Empir Software Eng (2010) 15:523–555
DOI 10.1007/s10664-010-9128-6

LSEbA: least squares regression and estimation by analogy in a semi-parametric model for software cost estimation

Nikolaos Mittas · Lefteris Angelis

Published online: 6 May 2010

Editor: Emilia Mendes

Abstract The importance of Software Cost Estimation at the early stages of the development life cycle is clearly portrayed by the utilization of several models and methods, appeared so far in the literature. The researchers' interest has been focused on two well known techniques, namely the parametric Regression Analysis and the non-parametric Estimation by Analogy. Despite the several comparison studies, there seems to be a discrepancy in choosing the best prediction technique between them. In this paper, we introduce a semi-parametric technique, called LSEbA that achieves to combine the aforementioned methods retaining the advantages of both approaches. Furthermore, the proposed method is consistent with the mixed nature of Software Cost Estimation data and takes advantage of the whole pure information of the dataset even if there is a large amount of missing values. The paper analytically illustrates the process of building such a model and presents the experimentation on three representative datasets verifying the benefits of the proposed model in terms of accuracy, bias and spread. Comparisons of LSEbA with linear regression, estimation by analogy and a combination of them, based on the average of their outcomes are made through accuracy metrics, statistical tests and a graphical tool, the Regression Error Characteristic curves.

Keywords Estimation by analogy · Regression analysis · Semi-parametric model · Software cost estimation

1 Introduction

The plethora of *Software Cost Estimation* (SCE) models proposed in the literature (Jorgensen and Shepperd 2007) reveals that the prediction of the cost for a new software

N. Mittas · L. Angelis (✉)
Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki 54124, Greece
e-mail: lef@csd.auth.gr

N. Mittas
e-mail: nmittas@csd.auth.gr

project is a vital task affecting the well-balanced management of the development process. The overestimation of a project may lead to the canceling and loss of a contract, whereas the underestimation may affect the earnings of the development organization. Hence, there is an ongoing research in the SCE area attempting to build prediction models that provide accurate estimates of the cost.

The methods appeared so far in the literature, vary from expert judgment techniques to statistical and machine learning models. Several studies investigate issues of accuracy and efficacy trying to assess the benefits and drawbacks of each method and contributing to the selection of the "best" prediction technique. For a comprehensive account of the research on cost estimation techniques appeared in literature so far, we refer to the systematic review by Jorgensen and Shepperd (2007).

Although there is a practical need for a manager to decide whether a model predicts accurately the cost of a new project, there seems to be no global answer. There are contradictory results and one reason for this, is the lack of standardization in software research methodology which leads to heterogeneous sampling, measurement and reporting techniques. Another reason is the appropriateness of the prediction techniques on the available data (Mair and Shepperd 2005). Indeed, linear models can be fitted efficiently in cases where there are predictors linearly associated with the response variable. Most often, the relationship becomes linear after some simple transformation of the original variables. On the other hand, non-linear models also seem to be a plausible choice since the software cost datasets contain significant information in categorical form (*nominal* and *ordinal* variables).

Despite the fact that there are several forms of Regression Analysis, *Ordinary Least Squares Regression* appears to be one of the most popular techniques. It is used for fitting a linear parametric model for the cost variable by minimizing the sum of squared residuals (Jorgensen and Shepperd 2007; Mair and Shepperd 2005). Although the specific form of parametric model is widely applicable, its main drawback remains the requirement that both the structural model and the error distribution have to be correctly specified. Furthermore, there are many assumptions that have to hold in order to apply the procedure and certain limitations in the manipulation of categorical attributes that are usually contained in the SCE datasets (see Section 3.2).

Estimation by Analogy (EbA) is an alternative method that can be used to obtain predictions. It is based on finding a few similar or neighbor projects to the one that has to be estimated, imposing no restrictions on the data (Shepperd and Schofield 1997). Contrary to regression analysis, EbA is free of strict distribution assumptions and it can be easily applied to all types of data (numerical and categorical). Despite the simplicity in the application of EbA and the intuitively appealing interpretation of the whole procedure (Shepperd and Schofield 1997), the method involves several design decisions. It also requires specific adjustments (such as the number of neighbors, the distance metric etc) that have to be examined in order to calibrate the procedure and produce accurate predictions (see Section 3.1).

The main goal of this study is to further extend a previous conference paper (Mittas and Angelis 2008a), concerning the possibility of improving the estimation of the cost function by a model that combines Least Squares (LS) regression and EbA. We call this new method LSEbA in order to show the mixing of LS and EbA in a single model. The methodology is inspired by the formal statistical model known as "*Partially Linear Model*" (Hardle et al. 2000) and can handle in a sophisticated manner the parametric and non-parametric information exploiting the merits of both regression and EbA.

Having in mind the mixed-type nature of SCE datasets, the proposed semi-parametric model seems to be a more reasonable approach. Indeed, in datasets with software projects there is a certain set of continuous variables that are linearly correlated with the cost. However, there are also several categorical variables having significant impact on effort which cannot be expressed through a linear relationship. In general, the proposed methodology addresses a number of problems encountered in the construction of a cost estimation model including problems with the mixed type of data and the missing values.

The remainder of the paper is organized as follows: In Section 2, we summarize related work and discuss the contribution of the paper. Section 3 presents LS regression and EbA in the form of parametric and non-parametric modeling techniques as applied in SCE. Section 4 presents LSEbA as a semi-parametric model and describes the algorithm for predicting the cost of a new project. In Section 5, we describe the methodology applied for the statistical and graphical comparison of LSEbA with LS, EbA and a combination of them based on their average, using three datasets. In Section 6, we present the results of the experimentation and finally in Section 7 we conclude with a discussion of the results and some directions for future research.

2 Related Work and Contribution

During the last decades, the crucial issue that SCE community had to face was the necessity for the development of accurate cost estimation models. The problem remains critical today, due to the continuous changes in the organizations' requirements and the evolvement of the software development processes. The interest of the research community in the subject is reflected to the large number of new methods and improvements of classical methods that appear annually in the literature.

The systematic review by Jorgensen and Shepperd (2007) considered 304 studies describing research on SCE. The distribution of topics revealed that the most common research topic is the introduction and evaluation of estimation methods. Moreover, it was found that regression-based techniques prevail in terms of their frequency of application. The application of analogy-based methods seems to steadily increase. These results, along with the findings of another systematic review by Mair and Shepperd (2005), support the belief that LS regression and EbA are still two of the most widely used cost estimation methodologies.

Despite the evolving research dealing with the performance of regression and EbA, the question about the superiority of one method against the other remains an open problem. The systematic review of 20 studies by Mair and Shepperd (2005) which used regression and analogy-based techniques, covering the period from 1997 to 2004, reveals that there is approximately equal evidence in favor of and against analogy-based techniques. More precisely, it is mentioned that 45% of the studies offer some support for EbA, 35% for regression while 20% of the studies do not conclude about the superiority of a method against the comparative one. Mair and Shepperd also pointed out that only six of the papers (30%) made use of the traditional hypothesis testing in order to evaluate the predictive performance of the two comparative models. In these formal comparisons, four out of six studies indicated the superiority of regression, one study showed the superiority of EbA, whereas in one study there was no significant difference between the models. The researchers discussed as potential sources of disagreement in the findings the different dataset characteristics, which may favor different prediction systems.

The present study was motivated from the aforementioned results of the literature review and from the accuracy results obtained after trying different prediction methods on various data. The research question we posed was whether there is a unified way to handle simultaneously the linear and the nonlinear relationships in a dataset by combining LS regression and EbA.

The semi-parametric model proposed in this paper is based on an expression similar to the partially linear regression (Hardle et al. 2000). The main modification is the way in which the non-parametric part of the model is evaluated. The partially linear regression in its general form utilizes a quite complicated non-parametric procedure, namely the kernel weights, in order to estimate the non-linear portion of the model. Since EbA is simpler, widely spread in SCE area and can be applied to categorical data, we use this form of non-parametric approach for the evaluation of the semi-parametric model.

One of the major problems addressed by the proposed methodology is the modeling of several categorical variables appearing in a dataset. Angelis et al. (2001) point out that traditional prediction models are based mainly on numerical features and the presence of categorical attributes causes various problems in the suitability, reliability, interpretability and robustness of the method (see Section 3.2). Due to this fact, *Categorical Regression* is a possible approach for the quantification and manipulation of categorical variables (Angelis et al. 2001). However, significant drawbacks of the methodology are reported, such as the extensive statistical preprocessing of the categorical attributes and the interpretation of the quantified variables.

A relevant problem, arising during the construction of parametric prediction models, is that datasets contain many unbalanced factors (i.e. the possible combinations of all the levels of the categorical variable do not appear in a dataset). Kitchenham (1998) proposed a methodology, called *Forward Pass Residual Analysis*, in order to overcome this restriction. The method consists of a series of consecutive parametric ANOVA models, in which each model is built on the residuals of the previous one. The method is highly interactive and requires the construction of several models (Kitchenham 1998).

Although the aforementioned studies provide alternative choices for the manipulation of categorical variables, the most known technique is the definition of dummy variables (Briand et al. 2000; Lokan and Mendes 2006; Mendes et al. 2005). Lokan and Mendes (2006) discussed the problem arisen from the definition of dummy variables and the difficulties of merging the categorical levels, especially in small datasets.

