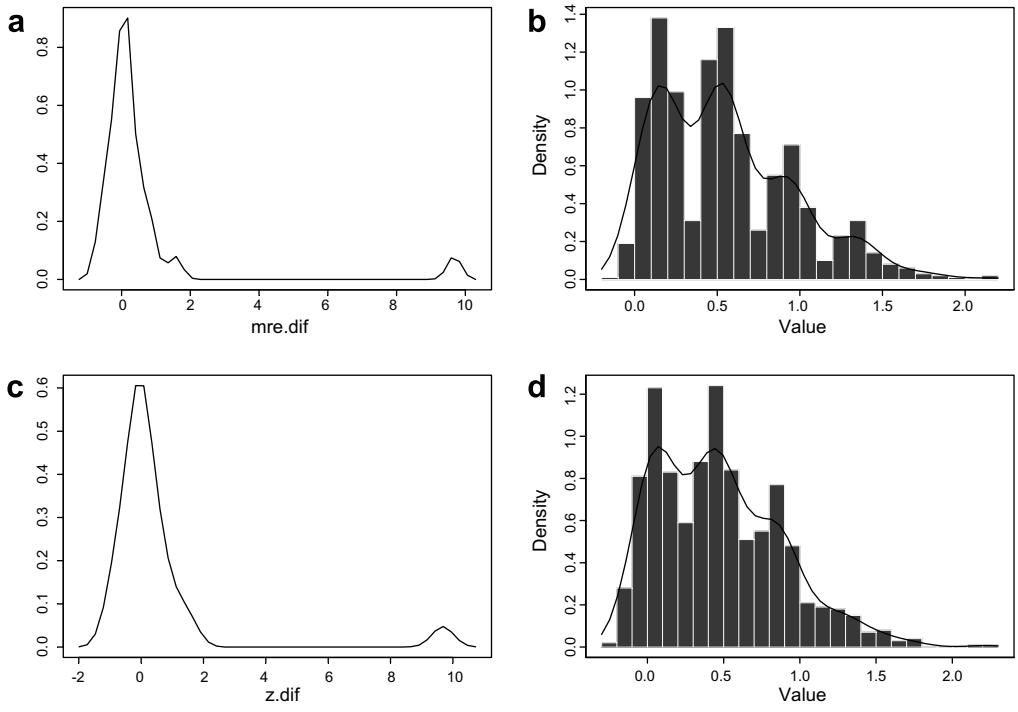

Fig. 2. (a–c) Density and (b–d) bootstrap estimation for the difference of the MRE_{dif} and z_{dif}.

Table 12
Confidence intervals for the difference of means

CI	MMRE		MMER		MSE		Mean(z)	
	90%	95%	90%	95%	90%	95%	90%	95%
Student's t-CI	(−0.156, 1.245)	(−0.301, 1.390)	(−0.069, 0.413)	(−0.119, 0.463)	(−401.6, 162.4)	(−460.0, 220.8)	(−0.258, 1.177)	(−0.407, 1.325)
t-CIbse	(−0.140, 1.229)	(−0.282, 1.371)	(−0.069, 0.412)	(−0.119, 0.463)	(−398.6, 159.3)	(−456.4, 217.1)	(−0.238, 1.156)	(−0.382, 1.301)
CIbp	(0.035, 1.335)	(0.007, 1.434)	(−0.058, 0.416)	(−0.103, 0.450)	(−439.8, 72.3)	(−486.7, 87.9)	(−0.081, 1.285)	(−0.123, 1.412)
CIbca	(0.103, 1.782)	(0.074, 2.206)	(−0.038, 0.430)	(−0.081, 0.468)	(−620.0, 52.6)	(−962.5, 61.6)	(0.016, 0.672)	(−0.030, 2.002)

EbA was used to predict the cost of each project in a dataset by finding the closest projects after some calibration regarding the distance metric, number of closest neighbours, standardization, and statistic for estimation (Angelis and Stamelos, 2000) in order to optimize as much as possible its predictive power.

OLS regression explains the relation between several independent variables (cost factors) and a dependent variable (effort or productivity) in the form of a linear relationship. Since the variables are usually non-normally distributed, they need some transformation in order to obtain a valid linear model. The most usual transformations are the logarithmic and the square root (in cases of zero values). In order to handle mixed data with categorical and continuous variables, we used the *general linear model* method that combines regression analysis and analysis of variance.

5. Application to real data

In this section we give the results of the comparisons described earlier as applied to four datasets from the literature. The results from the resampling methods are presented in tables along with results from the traditional tests.

5.1. The Abran–Robillard dataset

The first dataset contains 21 projects (Abran and Robillard, 1996) from a major Canadian financial organization with 10 independent continuous attributes and a dependent continuous variable (actual effort days). After applying the two prediction methods (OLS and EbA), we computed by jackknife the local measures of error for which we obtained the basic statistics. These are given in Table 3.

From the statistics of Table 3, we can see that the measures MMRE, MMER, MdMRE and MdMER are lower for the OLS model. The Mean(z) is closer to one for the OLS, whereas the opposite is true for the case of the Median(z). The PREDmre(25) is the same for the two models, whereas the PREDmer(25) is higher for the OLS model. On the other hand, the MSE is lower for the EbA.

5.1.1. Parametric and non-parametric paired samples tests

The obvious question arising from the previous global measures is whether we can distinguish the ''best'' model. The accuracy measures give contradictory results. The OLS outperforms the EbA for the six out of nine measures, whereas the EbA outperforms OLS for two accuracy measures. None of the models gives better results for the PREDmre(25). However, these statistics cannot be used on their own, in order to select the best model.

The paired t-test (Table 4) focuses on the means of the paired local errors of the two models and reports the significance of their difference. This procedure is utilized for the comparison of MMRE, MMER, MSE and Mean(z) accuracy measures. Since the significance value (or p-value) is

Table 13
Confidence intervals for the difference of medians

CI	MdMRE		MdMER		Median(z)	
	90%	95%	90%	95%	90%	95%
CIbp	(−0.098, 0.215)	(−0.121, 0.275)	(−0.038, 0.249)	(−0.059, 0.283)	(−0.230, 0.157)	(−0.266, 0.208)

Table 14
Confidence intervals for the difference of percentages

CI	PREDmre(25)		PREDmer(25)	
	90%	95%	90%	95%
Student's t-CI	(−0.294, 0.132)	(−0.333, 0.172)	(−0.294, 0.132)	(−0.333, 0.172)
t-CIbse	(−0.316, 0.149)	(−0.364, 0.197)	(−0.311, 0.145)	(−0.359, 0.192)
CIbp	(−0.292, 0.167)	(−0.333, 0.208)	(−0.292, 0.125)	(−0.333, 0.167)
CIbca	(−0.292, 0.167)	(−0.333, 0.208)	(−0.333, 0.125)	(−0.365, 0.167)

always greater than 0.10, we cannot conclude that the OLS performs significantly better than the EbA.

Likewise, the high p-values of the Wilcoxon tests for the MdMRE, MdMER and Median(z), cannot support existence of significant difference between the two comparative models.

Since the number of predictions that satisfy the rule MRE $\leqslant$ 25% is the same for the two models, the p-value of the McNemar test is 1.

Summarizing our findings, none of the traditional parametric and non-parametric tests reveals a statistically significant difference for the two comparative models. Despite the fact that the accuracy measures show a difference for eight out of nine cases, the traditional tests do not confirm these findings, so one could infer that the two comparative models have more or less identical predictive power.

5.1.2. Bootstrap confidence intervals

From Table 4, we cannot support any significant difference for the MMRE between the analogy and the regression models. On the other hand, the variable MRE_{dif} is right-skewed and does not seem to be normal (Fig. 1a). For these reasons, the results of t-test might be inaccurate and so an alternative technique needs to be used for the comparison of the two models. The bootstrap resampling offers such an option.

After drawing the bootstrap samples we can compute the corresponding confidence intervals using the equations of Section 4.1. From Table 5, we can infer that only the 90% CIbp and the 95% and 90% CIbca show significant difference between the means since they do not contain the zero. However, the CIbca is probably the most appropriate technique for our data since the bootstrap distribution of the variable MRE_{dif} has considerable skewness (Fig. 1b).

The conclusion is that although we have observed some superiority of OLS from the MMRE measure, the t-test could not detect a real statistically significant difference.

Table 15
Significance of all paired samples permutation tests for the global accuracy measures

	Permutation test
MMRE	0.066
MdMRE	0.676
PREDmre(25)	0.770
MMER	0.224
MdMER	0.174
PREDmer(25)	0.724
MSE	0.826
Mean(z)	0.316
Median(z)	0.748

Instead, the CIbca method provided some evidence for this difference.

Another interesting issue is arisen from the construction of the 90% and 95% CIbp for the difference of the Median(z) accuracy measures (Table 6). Despite the fact that the p-value of the Wilcoxon test is greater than 0.10 (0.759), both the 90% and 95% confidence intervals do not contain the zero and we can infer that there is a significant difference. Having in mind that the Median(z) is closer to one for the EbA, we can infer some superiority of EbA compared to the OLS model.

Table 7 shows that since all the confidence intervals contain the zero, there is no significant difference between the PREDmre(25) and PREDmer(25) measures.

5.1.3. Permutation tests

Both the 90% and 95% CIbca and CIbp confidence intervals reveal that there is a significant difference between the two comparative models for the cases of the MMRE and Median(z), respectively. Permutation test is an alternative method to assess whether the difference between two means or medians could reasonably occur just by chance in a random sample. The results of the permutation tests for the accuracy measures are presented in Table 8.