Another important issue addressed by LSEbA is the large amount of missing values usually contained in datasets. This is a very serious problem, especially for the parametric models, since the most common techniques involve the deletion of cases or variables with missing values. Despite the simplicity of the deletion techniques, the major disadvantages are the dramatic loss of information in cases with high percentages of missing values and the possible bias in the data when there is some type of pattern in the missing data. The task of imputation is another solution but it requires extensive and complicated analysis with advanced statistical or probabilistic tools (Cartwright et al. 2003; Myrtveit et al. 2001; Strike et al. 2001), whereas it also constitutes a source of additional prediction error.

Finally, the present study extends preliminary experimentation conducted on two small datasets with only numerical variables (Mittas and Angelis 2008a) where we showed that the semi-parametric model decreases significantly the error compared to the LS regression and EbA.

In the present study we first provide in detail the rationale, the theoretical background and the description of the proposed methodology and then we illustrate the application of

LSEbA on three different SCE datasets. These datasets incorporate common problems in SCE where there is a difficulty to decide on the prediction model. Specifically, the three datasets are representative of the following problems: (a) availability of only a few categorical and continuous variables, (b) availability of a large number of categorical variables and (c) availability of a several categorical variables but with many missing values.

The aforementioned datasets are used to compare LSEbA not only with LS and EbA separately, but also with another method that combines LS and EbA by computing the average of the estimates of these two methods. The decision to compare LSEbA with the average of LS and EbA was based on the work of MacDonell and Shepperd (2003) where the authors discuss the issue of optimizing effort predictions by combining techniques such as LS and EbA. One such technique is the average which is also referred as an approved estimation method in the work by Kitchenham et al. (2002).

The proposed LSEbA method is essentially a mixed model in the sense that it combines two models with different principles, a parametric and a non-parametric one. Their combination, although resulting in an additive model, is the result of a two-step procedure which uses a partition of the variables in two subsets. Since it is generally easy to construct a mixed model by averaging the results of any two or more methods (even LSEbA could participate in such an averaging procedure with any other methods) we decided that it would be interesting to compare two mixed models with exactly the same components, i.e. LSEbA with the simpler average of LS and EbA.

For the comparisons, we use well-known accuracy metrics, statistical tests and also graphical representation through *Regression Error Characteristic* (REC) curves, a recently proposed visualization tool (see Section 5.2). REC curves graphically present certain summary statistics (median or pred25) of errors measured by different functions. Therefore, they offer a straightforward basis for visual comparison of different models.

3 Modeling Techniques

3.1 EbA as Non-Parametric Regression

EbA can be easily portrayed as a procedure conducted in three steps: First, the project to be estimated is characterized by a set of attributes common to the ones characterizing completed projects in a historical database. Second, a predefined similarity criterion is used in order to retrieve one or more similar projects. The final decision-making concerns the reuse of knowledge derived from the similar cases or *analogies* by combining the cost values to produce the estimate of the new project.

Despite the fact that EbA seems to be a case-based reasoning technique (Shepperd and Schofield 1997), Mittas et al. (2008) showed that EbA can be expressed in the form of a mathematical model. In the statistical literature, the method is known as *Nearest Neighbor Non-parametric Regression* and appears to be an alternative choice to the more traditional parametric regression models.

We denote by Y the real random dependent variable representing in our context the cost of a project and by $\mathbf{X}$ a P-dimensional random vector with coordinates representing the variables or attributes which characterize the projects in the available historical data set. The attributes of $\mathbf{X}$ may come from various types of distributions, either discrete or continuous and these are assumed to be unknown.

Our goal is to find a regression function $f(\mathbf{X}_i) = E(Y_i|\mathbf{X}_i)$, (by $E(Y_\mathrm{i}/\mathbf{X}_\mathrm{i})$ we denote the conditional expected value of random variable Y_i given the random vector $\mathbf{X}_i$) which will serve for building a prediction model of the form

$$Y_i = f(\mathbf{X}_i) + \varepsilon_i \quad E(\varepsilon_i|\mathbf{X}_i) = 0 \quad (i = 1, \dots, n). \tag{1}$$

Y_i is the i-th observation of the dependent variable (i.e. the cost for the i-th project) to be explained by variations in the vector of predictors $\mathbf{X}_i$ with dimensions $1 \times p$. We assume that the random errors ε_i are independent with zero mean.

In the statistical literature there have been proposed various non-parametric approaches for the estimation of $f(\mathbf{X})$, such as the partitioning estimate, the kernel estimate and the k–nearest neighbor estimate (Hardle 1990). Non-parametric models use a more general and robust functional form that is not known in advance (unlike the parametric linear model) and therefore they have the ability to detect a complicated non-linear structure which sometimes remains undetected by traditional parametric estimation techniques.

As the scope of this study is the construction of a model that utilizes both regression and EbA, we have decided to focus on the nearest neighbor approach since it is the most closely related to the EbA method. An additional advantage of this method is that it is based on distance measures for finding the nearest neighbors and these measures can be easily calculated even for categorical variables.

Given the dataset with n observations (or projects), a k–NN estimate of $f(\mathbf{X}_i)$ can be evaluated by

$$\hat{f}(\mathbf{X}_i) = \frac{\sum\limits_{j=1}^{n} W_{kj}(\mathbf{X}_i) Y_j}{\sum\limits_{j=1}^{n} W_{kj}(\mathbf{X}_i)} \tag{2}$$

where $\left\{W_{kj}(\mathbf{X}_i)\right\}_{j=1}^{n}$ is a sequence of weights defined through the set of indices $J_{\mathbf{x}_i}$={j: $\mathbf{X}_j$ is one of the k nearest neighbors of $\mathbf{X}_i$}.

With this set of indices of neighbors (or else analogies), the simplest k-NN sequence of weights is constructed by

$$W_{kj}(\mathbf{X}_i) = \begin{cases} 1/k, & \text{if } j \in J_{\mathbf{x}_i} \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

There are three major issues that have to be inquired in order to calibrate the model. The first issue is the selection of an appropriate similarity function in order to find the most relevant projects. As historical datasets contain various types of variables that have to be treated with a different manner, we have to use a distance metric that takes into account the mixed-type variables. Hence, we used a special dissimilarity coefficient suggested by Kaufman and Rousseeuw (1990). We denote by $d\ (i, j)$ the distance or dissimilarity of projects i and j which will be computed by their vectors of attributes $\mathbf{X}_i$=(X_{i1},...,X_{ip}) and $\mathbf{X}_\mathrm{j}$=(X_{j1},...,X_{jp}), respectively. Then:

$$d(i,j) = \frac{\sum\limits_{m=1}^{p} \delta_{i,j}{}^{(m)} d_{i,j}{}^{(m)}}{\sum\limits_{m=1}^{p} \delta_{i,j}{}^{(m)}} \tag{4}$$

where:

$$\delta_{i,j}{}^{(m)} = \begin{cases} 1 \text{ if } X_{im}, X_{jm} \text{ nonmissing} \\ 0 \text{ otherwise} \end{cases} \tag{5}$$

- If the *m*-th variable is binary or nominal, the following formula can be used to compute dissimilarities:

$$d_{i,j}{}^{(m)} = \begin{cases} 1 \text{ if } X_{im} \neq X_{jm} \text{ nonmissing} \\ 0 \text{ if } X_{im} = X_{jm} \end{cases} \tag{6}$$

- If the *m*-th variable is of interval or ratio scale, the following formula can be used to compute dissimilarities:

$$d_{i,j}{}^{(m)} = \frac{|X_{im} - X_{jm}|}{R_m} \tag{7}$$

where $R_m = \max\limits_{1 \le i \le n}(X_{im}) - \min\limits_{1 \le i \le n}(X_{im})$ is the range of the variable.

- If the *m*-th variable is ordinal, the dissimilarities are computed as follows:

  1. The values X_{im} are replaced by their ranks $r_{im} \in \{1,\ldots,M_m\}$
  2. The ranks are transformed to values in the interval [0,1]: $z_{im} = \frac{r_{im}-1}{M_m-1}$
  3. The z_{im} values are treated as interval-scaled.

From the evaluation of the dissimilarity coefficient through Eq. 4, it is clear that there is a special treatment when a dataset has missing values. In these cases, the dissimilarity coefficient does not take into account the feature that contains a missing value (Eq. 5), while at the same time the vector with the missing values is not completely ignored, i.e. the non-missing information participates in the calculation of the coefficient. Furthermore, the division with the variable range in Eq. 7 ensures that the variables participating in the calculation of the coefficient are standardized.

Another important decision for EbA is the number of the nearest neighbors that the practitioner has to combine. Since EbA procedure needs to be locally calibrated on the available data, there is no rule of thumb to decide upon the number of analogies without experimentation.

The meaning of Eq. 2 is that after the selection of the neighbor projects, we need to compute a certain statistic based on weights assigned to the values of the dependent variable and which may represent an efficient approximation of the unknown function. Although a variety of statistics can be calculated, the selection of the arithmetic mean, essentially by the weights defined in Eq. 3, is a plausible choice.

The fact that users are more willing to accept solutions from analogy based techniques, rather than solutions derived from complicated prediction systems (Shepperd and Schofield 1997) renders EbA as a popular methodology for estimating the cost of a new software project. Although EbA is a more flexible model, it may neglect some existing linear dependencies between the response and predictors. Additionally, the method itself may require specific adjustments (i.e. choice of analogues projects) according to the needs of the new projects. Finally, although non-parametric regression is considered a more robust method, it is not free from precision problems (Anglin and Gencay 1996). Another inherent problem of the method is that the cost estimations of new projects are restricted by the range of the costs in the available dataset, since each new cost is estimated by averaging few costs of similar projects within the dataset. Parametric and semi-parametric models do not have this restriction since the independent variables participate in the computation of cost estimations.