Table 16
General results for the Albrecht dataset

Statistical test	Rejection of null hypothesis			
	MRE	MER	SQE	z
Paired t-test for the difference of means	No	No	No	No
90% t-CIbse for the difference of means	No	No	No	No
90% CIbp for the difference of means	Yes	No	No	No
90% CIbca for the difference of means	Yes	No	No	Yes
Permutation test for the difference of means	Yes	No	No	No
Wilcoxon test for the difference of medians	No	No	–	No
90% CIbp for the difference of medians	No	No	–	No
Permutation test for the difference of medians	No	No	–	No
McNemar test for the difference of percentages	No	No	–	–
90% t-CIbse for the difference of percentages	No	No	–	–
90% CIbp for the difference of percentages	No	No	–	–
90% CIbca for the difference of percentages	No	No	–	–
Permutation test for the difference of percentages	No	No	–	–

As far as the MMRE concerns, we could reject the null hypothesis of no difference between the two comparative models at the 0.10 level ($p = 0.098$). Furthermore, the p-value of the permutation test for the case of the Median(z) accuracy measure is lower than 0.10 (0.066) and we can also infer a significant difference. These findings are consistent with the results obtained from the bootstrap confidence intervals.

5.1.4. General results

The general results (Table 9) that are extracted from our statistical analysis is that the OLS has an improved performance compared with the EbA model in terms of MMRE, whereas the opposite is the case for the Median(z). The traditional parametric and non-parametric tests do not statistically signify these findings. On the other hand, the more robust resampling techniques verify the significant divergence of the measures for the two comparative models. As far as the MMER concerns, the paired t-test, the accurate CIbca confidence intervals, and the permutation tests for the difference of means do not signify statistically different results. Hence, the small divergence of the percentages of the accuracy measures could occur just by chance. This is also the case for the accuracy measures that are based on the SQE. Furthermore, the parametric paired t-test, the bootstrap confidence intervals and the permutation test do not signify a difference for the means of the z.

Table 18
Significance of all paired samples tests for the global accuracy measures

	Paired t-test	Wilcoxon signed rank test	McNemar test
MMRE	0.007	–	–
MdMRE	–	0.010	–
PREDmre(25)	–	–	0.035
MMER	0.390	–	–
MdMER	–	0.118	–
PREDmer(25)	–	–	0.170
MSE	0.129	–	–
Mean(z)	0.013	–	–
Median(z)	–	0.067	–

Table 19
McNemar's 2×2 table results for the PREDmre(25)

		Regression	
		0	1
Analogy	0	29 (46.8%)	17 (27.4%)
	1	6 (9.7%)	10 (16.1%)

5.2. The Albrecht dataset

The Albrecht dataset contains 24 software projects (Albrecht and Gaffney, 1983) with six independent continuous variables and a dependent continuous variable (actual effort man months). The aforementioned analysis was applied again and the basic statistics are presented in Table 10.

The global accuracy measures are clearly better for the OLS except from the cases of the MSE and Median(z) in which the EbA seems to outperform the OLS.

5.2.1. Parametric and non-parametric paired samples tests

All the p-values of the parametric and non-parametric procedures (Table 11) are greater than 0.10 and we cannot reject the null hypothesis of no difference between the statistics under consideration. The conclusion from these tests is that the OLS and EbA give similar predictions.

5.2.2. Bootstrap confidence intervals

Despite the major difference (54.46%) between the MMRE measures, the paired t-test does not signify a statistically difference between the two comparative models. However, the distribution is highly skewed at the right

Table 17
Descriptive statistics for local measures of error

	MRE_{EbA} (%)	MRE_{OLS} (%)	MER_{EbA} (%)	MER_{OLS} (%)	SQE_{EbA}	SQE_{OLS}	z_{EbA}	z_{OLS}
Mean	88.51	47.73	60.10	51.73	73,999,948	20,377,643	1.546	1.155
Median	49.27	27.84	53.75	31.62	–	–	1.038	1.070
Hit rate ($100p \leqslant 25\%$)	25.81	43.55	30.65	43.55	–	–	–	–

Table 20
Confidence intervals for the difference of means

CI	MMRE		MMER		MSE		Mean(z)	
	90%	95%	90%	95%	90%	95%	90%	95%
Student's t-CI	(0.162, 0.653)	(0.114, 0.702)	(−0.078, 0.245)	(−0.110, 0.277)	(−4,578,160, 111,822,769)	(−16,056,574, 123,301,184)	(0.134, 0.646)	(0.084, 0.697)
t-CIbse	(0.171, 0.645)	(0.124, 0.692)	(−0.073, 0.240)	(−0.104, 0.271)	(−3,304,642, 110,549,252)	(−14,551,903, 121,796,512)	(0.144, 0.636)	(0.096, 0.685)
CIbp	(0.184, 0.657)	(0.141, 0.707)	(−0.070, 0.229)	(−0.106, 0.257)	(9,438,042, 117,248,121)	(7,891,301, 140,417,346)	(0.168, 0.648)	(0.138, 0.721)
CIbca	(0.188, 0.661)	(0.151, 0.718)	(−0.072, 0.224)	(−0.110, 0.257)	(17,714,611, 166,412,183)	(14,435,905, 194,177,765)	(0.186, 0.696)	(0.162, 0.758)

(Fig. 2a) and the paired t-test is not suitable for the comparison of the means. This is also the case for the Mean(z) (Fig. 2c). Under these circumstances, it is preferable to conduct a hypothesis test for the difference of means through the construction of bootstrap confidence intervals.

Due to the fact that the skewness still remains in the bootstrap distributions of the variables MRE_{dif} and z_{dif} (Fig. 2b and d), the t-CIbse are not considered accurate enough. On the other hand, the bootstrap confidence intervals that are based on the percentiles of the bootstrap distribution give more robust results. The zero value is not contained, either in the 90% or in the 95% confidence intervals for both of the CIbp and CIbca techniques (Table 12). Furthermore, the 90% CIbca confidence interval does not contain the zero for the case of the Mean(z).

Observing the contradictory results between the paired t-test and the bootstrap confidence intervals, it is clear that there is a need for a further investigation for the cases of the MMRE and Mean(z).

Tables 13 and 14 show that since all the confidence intervals contain the zero, there is no significant difference between the measures that are based on medians and percentages, respectively.

5.2.3. Permutation tests

The p-value of the permutation test for the difference of MMREs is 0.066 (Table 15) and we can reject the null hypothesis of no difference between the two comparative models at the 0.10 level. Therefore, we can see here that the results of both resampling techniques are in accordance and support a significant difference between the two models, which was not revealed by the t-test.

On the other hand, the p-value of the permutation test for the difference of Mean(z) (0.316) agrees with the paired t-test's result. In this case, we have to rely only on the 90% CIbca in order to claim that there is a difference.

5.2.4. General results

The general result (Table 16) that is extracted from our statistical analysis is that the OLS has an improved performance compared with that of the EbA for the case of the MMRE. The paired t-test could not detect the superiority of the OLS due to the skewness of the distribution. The more robust resampling techniques statistically verify the large divergence of the means for the two comparative models.

On the other hand, the 90% CIbca for the case of the difference of Mean(z) signifies also an improved performance of the OLS, but this outcome is not verified either by the paired t-test or by the permutation test.

For the rest global accuracy measures, both the traditional and resampling techniques agree and so we can infer that the two alternative models give similar predictions.

Table 21
Confidence intervals for the difference of medians

CI	MdMRE		MdMER		Median(z)	
	90%	95%	90%	95%	90%	95%
CIbp	(0.024, 0.330)	(−0.012, 0.362)	(0.089, 0.337)	(0.064, 0.360)	(−0.109, 0.202)	(−0.137, 0.226)

Table 22
Confidence intervals for the difference of percentages

CI	PREDmre(25)		PREDmer(25)	
	90%	95%	90%	95%
Student's t-CI	(−0.302, −0.053)	(−0.327, −0.028)	(−0.265, 0.007)	(−0.292, 0.033)
t-CIbse	(−0.303, −0.051)	(−0.328, −0.027)	(−0.265, 0.007)	(−0.292, 0.034)
CIbp	(−0.307, −0.048)	(−0.323, −0.032)	(−0.258, 0.000)	(−0.290, 0.032)
CIbca	(−0.307, −0.048)	(−0.325, −0.032)	(−0.258, 0.000)	(−0.290, 0.032)

5.3. The Finnish dataset

The third dataset contains 62 projects from a commercial Finnish bank (Maxwell, 2002). In our analysis, we used the same concatenated categorical variables as Sentas et al. (2005), with few categories, instead of the original variables. There are 24 independent variables (3 continuous, 16 ordinal and 5 nominal) and the dependent variable is the actual effort hours. Since most of the predictor variables were categorical, we built a regression-ANOVA model.

Concerning the analogy model, there are certain distance metrics that are used when the variables in the dataset are not all of the same type. For the Finnish dataset we used the *daisy* function available in the statistical software SPLUS (Insightful Corporation, 2001).

From the statistics of Table 17, it is clear that all the measures except from the Median(z) are better for the OLS.

5.3.1. Parametric and non-parametric paired samples tests

The parametric paired t-test (Table 18) signifies a difference for the cases of the MMRE and Mean(z). The Wilcoxon tests show that the two comparative models give significantly different results for the cases of the MdMRE and Median(z).