3.2 Parametric Regression

Contrary to the non-parametric regression, parametric estimation techniques assume that the function $f(\mathbf{X}_i)$ of Eq. 1 can be estimated by a linear expression of $\mathbf{X}_i$,

$$f(\mathbf{X}_i) = \mathbf{X}_i\beta \tag{8}$$

where $\boldsymbol{\beta}$ is $p\times1$ vector of unknown parameters, the regression coefficients. Hence, it is assumed that the dependent variable is functionally expressed by the predictors and unobservable errors according to a function given by an explicit formula.

There are many techniques for the estimation of the explicit function $f(\mathbf{X})$, i.e. the estimation of its regression coefficients, which are essentially methods for the minimization of the total error. However, ordinary least squares regression, where the regression coefficients are estimated from the data by minimizing the overall sum of squared errors, is the most common method in SCE.

Despite the popularity of the method and the large amount of studies dealing with this specific form, there are many assumptions to consider when fitting a parametric model. The most important ones are: (a) the relationship between each of the predictor variables and the dependent variable is linear and (b) the residuals are normally distributed and uncorrelated with the predictors. These two assumptions are usually addressed in the case of software project data by using the logarithmic transformations of the cost (dependent) variable and the size (independent) variable (Kitchenham and Mendes 2004; Lokan and Mendes 2006; Mendes and Lokan 2008). In our experimentations (Section 6) we followed the same practice after analysis of the residuals and validation of the fitting of the regression model.

Furthermore, in the standard linear regression analysis, the variables have to be quantitative but this is not the usual situation when dealing with SCE data. In cases where categorical data should be included in the model, the corresponding variables need to be recoded to sets of dummy variables (binary or (0,1)-variables). The strategy of creating new dummy variables can be a plausible selection for datasets with only a few categorical variables, each with few categories. However, this task becomes extremely complicated for datasets with several categorical variables. The definition of dummy variables increases dramatically the dimensionality of the attribute space, especially when the number of categories is large (c-1 dummy variables are needed to be defined for each categorical variable with c categories) and makes a model impractical. Moreover, the outcome and the interpretation of the model may not be the same as in the case of quantitative continuous explanatory variables, whereas the existence of many dummy variables is known to be a source for multicollinearity (Wissmann et al. 2007).

Finally, as we have mentioned before, the strict linearity assumption between the response cost variable and independent predictors may not hold. There is always the possibility to have in a dataset variables (usually categorical) that are non-linearly related with the dependent variable (see Section 6.2.2). Due to this fact, a significant strong dependency may not be detected by the parametric linear model.

4 A Proposed Semi-Parametric Methodology

4.1 Rationale

LS assumes that there is a functional form between a dependent and a set of independent variables, fully described by a finite set of parameters. However, a pre-selected parametric model for all independent variables is an assumption too strong, especially for datasets with categorical variables. Such a model often fails to fit unexpected effects of the attributes. On the other hand, EbA can offer a flexible procedure in explaining unknown and complicated relationships as it is based on the concept of similarity which can be calculated even for categorical data. However, the inclusion of all independent variables in the procedure of computing similarities is also a strong assumption, which may mask strong parametric relations.

The aforementioned considerations led us to believe that a combination of methods appears to be a more realistic approach for the SCE datasets. These contain a portion of variables (for example the size) that is parametrically correlated with the cost variable and another portion that has a significant impact on cost but in an undefined and non-linear form. An example of such a relationship between complexity and effort is described in Section 6.2.2. So, the research question posed was how to combine these approaches.

The research question led us to the aggregation of the parametric LS regression and the non-parametric EbA in a semi-parametric model called LSEbA. The new model incorporates in a systematic way the linear and non-linear information obtained from the same dataset. It is not our intention to prove that the semi-parametric approach displaces the well-established regression and EbA models. Our goal is to reveal the benefits of utilizing both the abovementioned techniques under certain circumstances, so as to reduce the prediction error.

One of the main research questions of the present study was the building of a prediction model on the original mixed-type SCE datasets, without having to preprocess the categorical attributes. The LSEbA model is an approach that tries to address the various problems of the previous methods, especially those concerning the handling of categorical variables. First, the categorical data do not need preliminary analysis, transformations or definitions of new variables. Second, there is only one model constructed by an algorithm in two steps corresponding to the combined methods EbA and LS regression. Third, we avoid making strong assumptions by presupposing either a parametric model or a non-parametric model for all variables. On the contrary, the LSEbA methodology is free of strict assumptions allowing the linear component of the semi-parametric model to be partly dependent on the non-linear part of it (Robinson 1988).

LSEbA takes advantage of the similarity concept used in EbA, it is easily computed even for cases with missing data and so it does not need to consider methods for deletion or imputation. Of course, we have to make clear that there is no intention to compare LSEbA with the imputation techniques appeared in the literature but to show that the proposed methodology utilizes the initial raw information without having to perform a preliminary analysis.

Next, we present the theoretical background of the method, we describe the algorithm for the construction of the LSEbA model and we illustrate it by a numerical example.

4.2 Description of the Semi-Parametric Model

For the formal presentation of the method, we denote by Y_i the cost of project i and we assume that in our dataset there is a vector $\mathbf{X}_i=(X_{i1},...,X_{ip})$ of independent variables linearly correlated with Y_i and another vector $\mathbf{T}_i=(T_{i1},...,T_{id})$ of variables non-linearly correlated with Y_i. The term “linearly correlated” also implies relations that can be easily transformed to linear by a simple transformation (for example logarithmic or square root) of the variables. On the other hand, non-linear correlations are considered those which are complicated and not directly expressible by an explicit formula.

In general, for the identification of the variables which are linearly correlated with cost it is necessary to conduct a preliminary analysis. Correlation coefficients (e.g. Pearson's coefficient) are useful in detecting linear relationships. The type of the variables may also be a criterion for dividing them in "linear" and "non-linear". In our analysis, for example (Section 6), we considered as "linear" the continuous variables which were found to have strong linear correlation with cost (after transformation) and as "non-linear" the categorical variables.

The model we assume has the following form:

$$Y_i = \mathbf{X}_i\beta + g(\mathbf{T}_i) + \varepsilon_i \tag{9}$$

where $\boldsymbol{\beta}$ is an unknown vector of parameters (linear regression coefficients) and g(**T**) an unknown non-linear function of the vector $\mathbf{T}_i$. The error terms ε_i will be assumed uncorrelated with zero mean and variance σ^2. Robinson (1988) proved that the aforementioned model can be rewritten as

$$Y_i - E(Y_i|\mathbf{T}_i) = (\mathbf{X}_i - E(\mathbf{X}_i|\mathbf{T}_i))\beta + \varepsilon_i \tag{10}$$

indicating that $\boldsymbol{\beta}$ should be evaluated in two steps (Anglin and Gencay 1996):

1. Evaluation of the unknown conditional means, $E(Y_i|\mathbf{T}_i)$ and $E(\mathbf{X}_i|\mathbf{T}_i)$ by a non-parametric method.
2. Substitution of the estimates from step 1 in the place of the unknown functions in Eq. 10 and application of ordinary least squares for the estimation of parameters for vector $\boldsymbol{\beta}$.

4.3 Algorithm Description

Considering the aforementioned form of the model, our approach involves the utilization of EbA (as a non-parametric regression technique) for the evaluation of $E(Y_i|\mathbf{T}_i)$ and $E(\mathbf{X}_i|\mathbf{T}_\mathrm{i})$ in step 1. Then, the parametric LS can be used for the estimation of vector $\boldsymbol{\beta}$.

Our approach can be described in details by the following algorithm:

1. Define which of the independent variables in a dataset are linearly correlated with Y_i. These will form the "LS-set" while the rest of the variables will form the "EbA-set".
2. For every project i, let $\mathbf{X}_i=(X_{i1},...,X_{ip})$ be the LS-set (X-variables) and $\mathbf{T}_i=(T_{i1},...,T_{id})$ the EbA-set (T-variables). The objective is that for a new project, given the combined vector $(\mathbf{X}_{new}, \mathbf{T}_{new})$, we have to estimate the cost Y_{new}.
3. For the new project, apply first the EbA: Find the k–nearest neighbors of $\mathbf{T}_{new}$ among all vectors $\mathbf{T}_i$ using the dissimilarity coefficient described in Section 3.1. Denote the set of neighbors of $\mathbf{T}_{new}$ by J_{new}.
4. For all projects in the data set i=1,...,n compute:

$$\text{a.}\quad \widetilde{Y}_i = Y_i - \frac{1}{k}\sum_{j\in J_i} Y_j \tag{11}$$

$$\text{b.}\quad \widetilde{\mathbf{X}}_i = \mathbf{X}_i - \frac{1}{k}\sum_{j\in J_i} \mathbf{X}_j \tag{12}$$

i.e. estimate by EbA all values of the dependent variable and also all values for each of the independent variables in the LS-set and then subtract these estimations from the original observed values. The set denoted by J_i in Eq. 11 and 12 is the set of all neighbors of vector $\mathbf{T}_i$.

5. Fit the regression model $\widetilde{Y}_i = \widetilde{\mathbf{X}}_i\beta + \varepsilon_i$, i.e. estimate the LS regression coefficients by $\boldsymbol{\beta}_{LS}$.