Table 23
Significance of all paired samples permutation tests for the global accuracy measures

	Permutation test
MMRE	0.004
MdMRE	0.088
PREDmre(25)	0.036
MMER	0.402
MdMER	0.026
PREDmer(25)	0.156
MSE	0.006
Mean(z)	0.006
Median(z)	0.692

Analyzing the McNemar's 2×2 table (Table 19), we can observe that the OLS model predicts more accurately 17 projects in terms of the PREDmre(25), whereas the EbA only 6 of the 62 projects. The discordant cells contain a satisfactory number of observations and the p-value is 0.035. Hence, we can reject the null hypothesis that the proportions of positive results are the same for the two comparative models.

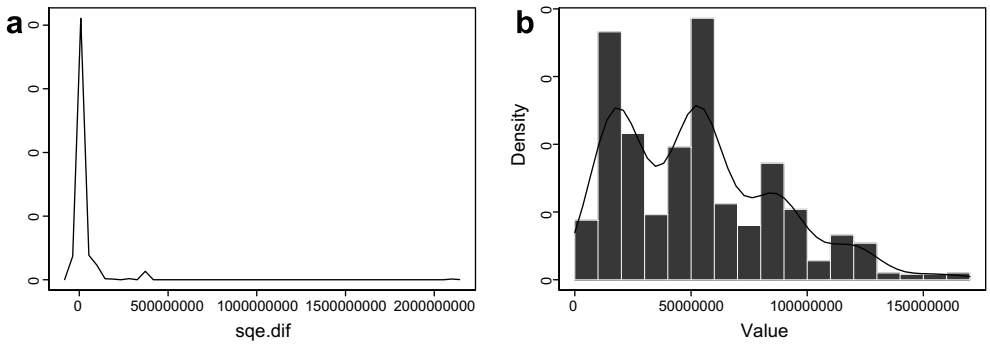

Fig. 3. (a) Density and (b) bootstrap estimation for the difference of the SQE_{dif}.

Table 24
General results for the Finnish dataset

Statistical test	Rejection of null hypothesis			
	MRE	MER	SQE	z
Paired t-test for the difference of means	Yes	No	No	Yes
90% t-CIbse for the difference of means	Yes	No	No	Yes
90% CIbp for the difference of means	Yes	No	Yes	Yes
90% CIbca for the difference of means	Yes	No	Yes	Yes
Permutation test for the difference of means	Yes	No	Yes	Yes
Wilcoxon test for the difference of medians	Yes	No	–	Yes
90% CIbp for the difference of medians	Yes	Yes	–	No
Permutation test for the difference of medians	Yes	Yes	–	No
McNemar test for the difference of percentages	Yes	No	–	–
90% t-CIbse for the difference of percentages	Yes	No	–	–
90% CIbp for the difference of percentages	Yes	No	–	–
90% CIbca for the difference of percentages	Yes	No	–	–
Permutation test for the difference of percentages	Yes	No	–	–

Having in mind the large divergence of the global accuracy measures, we can infer that the OLS gives more accurate predictions as the traditional procedures can detect statistically significant different results for four out of nine tests. On the other hand, the EbA outperforms the OLS ($p = 0.067$) for the case of the Median(z). The resampling techniques should be utilized in order to confirm the aforementioned findings. Furthermore, they might also reveal some other differences that the traditional tests could not detect due to the potential skewness of the data.

5.3.2. Bootstrap confidence intervals

Tables 20–22 show that the bootstrap confidence intervals for MMRE, MdMRE, PREDmre(25) and Mean(z) give the same results with the corresponding parametric and non-parametric tests and verify that the OLS gives more accurate predictions.

Moreover, in three cases the results of traditional tests are not consistent with the bootstrap confidence intervals. Specifically:

- The 90% and 95% CIbp in Table 21 do not contain the zero, so we assess a significant difference between the medians of MER for the two comparative models.
- Both the 90% and 95% CIbp and CIbca in Table 20 do not contain the zero, so we can reject the null hypothesis of no difference between the OLS and the EbA for the MSE. This is probably due to the high skewness of the SQE_{dif} (Fig. 3a). Since, the skewness still remains in the bootstrap distribution of the SQE_{dif} (Fig. 3b) the CIbca can be considered more accurate.
- The 90% and 95% CIbp (Table 21) contain the zero for the Median(z), showing no significant difference.