The final estimation of the dependent variable for the new project will be:

$$\hat{Y}_{new} = \mathbf{X}_{new}\beta_{LS} + \frac{1}{k}\sum_{j\in J_{new}}\left(Y_j - \mathbf{X}_j\beta_{LS}\right) \tag{13}$$

Steps 4 and 5 of the algorithm are essentially used to estimate the terms of Eq. 10. The general idea behind this approach is to remove first the non-linear effect of the T-variables on the dependent and the X-variables and then to fit the linear model. The final estimation of the cost for the new project Y_{new} combines the information from its neighbors according to the T-variables and the linear model fitted according to the X-variables.

4.4 An Example for the Construction of the Semi-Parametric Model

In this section, we present a detailed example illustrating the construction of the proposed semi-parametric model using a small sample of artificial projects. The dataset contains 10 projects with 5 independent variables (Table 1): one continuous (*ind_c1*), two nominal (*ind_n1* and *ind_n2*) and two ordinal (*ind_or1* and *ind_or2*). The dependent variable is *dep.* Based on the training set (p1-p10), we want to estimate by LSEbA the cost of a *new* project, given the vector of independent variables (last row of Table 1).

Step 1 In the first step of the LSEbA algorithm, we have to define which of the independent variables are linearly correlated with Y_i (*dep*). After some preliminary analysis (for example by calculating the Pearson's correlation coefficient), we decide to assign the only independent continuous variable (*ind_c1*) to the LS-set and all the categorical variables (*ind_n1*, *ind_n2*, *ind_or1, ind_or2*) to the EbA-set.

Table 1 Variables of the artificial dataset

Projects	*dep*	*ind_c1*	*ind_n1*	*ind_n2*	*ind_or1*	*ind_or2*
p1	10.9	234	1	1	1	3
p2	11.2	241	1	1	1	3
p3	12.1	262	4	4	3	1
p4	17.4	271	4	4	3	1
p5	19.9	285	2	2	1	3
p6	21.2	288	3	2	1	3
p7	21.4	291	3	3	3	1
P8	21.8	296	3	2	3	1
p9	27.5	301	2	3	2	2
p10	28.7	312	2	3	2	2
new	?	292	3	3	3	1

Step 2 The LS-set and the EbA-set are now expressed as vectors $\mathbf{X}_i=(ind\_c1_i)$ and $\mathbf{T}_i=(ind\_n1_i, ind\_n2_i, ind\_or1_i, ind\_or2_i)$, respectively for each project i. The objective is to estimate the unknown cost value dep_{new} of the *new* project, given the combined vector

$$(ind\_c1_{new}, ind\_n1_{new}, ind\_n2_{new}, ind\_or1_{new}, ind\_or2_{new}) = (292, 3, 3, 3, 1).$$

Step 3 We assume now that we decide to find the *closest nearest neighbor* (or k=1) of $\mathbf{T}_{new}$ among all vectors $\mathbf{T}_i$, using the dissimilarity coefficient defined by Eq. 4. After calculating and sorting the dissimilarities between $\mathbf{T}_{new}$ and each one of the vectors $\mathbf{T}_i$, i=1,...,10, we find that the closest nearest neighbor of the new project is p7 that is the 7th project of the training set. So, according to the notation used, $J_{new}=\{p7\}$.

Step 4 In this step of the algorithm, we estimate by EbA all values in the training set (p1-p10) of the dependent variable (*dep*) and all values of the independent variable (*ind_c1*) in the LS-set by Eq. 11 and 12, respectively. These values form the new variables $\widetilde{Y}_i$ and $\widetilde{\mathbf{X}}_i$ (Table 2). For these estimations, we use again the T-variables and the dissimilarity coefficient in order to find the closest nearest neighbor of each project in the dataset.

Step 5 Based on the new variables $\widetilde{Y}_i$ and $\widetilde{\mathbf{X}}_i$ of Step 4, we fit the regression model $\widetilde{Y}_i = \widetilde{X}_i\beta + \varepsilon_i$ and estimate the LS regression coefficient. Here, we find $\boldsymbol{\beta}_{LS}$= 0.242. The final estimation of the cost for the *new* project (dep_{new}), based on Eq. 13, is:

$$dep_{new} = 292{*}0.242 + (21.4 - 291{*}0.242) = 21.6$$

5 Validation, Accuracy Measures and Graphical Comparison

In this section, we present the basic error measures, statistical methods and graphical tools that we used in the experimentation procedure with the datasets in order to compare LSEbA with LS and EbA separately, and with their combination based on the mean value of their predictions (Kitchenham et al. 2002).

Table 2 Evaluation of the closest neighbor projects and the new variables of Step 4

Project	Closest neighbor	Y_i	$\mathbf{X}_i$	$Y_{neighbor}$	$\mathbf{X}_{neighbor}$	$\widetilde{Y}_i$	$\widetilde{X}_i$
p1	p2	10.9	234	11.2	241	−0.3	−7.0
p2	p1	11.2	241	10.9	234	0.3	7.0
p3	p4	12.1	262	17.4	271	−5.3	−9.0
p4	p3	17.4	271	12.1	262	5.3	9.0
p5	p6	19.9	285	21.2	288	−1.3	−3.0
p6	p5	21.2	288	19.9	285	1.3	3.0
p7	p8	21.4	291	21.8	296	−0.4	−5.0
p8	p7	21.8	296	21.4	291	0.4	5.0
p9	p10	27.5	301	28.7	312	−1.2	−11.0
p10	p9	28.7	312	27.5	301	1.2	11.0

5.1 Prediction Accuracy Measures

The prediction accuracy for each model was evaluated through the *leave-one-out cross-validation* procedure which estimates the cost of each project from all the others. According to this, a completed project with actual cost Y_A is removed from the dataset and the remaining projects are used as a basis for a cost estimation Y_E.

Based on the actual Y_A and the estimated value Y_E, various error functions for the evaluation of the predictive power of the models have been proposed in the SCE literature (Foss et al. 2003). However, there is a lack of convergence of the researchers' opinions regarding the appropriateness of the error measures for the comparison of alternative models (Kitchenham et al. 2001).

Taking into account the discussions in the related literature, we decided to utilize three measures of "local" error (the term "local" refers to the individual error generated by the estimation of a single project) measuring three different aspects of prediction performance (Table 3). Specifically, we calculated for each project i, $i=1,...,n$, which is estimated by the leave-one-out cross-validation procedure:

1. The *absolute error* (AE).
2. The *error ratio* (z).
3. The *magnitude of relative error* (MRE).

These error functions have different interpretation regarding the prediction performance of a model. More precisely, *AE* measures the *accuracy* of the estimation, whereas the error ratio *z* has been proposed by Kitchenham et al. (2001) as a measure of *bias* accounting for underestimation or overestimation of a prediction. Ideally, an optimum estimation is equal to the actual value and then z=1. Values greater or less than 1 show overestimation or underestimation, respectively.

The local error measures of Table 3 are computed for each individual project with the leave-one-out cross-validation procedure and then a "global" statistic is computed from all the error values in order to evaluate the overall prediction performance of a SCE model. These global measures are defined in Table 4.

MRE is the most commonly used measure (Foss et al. 2003; Kitchenham et al. 2001; Korte and Port 2008), however the derived MMRE has been criticized (Foss et al. 2003) for inability to select the "best" model. Moreover, Kitchenham et al. (2001) suggested that MMRE measures the *spread* of the variable z which is clearly related to the distribution of the residuals. Summarizing, the three error functions presented in Table 3 are used in our analysis to measure the following three aspects of a prediction model: (a) the *accuracy,* (b) the *bias* and (c) the *spread.*

Since our purpose is to compare LSEbA with LS, EbA and a simple mixture of them, in the experimentation section (Section 6) we present: (a) all these global error measures for all the models, (b) the percentage of the improvement in each measure achieved by LSEbA and (c) statistical tests for significant differences between the methods.

For the statistical tests, we used only the AEs following the recommended analysis in the study by Kitchenham et al. (2001). Since the distributions of AEs were found to be highly skewed, with many outliers and far from being normal (see also Kitchenham et al. 2001;

Table 3 Local accuracy measures

$AE_i = \lvert Y_{A_i} - Y_{E_i}\rvert$	$z_i = \frac{Y_{Ei}}{Y_{Ai}}$	$MRE_i = \frac{\lvert Y_{Ai} - Y_{Ei}\rvert}{Y_{Ai}}$

Table 4 Global accuracy measures

$MAE = \frac{1}{n}\sum_{i=1}^{n} AE_i$	$MdAE = median\{AE_i\}$
$Meanz = \frac{1}{n}\sum_{i=1}^{n} z_i$	$Medianz = median\{z_i\}$
$MMRE = \frac{1}{n}\sum_{i=1}^{n} MRE_i$	$MdMRE = median\{MRE_i\}$
$pred25 = \frac{\#(\text{projects with } MRE \leq 0.25)}{\#(\text{projects})}$	

Mittas and Angelis 2008b), we applied the non-parametric Wilcoxon sign rank test for paired samples.

The paired-samples Wilcoxon test, unlike the paired t-test, is free of strong assumptions on normality and homogeneity of the underlying populations while its null hypothesis concerns the equality of medians of AEs (i.e. it compares the MdAEs). The paired-samples t-test on the other hand, is completely unreliable in our case, since it compares means under the strong assumptions of normality and homogeneity of variances of both samples. For a discussion on the unsuitability of the t-test in cases where the aforementioned assumptions are violated, we refer to the book of Sheskin (2004). Furthermore, for the choice to compare the models using the AEs and the Wilcoxon test, we refer to the studies by Kitchenham and Mendes (2004); Mendes and Kitchenham (2004); Mendes and Lokan (2008).

In all tests we consider as statistically significant a difference with *p*-value (significance) smaller than 0.05. All the tests conducted are two-tailed (non-directional) in the sense that the alternative hypothesis is that the measures tested are not equal.

5.2 Graphical Comparison

In addition to the accuracy measures and statistical tests described in the previous section, we also used in our analysis a graphical tool for visual comparison of the prediction models, the *Regression Error Characteristic* (REC) curves. REC curves plot simultaneously the *Cumulative Distribution Functions* (CDF) of the prediction errors, obtained by different models, offering a graphical technique to estimate for each model the probability the error to be less or equal than a certain value.

REC curves were introduced by Bi and Bennet (2003) as a generalization of the well-known *Receiver Operating Characteristic* (ROC) curves which are widely used in the visual comparison of classification models. A REC curve is a two-dimensional plot where the horizontal axis (*x*-axis) represents the error tolerance, i.e. all possible values of errors as expressed by a predefined measure, and the vertical axis (*y*-axis) represents the accuracy of a prediction model. Accuracy is defined by Eq. 14 as the percentage of projects that are predicted within the error tolerance *e*. The important feature of REC curves is that they visualize the whole error distribution and not just a single indicator (statistic) of the errors.

$$accuracy\,(e) = \frac{\#(\text{projects with error} \leq e)}{\#(\text{projects})} \tag{14}$$

The interpretation of a plot containing several REC curves, each corresponding to the distribution of errors of a prediction model, is generally simple. If a curve is placed in higher position with respect to all other curves in the plot, then the corresponding prediction model outperforms. The REC curves visualize the overall behavior of prediction errors produced by comparative models and provide useful information regarding their distribution, for example existence of outliers.

In a recent study (Mittas and Angelis 2008c), REC curves were suggested for the visual comparison of SCE models since they can reinforce the decision-making process based on single accuracy indicators or statistical comparisons. Although REC curves are not a method for assessing directly statistically significant differences, they are able to depict the overall dominance of a SCE model against others. An interesting issue of their use in SCE is their ability to represent certain accuracy measures, such as the median of any error measure (e.g. the MdAE) or the pred25. Further details for their properties and the algorithm for their construction can be found in the paper by Mittas and Angelis (2008c).

In the present study, REC curves are used to visualize the relative distributions of certain error measures produced by the comparative models. It is obvious from Eq. 14 that the y-axis of a REC plot ranges from 0 to 1. An entire REC curve begins from $y=0$ and extends until it reaches $y=1$. However, some outliers (extremely high values on the x-axis) can prolong dramatically the shape of a curve until it reaches $y=1$. For reasons of better graphical representation and visual comparison, in the following sections we provide REC curves in two panels: over a focused range of error values in order to show the differences of certain measures like (a) the MdAE and (b) the pred25.

6 Experimentation

In this section, we present the results of the experiments conducted in order to investigate the predictive power of the semi-parametric LSEbA model in comparison to LS and EbA, when applied separately. Furthermore, we compare the estimations of LSEbA with the average of the estimations of the two methods as we described in Section 2. We denote this combination by Mean(EbA, LS) from now on.

For the comparisons, we used three representative datasets that portray different common situations in SCE: (a) a large dataset containing both continuous variables and few factors (the term “factors” is used for independent categorical variables) with no missing values, (b) a relatively smaller dataset with continuous variables and several factors without missing values and (c) a dataset with continuous variables and factors with missing values.

The methodology followed for the construction of the LSEbA model is extensively presented in Section 4.3. Regarding the first step of the algorithm, we have to define in each dataset a partition of the independent variables into an LS-subset, linearly correlated with the cost, and an EbA-subset which will be used for finding the nearest neighbors in steps 3 and 4.

After some preliminary analysis of the relations with the cost variable, we decided to assign all continuous variables to the LS-set and all factors to the EbA-set. This decision was made because the continuous independent variables, after logarithmic transformations, were found to be linearly correlated with the cost variables. More precisely, we evaluated the two-tailed Pearson’s correlation test, whereas a relationship was confirmed as statistically significant when the p-value of the test was smaller than 0.05. As for the factors, we took into account all the difficulties and limitations discussed in Section 3.2 regarding the manipulation of categorical variables in a parametric model. In the following sections, we describe in detail these difficulties in each one of the datasets used for experimentation. It is interesting to note that analogous partitions of datasets in continuous and categorical variables were also used in other analyses (Briand et al. 2000; Kitchenham 1998).

As far as the EbA model concerns, the statistic for the evaluation of the dependent variable through neighbor projects was the arithmetic mean (Eq. 2 and 3). An important issue that had to be addressed was the number of analogies that should be used for the estimation. In order to select a sort of optimal number of analogies, we had to minimize a criterion (Myrtveit et al. 2005), so we decided to use as criterion the MdAE. More specifically, we evaluate the MdAE for a range of possible nearest neighbors (k=1 to k=20 in our example) through the leave-one-out cross-validation procedure assuming that the number that minimizes MdAE gives the best model. The same procedure was used for the first step of the LSEbA method.

6.1 ISBSG Dataset 1

6.1.1 Description of the Dataset

The first dataset is derived from the International Software Benchmarking Standards Group (ISBSG, release 10) (ISBSG 2007) which is a well-known, multi-organizational and international repository of completed software projects. As the ISBSG guidelines suggest, we have to select a suitable subset of projects in order to make our analysis meaningful and so, we chose to work with the proposed dataset that is also utilized by the ISBSG application “Early Estimate Checker V5.0” (ISBSG 2007). This tool fits regression models to ISBSG data with dependent one of the variables *effort*, elapsed time or *duration*, *project delivery rate* and *speed of delivery*. As the last two are computed by ratios of the size, we decided to work with a plain dependent variable in order to include the size in the model. We chose to work with *duration* because the *effort* gave similar results for all methods and generally poor models. Hence, the experimentation dataset contains 1,530 completed projects with 4 independent predictors (3 nominal and 1 continuous), whereas the dependent variable is *Duration* (Table 5).

In a recent paper (Liebchen and Shepperd 2008), the researchers discuss the problem of the quality of datasets in Software Engineering reporting their concerns about the quality of some of the datasets that are used to learn, evaluate and compare prediction models. Taking into account the importance of data quality, we removed from the dataset the projects which are of low quality (rated with C and D by ISBSG) and from the remaining (rated with A and B) we used only those projects measured by the same sizing method. More specifically, the database of ISBSG has two fields, *Data Quality Rating* and *UFP Rating,* which contain ratings regarding the quality of measurements for every project. The recommendation of ISBSG is to avoid using in statistical analyses projects rated with C and D. On the other hand, ISBSG underlines that it is necessary to work with projects with the same sizing method. So, from the field *FP Standards*, we selected all variants of IFPUG 4 or NESMA which can be utilized

Table 5 Variables of the ISBSG dataset 1

Variable	Scale	Description	Levels definition
Duration	Ratio	Total elapsed time for the project in calendar months	
Ufp	Ratio	Application size in unadjusted function points	
DevType	Categorical	Development Type	Enhancement, New Development, Re-development
DevPlat	Categorical	Development Platform	MF, MR, Multi, PC
LangType	Categorical	Language Type used for the project	2GL, 3GL, 4GL, ApG

together in statistical analysis according to ISBSG. After the deletion of cases with missing values, the dataset was restricted to 759 projects.

6.1.2 The Parametric Model

The parametric LS model was fitted after preliminary analysis of the dataset. The two ratio-scaled variables (*Duration* and *Ufp*) were transformed to the natural logarithmic scale in order to achieve better fitting. On the other hand, the dataset also contains three nominal factors (Table 5) that had to be replaced by binary (or dummy) variables in order to be included in the LS model. Due to the fact that the size of dataset is large and that there are only three factors, the utilization of dummy variables seems quite reasonable. The final set of variables, used for the construction of the LS model, is presented in Table 6. As we have already mentioned, for a nominal variable with *c* possible values, we always need *c*-1 dummy variables. So, if for example a nominal variable has *c*=3 categories, these can be represented by the ordered pairs (0,1), (1,0) and (0,0). These pairs require only 2 binary variables and therefore the nominal variable can be replaced by two dummy variables.

Despite the small number of factors, there seems to exist a problem of multicollinearity, since the resulting from the stepwise procedure model in Eq. 15 excludes all of the dummy variables except *LangType1* and *LangType2*. After running the stepwise procedure, three independent variables that explained 29.9% of *lnduration* (r^2) were identified as statistically significant and were used in order to build the LS model:

$$\begin{aligned}\ln duration = & -0.202 + 0.348 \times 1\mathrm{n}\, ufp \\ & + 0.907 \times LangType1 + 0.356 \times LangType2\end{aligned} \tag{15}$$

6.1.3 The Non-Parametric Model

As the non-parametric model is a free of assumptions method, we used all of the original variables presented in Table 5 for estimating the model.

$$duration = f(Ufp, DevType, Devplat, LangType) \tag{16}$$

The abstract form of Eq. 16 shows that the unknown function $f(\mathbf{X})$ was estimated using all independent variables (continuous and factors) which participated in the computation of

Table 6 Variables used in the stepwise regression

Variable	Meaning
lnduration	Natural logarithm of *Duration*
lnufp	Natural logarithm of *Ufp*
DevType1	Dummy variable where “Enhancement” type is coded as 1 and all other types as 0
DevType2	Dummy variable where “New Development” type is coded as 1 and all other types as 0
DevPlat1	Dummy variable where “MF” platform is coded as 1 and all other platforms as 0
DevPlat2	Dummy variable where “MR” platform is coded as 1 and all other platforms as 0
DevPlat3	Dummy variable where “Multi” platform is coded as 1 and all other platforms as 0
LangType1	Dummy variable where “2GL” language is coded as 1 and all other languages as 0
LangType2	Dummy variable where “3GL” language is coded as 1 and all other languages as 0
LangType3	Dummy variable where “4GL” language is coded as 1 and all other languages as 0

the dissimilarity coefficient (Section 3.1) and the finding of the analogue projects. Following the procedure for finding the optimal number of analogies, we observed that seven analogies gave the best MdAE value and therefore $k=7$ was used in Eq. 2 and 3.