Table 26
Significance of all paired samples tests for the global accuracy measures

	Paired t-test	Wilcoxon signed rank test	McNemar test
MMRE	0.143	–	–
MdMRE	–	0.043	–
PREDmre(25)	–	–	1
MMER	0.434	–	–
MdMER	–	0.418	–
PREDmer(25)	–	–	1
MSE	0.145	–	–
Mean(z)	0.041	–	–
Median(z)	–	0.003	–

5.3.3. Permutation tests

The p-value of the permutation test for the difference of the medians of MER is 0.026 (Table 23). This result is consistent with the bootstrap confidence intervals. On the other hand, the Wilcoxon test could not detect the difference between the medians. Observing the large divergence of the MdMER values, the p-value of the Wilcoxon test (0.118) which is close to the rejection area (0.10) and the findings of the resampling techniques that reveal a significant difference, we can claim that the OLS seems to give more accurate predictions.

The p-value of the permutation test is lower than 0.05 for the MSE and verifies the result of CIbp and CIbca. This difference was not revealed by the paired t-test procedure.

Furthermore, the p-value of the permutation test for the difference of the Median(z) is 0.692 so we cannot reject the null hypothesis of no difference. Summarizing our findings for the Median(z), we can notice that only the Wilcoxon test signifies a difference between the medians, a result which is not supported by the resampling methods.

5.3.4. General results

From Table 24, we can observe that both the traditional and the resampling techniques show that the OLS outperforms the EbA model for MMRE, MdMRE, PREDmre(25) and Mean(z). Moreover, the superiority of the OLS can only be detected by the resampling techniques for the MdMER and MSE. On the other hand, the Wilco-

Table 25
Descriptive statistics for local measures of error

	MRE_{EbA} (%)	MRE_{OLS} (%)	MER_{EbA} (%)	MER_{OLS} (%)	SQE_{EbA}	SQE_{OLS}	z_{EbA}	z_{OLS}
Mean	71.85	56.02	77.76	68.64	681,358	581,655	1.454	1.225
Median	59.08	40.62	46.60	34.49	–	–	1.336	1.150
Hit rate ($100p \leqslant 25\%$)	27.45	27.45	25.49	27.45	–	–	–	–

Table 27
Confidence intervals for the difference of means

CI	MMRE		MMER		MSE		Mean(z)	
	90%	95%	90%	95%	90%	95%	90%	95%
Student's t-CI	(−0.035, 0.337)	(−0.055, 0.372)	(−0.103, 0.285)	(−0.141, 0.324)	(−13263.3, 212669.0)	(−35686.1, 235091.9)	(0.046, 0.412)	(0.010, 0.449)
t-CIbse	(−0.013, 0.329)	(−0.047, 0.364)	(−0.093, 0.256)	(−0.128, 0.291)	(−13330.4, 212736.2)	(−35806.7, 235212.5)	(0.042, 0.416)	(0.005, 0.453)
CIbp	(−0.012, 0.323)	(−0.052, 0.367)	(−0.079, 0.259)	(−0.117, 0.306)	(−1766.1, 218243.1)	(−11375.8, 242841.6)	(0.043, 0.411)	(0.005, 0.441)
CIbca	(−0.019, 0.332)	(−0.055, 0.357)	(−0.068, 0.281)	(−0.090, 0.333)	(18536.1, 254874.7)	(5810.1, 302478.5)	(0.030, 0.406)	(0.001, 0.436)

xon test signifies a superiority of the EbA for the case of the Median(z), a result that is not verified by either of the resampling techniques.

5.4. The ISBSG7 dataset

The ISBSG is a non-profit organization that helps software developers to produce more accurate predictions. The ISBSG7 (ISBSG, 2001) dataset contains 1239 projects but the initial database is considerably reduced due to the large number of missing values and the different methods of measuring the work effort and size of projects. It contains 11 independent variables (1 continuous and 10 nominal), whereas the dependent variable that is used to construct prediction models is productivity. The procedure that is followed for the selection of the appropriate data and the concatenation of the initial variables were presented in Sentas et al. (2005).

From Table 25, we can clearly notice that OLS outperforms EbA for all except one accuracy measure (PREDmre(25)), in which the percentages are equal. The next issue that we have to deal with, is the significance of the differences between the global accuracy measures.

5.4.1. Parametric and non-parametric paired samples tests

Regarding the p-values of the parametric and non-parametric procedures (Table 26), we can infer that the Mean(z), MdMRE and Median(z) give statistically different results for the two comparative models.

5.4.2. Bootstrap confidence intervals

The bootstrap intervals provide similar results with the traditional tests. However, there are two cases where there is a disagreement:

- For MSE, the 90% and 95% CIbca (Table 27) do not contain the zero so they give some indication of difference between the OLS and EbA.
- The 90% and 95% CIbp (Table 28) contain the zero value while the Wilcoxon test reports a statistically significant difference for the medians of z.

These contradictory results need further analysis.

Table 29 shows that since all the confidence intervals contain the zero, there is no significant difference between the PREDmre(25) and PREDmer(25) measures.

5.4.3. Permutation tests

The p-values of permutation tests (Table 30) are close to the corresponding p-values of the parametric and non-parametric procedures. It is interesting to note that in the two aforementioned contradictory cases, the permutation tests confirmed the traditional tests.

5.4.4. General results

Summarizing our findings (Table 31), the difference between the accuracy measures that are based on MdMRE

Table 28
Confidence intervals for the difference of medians

CI	MdMRE		MdMER		Median(z)	
	90%	95%	90%	95%	90%	95%
CIbp	(0.023, 0.401)	(−0.020, 0.426)	(−0.043, 0.192)	(−0.074, 0.197)	(−0.009, 0.505)	(−0.028, 0.553)

Table 29
Confidence intervals for the difference of percentages

CI	PREDmre(25)		PREDmer(25)	
	90%	95%	90%	95%
Student's t-CI	(−0.133, 0.133)	(−0.159, 0.159)	(−0.164, 0.125)	(−0.193, 0.154)
t-CIbse	(−0.135, 0.135)	(−0.162, 0.162)	(−0.165, 0.126)	(−0.194, 0.155)
CIbp	(−0.118, 0.137)	(−0.157, 0.176)	(−0.177, 0.118)	(−0.196, 0.157)
CIbca	(−0.118, 0.137)	(−0.157, 0.177)	(−0.157, 0.118)	(−0.196, 0.157)

Table 30
Significance of all paired samples permutation tests for the global accuracy measures

	Permutation test
MMRE	0.14
MdMRE	0.04
PREDmre(25)	1
MMER	0.500
MdMER	0.244
PREDmer(25)	0.926
MSE	0.212
Mean(z)	0.028
Median(z)	0.098

Table 31
General results of ISBSG7 dataset

Statistical test	Rejection of null hypothesis			
	MRE	MER	SQE	z
Paired t-test for the difference of means	No	No	No	Yes
90% t-CIbse for the difference of means	No	No	No	Yes
90% CIbp for the difference of means	No	No	No	Yes
90% CIbca for the difference of means	No	No	Yes	Yes
Permutation test for the difference of means	No	No	No	Yes
Wilcoxon test for the difference of medians	Yes	No	–	Yes
90% CIbp for the difference of medians	Yes	No	–	No
Permutation test for the difference of medians	Yes	No	–	Yes
McNemar test for the difference of percentages	No	No	–	–
90% t-CIbse for the difference of percentages	No	No	–	–
90% CIbp for the difference of percentages	No	No	–	–
90% CIbca for the difference of percentages	No	No	–	–
Permutation test for the difference of percentages	No	No	–	–

show that the OLS outperforms the EbA, whereas it seems that there is not a significant difference between the means of the two comparative models and these results are statistically verified from the traditional and the resampling techniques. In the case of MER, both traditional tests and resampling techniques statistically signify that there is no difference between the accuracy measures of the two comparative models. The results are not clear for the case of the SQE local accuracy measure. The regression model seems to outperform the analogy but this result is only verified by the construction of the 90% CIbca confidence interval. Finally, the traditional and resampling techniques evidence that the OLS gives more accurate results for the case of the z.

6. Conclusions

In this paper, we examined a crucial issue in the software cost estimation area, concerning the selection of the "best" model between two comparative models. More precisely, we considered two prediction methods; the estimation by analogy and the regression analysis. However, our purpose was not to conclude about the superiority of the one prediction method against the other, but rather to show how formal comparisons can be performed using alternative statistical techniques.

The extensive examination of nine accuracy measures that have been proposed in the literature showed that the comparison of the accuracy of cost estimation methods should not be based just on the accuracy indicators but it is necessary to evaluate their differences through statistical procedures. The traditional parametric and non-parametric procedures offer such options. On the other hand, in some circumstances, traditional methods might lead to erroneous inference when the dataset is considerably small and skewed or when the parametric assumptions do not hold.

Alternatively, two computer intensive techniques can be used in order to obtain reliable and accurate results. In particular, we utilized the bootstrap and the permutation tests that are free from the normality assumptions. These techniques repeat the data analysis a large number of times on replicated datasets, all drawn by resampling from the original observed set of data. The resampling techniques can be used on their own in carrying out a hypothesis test

without worrying about the distribution of the variables or they can also be utilized with the traditional procedures in order to reinforce their results.

Regarding the extensive experimentation that we performed by applying several comparison tests to the four datasets, we have to discuss some validity issues. First of all, the applications are used as means for illustrating that we cannot rely solely on any accuracy measure for taking decisions about which model is the best. In that sense, the results of the comparisons we made are not generalised to any population of projects, but on the contrary, they show that for different datasets and different accuracy measures the results can be quite contradictory. Thus, the various validity issues, usually raised when we test hypotheses, are addressed in our research by the large variety of accuracy measures, the different datasets, especially the ISBSG dataset which is multi-organizational, and most important by the plethora of confidence intervals and tests we implemented. Furthermore, the limitations of each statistical test used for comparison is discussed explicitly. These limitations are related either to the underlying theoretical distribution of the original sample or the distribution of the resampling estimates. For example, the resampling techniques are subject to two sources of variation: the randomness of drawing the original sample from the population and the randomness of the repeated sampling from the original sample. However, this added variation is considered small and can be overcome by increasing the number of the repeated samples.