6.1.4 The Semi-Parametric Model

In the specification of the LSEbA model, a similar approach to the one presented for the case of EbA was adopted for the selection of the analogue projects. The number of neighbors that minimized MdAE was sixteen ($k=16$). As we can observe from Eq. 17, the form of LSEbA model depicts that the only continuous variable was inserted in the linear part, whereas all factors were used for the evaluation of the unknown non-linear function g ($\mathbf{T}$) through the steps of the algorithm proposed in Section 4.3.

$$\ln duration = 0.332 \times \ln ufp + g(DevType, Devplat, LangType) \tag{17}$$

6.1.5 General Results

The overall prediction performance for each of the four comparative models is presented in Table 7. We can see that the LSEbA model outperforms both LS and EbA in terms of the measures of accuracy MAE and MdAE (see Section 5.1). The comparison of LSEbA with Mean(EbA,LS) shows that there is no difference in terms of MAE but this does not hold for the case of MdAE, where LSEbA appears smaller value of median AEs than Mean(EbA,LS). Concerning the bias of the predictions through the evaluation of error ratio z that inspects the balance between overestimation and underestimation, we can clearly observe that the mean values of the error z-ratio indicate that all methods are prone to overestimation since all values are higher than the optimum value of 1 (Table 7). Due to the fact that the mean statistic is affected by the presence of extreme outliers and the data is highly skewed, we should also examine the median as a more robust measure of central tendency for their distributions. Using the medians, it is clear that LSEbA and LS have similar median values which are closer to 1 compared to that of EbA and Mean(EbA,LS) and therefore they present lower bias. Finally, LSEbA appears smaller spread indicators (MMRE and MdMRE) than LS, EbA and Mean(EbA,LS) (Table 7).

Table 7 Accuracy measures for ISBSG dataset 1

	LS	EbA	Mean(EbA,LS)	LSEbA
MAE	4.03 (2.23%)	4.19 (5.97%)	3.94 no improvement	3.94
MdAE	2.78 (8.99%)	3.00 (15.67%)	2.84 (10.92%)	2.53
MMRE (%)	62.27 (4.93%)	76.44 (22.55%)	67.17 (11.87%)	59.20
MdMRE (%)	40.06 (4.72%)	42.86 (10.94%)	39.21 (2.65%)	38.17
pred25 (%)	30.57 (16.78%)	32.54 (9.71%)	31.88 (11.98%)	35.70
Meanz	1.26 (15.38%[a])	1.49 (55.10%[a])	1.37 (40.54%[a])	1.22
Medianz	0.95 no improvement	1.13 (53.85%[a])	1.08 (25.00%[a])	0.94

[a] The % improvement for the case of error ratio z is evaluated on the basis of its distance from the optimum value of 1 (i.e. improvement means that a mean or median value is getting closer to 1)

In each cell of Table 7, we can also see the percentage of the improvement in all measures of LSEbA, compared to that of LS, EbA and Mean(EbA,LS). The improvement achieved by LSEbA compared to LS, ranges from 2.23% (MAE) up to 16.78% (pred25), whereas the comparisons between LSEbA and EbA show an improvement that varies from 5.97% (MAE) up to 55.10% (Meanz). Finally, our approach seems also to present better accuracy indicators than Mean(EbA,LS) with an improvement that varies from 2.65% (MdMRE) up to 40.54% (Meanz).

In order to examine the significance of the accuracy improvement obtained by LSEbA, we used the Wilcoxon signed rank test for matched pairs which tests the values of MdAE for significant differences. The tests show (Table 8) that there is a statistically significant difference between the distributions of AEs obtained by LSEbA and LS and also by LSEbA and EbA. The p-value for the case of LSEbA and Mean(EbA,LS) is higher than 0.05 signifying no difference between the comparative models. Regarding the comparison between LS and EbA, we can infer that there is no significant difference, since the p-value of the test is higher than 0.05.

In the two panels of Fig. 1, we can see the REC curves for AE and MRE for the four models. The left panel focuses on the AEs corresponding to the lower 50% of the accuracy, whereas the right panel focuses on the region corresponding to values of MRE (x-axis) less than 0.30.

In Fig. 1(a) the median value of AEs (MdAE) for each prediction model can be evaluated by drawing first a horizontal line from 0.5 of the accuracy (y) axis towards the REC curves and then by projecting on the x-axis the points of intersection with the REC curves. Therefore, the points m_1, m_2 m_3 and m_4 visualize the position of MdAE for the four models. Clearly, the MdAE of LSEbA (m_1) has the lowest value compared with the median values of LS, EbA and Mean(EbA,LS) (m_2, m_3 and m_4, respectively).

In Fig. 1(b), the values of pred25 for the four models are visualized by drawing first a reference vertical line from 0.25 of the x-axis and then from the intersecting point of the REC curve, a horizontal line which meets the accuracy axis. From the relative positions of the four pred25 values (p_1, p_2, p_3 and p_4), we can infer that LSEbA achieves the highest (and hence the best) pred25 measure (p_1), whereas the pred25 of EbA (p_3) is better than LS (p_2) and Mean(EbA,LS) (p_4).

The results derived from the experimentation on a large dataset with a continuous independent variable and a small number of factors with no missing values suggest that LSEbA improves significantly the accuracy of the predictions. Almost all of the global error measures are improved, the graphical inspection of the distributions shows a clear superiority of LSEbA and the statistical tests imply that the improvement in the accuracy, compared to both LS and EbA, is statistically significant. Finally, LSEbA model presents better accuracy indicators than the simple mixture of them but the statistical comparison does not signify the difference for the case of MdAE measure.

Table 8 Significance of Wilcoxon tests for ISBSG dataset 1 (MdAE)

Comparison	p-value
LSEbA-LS	**0.043/Sig.**
LSEbA-EbA	**0.000/Sig.**
LSEbA-Mean(EbA,LS)	0.425
EbA-LS	0.080

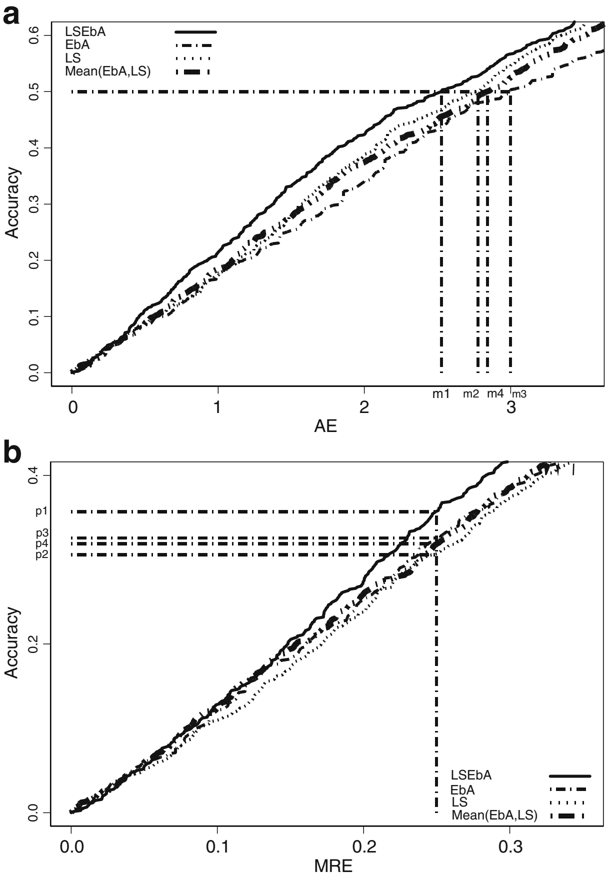

Fig. 1 REC curves over a limited range for **a** AE to visualize differences between the MdAE values (m_1, m_2, m_3 and m_4 correspond to LSEbA, LS, EbA and Mean(EbA,LS)) and **b** MRE to visualize differences between the pred25 values (p_1, p_2, p_3 and p_4 correspond to LSEbA, LS, EbA and Mean(EbA,LS))

6.2 NASA93 Dataset

6.2.1 Description of the Dataset

The second dataset, used in our experimentation, is NASA93 (NASA93 2007). This dataset was selected because it is publicly available and because it has interesting structure which is challenging for the testing of the proposed semi-parametric method. More precisely, the dataset is relatively small and contains both continuous variables and several factors with no missing values.

The initial dataset contains 93 NASA projects from different centers with 23 independent predictors (22 categorical and 1 continuous) while the dependent variable is the actual effort measured in person months. After the removal of five independent variables that are not related to effort (i.e. unique id, project name etc), we finally have a dataset with variables presented in Table 9. We have to clarify that as in the study by Kitchenham (1998), it is not our purpose to gain new insights into the NASA93 dataset but

Table 9 Variables of the NASA93 dataset

Variable	Scale	Description	Levels definition
Effort	Ratio	Development effort in months	
SLOC	Ratio	Equivalent physical 1000 lines of source code	
forg	Categorical	Flight or ground system	Flight, Ground
mode	Categorical	Development mode	Embedded, Organic, Semidetached
acap	Categorical	Analysts capability	Very low Low Nominal High Very high Extra high
pcap	Categorical	Programmers capability	
aexp	Categorical	Application experience	
modp	Categorical	Modern programming practices	
tool	Categorical	Use of software tools	
vexp	Categorical	Virtual machine experience	
lexp	Categorical	Language experience	
sced	Categorical	Schedule constraint	
stor	Categorical	Main memory constraint	
data	Categorical	Database size	
time	Categorical	Time constraint for cpu	
turn	Categorical	Turnaround time	
virt	Categorical	Machine volatility	
cplx	Categorical	Process complexity	
rely	Categorical	Required software reliability	

to illustrate the merits of the proposed methodology on a dataset with features common in SCE.

6.2.2 The Parametric Model

The same procedure as in Section 6.1.2 was followed in order to fit the parametric LS model. Again, the initial variables were logarithmically transformed in order to achieve better fitting. On the other hand, this dataset contains a large amount of categorical variables. More specifically, there are 17 factors, most of them with 6 different levels (Table 9) that have to be replaced by dummy variables. Having in mind that it is required to define 15×5=75 new dummy variables for the last 15 factors with 6 levels, 1 dummy variable for *forg* and 2 new dummy variables for *mode* (overall 78 new dummy variables); it is clear that the strategy of creating so many binary variables is impracticable for this specific dataset due to the small size of the sample of projects.

Another plausible choice would be to attempt to concatenate the categories of the predictors into homogeneous groups for each factor. For each one of these categorical variables, we can perform one-way ANOVA in order to test the impact of every factor on the original dependent variable. Every factor with significance less than 0.05 can be considered important and has to be included in the analysis, whereas post-hoc tests can be performed in order to identify the various homogeneous categories that have to be concatenated in every factor. Detailed description of the whole procedure can be found in the study by Sentas et al. (2005). On the other hand, this strategy is also not so simple and important design decisions have to be made in order to carry out a meaningful analysis. The following example is quite representative about the pitfalls of the procedure and portrays the difficulties in the construction of a reliable and flexible prediction model.

Let us assume that our goal is to concatenate the levels of the factor *cplx*. The significance value of the *F*-test in the ANOVA table is less than 0.0005 and indicates that there is a statistically significant relationship between *lneffort* and *cplx*. In Fig. 2, we can see a complicated non-linear relationship from which it is clear that the mean of *lneffort* initially decreases for low, nominal and high levels and then increases for very high and extra high levels. This type of relationship does not allow merging levels so as to preserve the ordinal nature of the factor and to interpret its effect. The shape of Fig. 2 suggests that the effect of the factor can be detected more easily by the non-parametric EbA method.

The problem with the manipulation of the categorical variables in this specific dataset has also been studied by Korte and Port (2008) and by Port and Korte (2008). Both studies investigate the construction of five different models for the NASA93 dataset with a variety of regression models trying to handle the mixed nature of independent variables. The authors pointed out that a specific form of regression model (categorical least squares regression) allowing the parameters to vary unconstrained, was fitted to the data very well and gave low percentages of errors. However, this model estimated the parameter values for *cplx* to be negative for level “low”, positive for the levels “nominal and high” and again negative for the level “very high”, not taking into account the ordered nature of the variable. The problem is that such quantification is contrary to the general intuitive belief that higher complexity (*cplx*) requires higher *effort*. Moreover, Port and Korte (2008) noted that although someone can be confident that by adding effort multipliers (essentially by incorporating categorical variables in the model) the accuracy of a model is improved, this is not true for the NASA93 dataset. So, it is suggested that a linear model, with *lneffort* as the dependent and *lnSLOC* as the independent variable, gives results that do not differ significantly from the results obtained by a model constructed with the utilization of other categorical variables.

Having in mind the abovementioned discussion, we finally built a simple linear parametric model with *lneffort* as the dependent variable and *lnSLOC* as the independent variable which explains a quite high percentage of the variability (r^2=71.1%). This model

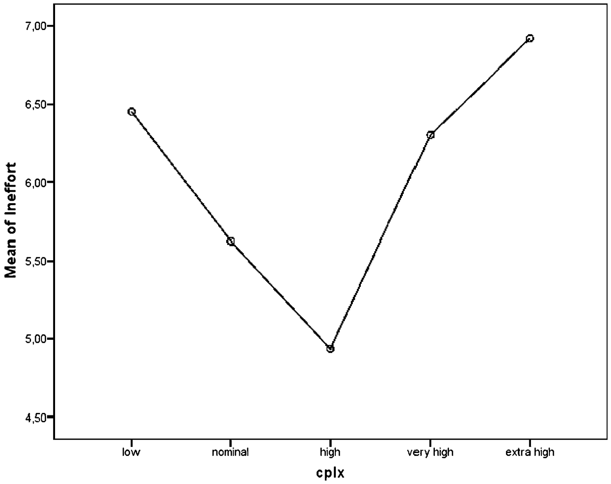

Fig. 2 Means plot for *lneffort* and *cplx* factor

has quite similar performance in terms of accuracy with the models built by Korte and Port (2008). However, we have to point out that our accuracy indicators in Table 10 were obtained by the leave-one-out cross-validation procedure and concern the predictive accuracy of our model in Eq. 18 whereas the results by Korte and Port (2008) were obtained by a bootstrapping procedure using the whole data set without cross-validation.

$$\ln \mathit{effort} = 1.977 + 0.920 \times \ln \mathit{SLOC} \tag{18}$$

6.2.3 The Non-Parametric Model

The same procedure as in Section 6.1.3 was followed for the calibration of EbA. All of the initial variables (Table 9) participated in the computation of the dissimilarity coefficient and the finding of the analogies in the procedure of estimating the non-parametric model.

$$\begin{aligned}\mathit{effort} = {} & f(\mathit{SLOC}, \mathit{forg}, \mathrm{mod}e, \mathit{acap}, \mathit{pcap}, a\exp, \mathrm{mod}\, p, \\ & \mathit{tool}, v\exp, l\exp, \mathit{sced}, \mathit{stor}, \mathit{data}, \mathit{time}, \mathit{turn}, \mathit{virt}, \mathit{cplx}, \mathit{rely})\end{aligned} \tag{19}$$

The number of analogue projects that minimized MdAE was found to be one (k=1), so in this case we used only the closest neighbor.

6.2.4 The Semi-Parametric Model

The specification of the LSEbA model was also conducted through the procedure described in Section 6.1.4. More precisely, we inserted *lnSLOC* in the linear part and all the categorical variables in the non-linear part of Eq. 20. Seven analogies (nearest neighbors) were found to be the best choice for applying the EbA part of the algorithm.

$$\begin{aligned}\ln \mathit{effort} = {} & 0.995 \times \ln \mathit{SLOC} + g(\mathit{forg}, \mathrm{mod}e, \mathit{acap}, \mathit{pcap}, a\exp, \mathrm{mod}\, p, \\ & \mathit{tool}, v\exp, l\exp, \mathit{sced}, \mathit{stor}, \mathit{data}, \mathit{time}, \mathit{turn}, \mathit{virt}, \mathit{cplx}, \mathit{rely})\end{aligned} \tag{20}$$

6.2.5 General Results

The performance of each one of the comparative models is shown in the summary results of Table 10. The LSEbA model seems to have the best performance in terms of all error measures. The improvement achieved by LSEbA compared to LS, ranges from 19.82%

Table 10 Accuracy measures for NASA93

	LS	EbA	Mean(EbA,LS)	LSEbA
MAE	353.23 (36.90%)	445.54 (49.98)	320.25 (30.40%)	222.88
MdAE	73.48 (40.54%)	142.00 (69.23%)	96.08 (54.53%)	43.69
MMRE (%)	67.88 (19.82%)	86.93 (37.47%)	67.94 (19.99%)	54.36
MdMRE (%)	36.78 (29.36%)	42.86 (39.38%)	38.13 (31.86%)	25.98
pred25 (%)	32.26 (46.65%)	35.48 (33.34%)	33.33 (41.94%)	47.31
Meanz	1.34 (35.29%[a])	1.39 (43.59%[a])	1.36 (38.88%[a])	1.22
Medianz	1.07 (57.14%[a])	0.94 (50.00%[a])	1.10 (70.00%[a])	0.97

[a] The % improvement for the case of error ratio z is evaluated on the basis of its distance from the optimum value of 1 (i.e. improvement means that a mean or median value is getting closer to 1)

(MMRE) up to 57.14% (Meanz), from 33.34% (pred25) up to 69.23% (MdAE) compared to EbA and from 19.99% (MMRE) up to 70.00% (Medianz) for the comparison with the Mean(EbA,LS) model. The statistical tests in Table 11 show significant differences between all pair-wise comparisons of AEs obtained by the four models.

Figure 3(a) depicts the large difference between the MdAE of LSEbA (m_1) and of the other three methods (m_2, m_3 and m_4 for LS, EbA and Mean(EbA,LS), respectively). Furthermore, LS model seems to be the second “best” choice for building a prediction model since it significantly outperforms EbA.