Some interesting issues arisen from our work deserve further research. For example, the introduction of other bootstrap methods and the systematic identification of differences between several cost estimation methods. Since there are a lot of models that can be fitted to a certain dataset, the point is the selection of the ''best'' model through formal statistical procedures.

References

Abran, A., Robillard, P.N., 1996. Function point analysis: an empirical study of its measurement processes. IEEE Trans. Softw. Eng. 22 (12), 895–909.

Albrecht, A.J., Gaffney, J.E., 1983. Software function, source lines of code, and development effort prediction: a software science validation. IEEE Trans. 6, 639–648.

Angelis, L., Stamelos, I., 2000. A simulation tool for efficient analogy based cost estimation. Empirical Softw. Eng. 5, 35–68.

Briand, L., Langley, T., Wieczorek, I., 2000. A replicated assessment and comparison of common software cost modeling techniques. In: Proceedings of the International Conference on Software Engineering (ICSE 22), pp. 377–386.

Conte, S., Dunsmore, H., Shen, V.Y., 1986. Software Engineering Metrics and Models. Benjamin Cummings, Menlo Park, CA.

Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman & Hall.

Insightful Corporation, 2001. SPLUS 6 for Windows, Guide to statistics, vol. 2. Insightful Corporation, Seattle, Washington.

ISBSG, 2001. Data Disk, Release 7, June.

Jeffery, R., Ruhe, M., Wieczorek, I., 2001. Using public domain metrics to estimate software development effort. In: Proceedings of the METRICS 2001 Conference, pp. 16–27.

Kitchenham, B., 1998. A procedure for analyzing unbalanced datasets. IEEE Trans. Softw. Eng. 24 (4).

Kitchenham, B., MacDonell, S.G., Pickard, L.M., Shepperd, M.J., 2001. What accuracy statistics really measure. IEEE Proc. Softw. 148 (3), 81–85.

Mair, C., Kadoda, G., Lefley, M., Phalp, K., Schofield, C., Shepperd, M., Webster, S., 2000. An investigation of machine learning based prediction systems. J. Syst. Softw. 53, 23–29.

Maxwell, K., 2002. Applied Statistics for Software Managers. Prentice-Hall, Englewood Cliffs, NJ.

Moore, D.S., McCabe, G.P., Duckworth II, W.D., Sclove, S.L., 2003. The Practice of Business Statistics. W.H. Freeman and Company, New York.

Myrtveit, I., Stensrud, E., 1999. A controlled experiment to assess the benefits of estimating with analogy and regression models. IEEE Trans. Softw. Eng. 25 (4), 510–525.

Myrtveit, I., Stensrud, E., Shepperd, M., 2005. Reliability and validity in comparative studies of software prediction models. IEEE Trans. Softw. Eng. 31 (5), 380–391.

Sentas, P., Angelis, L., Stamelos, I., Bleris, G., 2005. Software productivity and effort prediction with ordinal regression. Inform. Softw. Technol. Trans. 47, 17–29.

Shan, Y., McKay, R., Lokan, C., Essam, D., 2002. Software project effort estimation using genetic programming. In: Proceedings of the ICCCAS'02 International Conference on Communications, Circuits and Systems, Chengdu, China, July, pp. 1108–1112.

Shepperd, M., Schofield, C., 1997. Estimating software project effort using analogies. IEEE Trans. Softw. Eng. 23 (12), 736–743.

Sheskin, D.J., 2004. Handbook of Parametric and Nonparametric Statistical Procedures, third ed. Chapman & Hall/CRC.

Shukla, K.K., 2000. Neuro-genetic prediction of software development effort. Inform. Softw. Technol. 42, 701–713.

Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S, fourth ed. Springer-Verlag, New York.

Nikolaos Mittas received his B.Sc. in Mathematics from University of Crete, Greece and M.Sc. degree in Informatics from Aristotle University of Thessaloniki (A.U.Th.). He is currently a Ph.D. student, working on statistical methods with applications to software engineering.

Lefteris Angelis received his B.Sc. and Ph.D. degree in Mathematics from Aristotle University of Thessaloniki (A.U.Th.). He is currently an Assistant Professor at the Department of Informatics of A.U.Th.. His research interests involve statistical methods with applications in information systems and software engineering, computational methods in mathematics and statistics, planning of experiments and simulation techniques.