The mean values of error z-ratios (Table 10) indicate that all models are prone to overestimations. However, the median values of all methods are quite close to 1 with LSEbA having the “least” bias. The mean and median statistics of MRE also bring to light certain improvement of LSEbA model in terms of spread. Finally, in Fig. 3(b) is clearly portrayed the best performance of LSEbA in terms of pred25 accuracy measure.

Summarizing the findings of the experimentation on the second dataset, we examined a specific case study in order to better present the potential utilization of the LSEbA procedure. The dataset contained both continuous variables and a large number of categorical predictors which caused many design problems in the evaluation of the parametric linear model and poor prediction performance of the non-parametric model. Since the dataset was not particularly large, the usual strategy of creating dummy variables proved to be inappropriate. Moreover, the existence of many levels in each factor and the non-linear relationship between them and the means of *lneffort* rendered the one-way ANOVA procedure an extremely difficult task.

The evaluation of the semi-parametric model showed that the new proposed methodology achieved to combine in an easily adaptable way both the parametric and the non-parametric components. The results obtained from the improvement in all of the error measures, from the statistical tests and from the graphical inspection of the REC curves provide evidence for the superiority of LSEbA in this specific dataset.

6.3 ISBSG Dataset 2

6.3.1 Description of the Dataset

The last dataset used in our experimentation was also derived from the ISBSG repository. This dataset contains different projects from the ones used in the first experiment, presented in Section 6.1.1. More specifically, our scope was to utilize a dataset which is relatively small with both continuous variables and several factors with missing values.

Following again the recommendations of ISBSG, proposing the utilization of subsets in which the same sizing method is used, we initially selected to work with all projects with COSMIC-FFP as *FP Standards* sizing method. Secondly, we ignored the projects with C and D from the variables *Data Quality Rating* and *UFP Rating* and worked with the

Table 11 Significance of Wilcoxon tests for NASA93 (MdAE)

Comparison	p-value
LSEbA-LS	0.001/Sig.
LSEbA-EbA	0.000/Sig.
LSEbA-Mean(EbA,LS)	0.000/Sig.
EbA-LS	0.047/Sig.

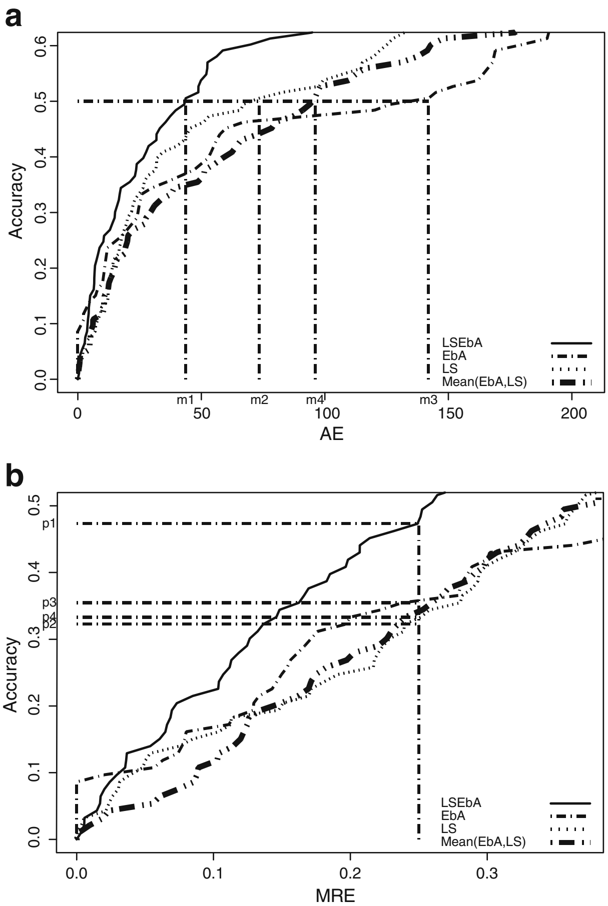

Fig. 3 REC curves over a limited range for **a** AE to visualize differences between the MdAE values (m_1, m_2, m_3 and m_4 correspond to LSEbA, LS, EbA and Mean(EbA,LS)) and **b** MRE to visualize the differences between the pred25 values (p_1, p_2, p_3 and p_4 correspond to LSEbA, LS, EbA and Mean(EbA,LS))

remaining A and B ratings. Another important issue that had to be addressed was the variables that should be included in our analysis. The ISBSG also points out to utilize the most important criteria for selecting projects such as *Size, Development Type, Primary Programming Language, Development Platform* and other general *Development techniques*. After a careful examination of all variables, we chose to include both subsets of variables under the names "*Grouping Attributes*" and "*Project Attributes*" and build the comparative models on a final dataset of 110 projects (Table 12). The names "*Grouping Attributes*" and "*Project Attributes*" were given by ISBSG (2007) in order to describe groups of variables with similar characteristics. For example, the first group includes variables such as *DevType*, *OrgType*, *BusAreaType*, *AppType* etc, whereas the second group includes variables such as *DevPlat*, *LangType*, *PrimLangType*, *FLang* etc.

Most of the categorical variables contain a large number of different levels for each factor (indicatively *PrimLangType* has 21 different levels). The factors contain in general several missing values. In the last column of Table 12 the number of missing values for each variable is also given.

Table 12 Variables of the ISBSG dataset 2

Variable	Scale	Description	#missing values
Effort	Ratio	Total effort in hours recorded	0
Ufp	Ratio	Application size in unadjusted function points	0
MaxTeamSize	Ratio	The maximum number of people that worked at any time on the project	50
DevType	Categorical	Development Type	0
OrgType	Categorical	Organization Type	10
BusAreaType	Categorical	Business Area Type	81
AppType	Categorical	Application Type	6
Architecture	Categorical	Architecture	25
ClientSer	Categorical	Client Server	33
DevPlat	Categorical	Development Platform	7
LangType	Categorical	Language Type	19
PrimLangType	Categorical	Primary Language Type	15
FLang	Categorical	First Language Type	16
FDataBaseSys	Categorical	First Data Base System	36
FCompSer	Categorical	First Component Server	89
CaseTool	Categorical	Case Tool	61

6.3.2 The Parametric Model

The application of the parametric LS model is an extremely complicated task for this specific dataset. First, the large number of different levels in each factor makes the definition of new dummy variables impossible and meaningless. Moreover, the one-way ANOVA for merging categories can not be performed due to fact that there are levels corresponding to only one observation. In addition, the large amount of missing values can not be manipulated by the specific model since LS performs a list-wise deletion of a case with at least one missing value. If the deletion of missing values is accomplished, the size of dataset becomes extremely small.

As the size of projects is the most important cost driver and there is no missing value for this independent variable (*Ufp*), we finally built a linear parametric model with the logarithmic transformation of the response variable (*lneffort*) and the *Ufp* independent variable (*lnufp)* which explains 37.3% of the variability (r^2=37.3%).

$$\ln effort = 1.753 + 1.079 \times \ln ufp \tag{21}$$

6.3.3 The Non-Parametric Model

All the initial variables (Table 12) participated in the computation of the dissimilarity coefficient and the finding of the analogies in the procedure of estimating the EbA non-parametric model

$$\begin{aligned} effort = f(&Ufp, MaxTeamSize, DevType, OrgType, BusAreaType,\\ &AppType, Architecture, ClientSer, DevPlat, LangType,\\ &\Pr imLangType, FLang, FDataBaseSys, FCompSer, CaseTool) \end{aligned} \tag{22}$$

The number of analogue projects that minimized MdAE was found to be one (k=1), so in this case we used only the closest neighbor.

6.3.4 The Semi-Parametric Model

The problems discussed in Section 6.3.1, concerning the large amount of levels for each factor and the missing values, are addressed by inserting all the categorical variables and *MaxTeamSize* (which has also several missing values), in the EbA-set. The main continuous independent variable (*lnufp*) is assigned to the LS-set. After the evaluation of leave-one-out cross-validation procedure, two analogies were found to be the best choice through the calibration procedure.

$$\begin{aligned}\ln \mathit{effort} = \;& 0.934 \times \ln \mathit{ufp} + g(\mathit{MaxTeamSize}, \mathit{DevType}, \mathit{OrgType}, \\ & \mathit{BusAreaType}, \mathit{AppType}, \mathit{Architecture}, \mathit{ClientSer}, \mathit{DevPlat}, \mathit{LangType}, \\ & \Pr \mathit{imLangType}, \mathit{FLang}, \mathit{FDataBaseSys}, \mathit{FCompSer}, \mathit{CaseTool})\end{aligned} \tag{23}$$

6.3.5 General Results

Table 13 summarizes the overall prediction performance for each one of the four comparative models. The improvement achieved by LSEbA in all measures is generally high.

The accuracy of LSEbA in terms of AE is the "best" compared to LS, EbA and Mean (EbA,LS) whereas the statistical tests (Table 14) show that the differences between the medians are statistically significant. Considering the comparison between LS and EbA, the AE summary measures suggest that EbA is the second "best" choice in terms of accuracy but the Wilcoxon test (fourth row in Table 14) does not support a statistically significant difference. The differences among the values of MdAE of the four models are shown in Fig. 4(a), whereas the examination of the pred25 accuracy indicators (Table 13 and Fig. 4(b)) shows the best prediction performance of LSEbA in terms of this specific measure.

The mean and the median error ratio z of LSEbA appear to have the closest to 1 values, indicating generally unbiased predictions (Table 13). Finally, the measures computed from MRE are generally very high for all models and the large divergence between means and medians (Table 13) reveals the existence of outliers.

In summary, we examined the behavior of the proposed methodology in a very difficult situation, which is common in SCE. More precisely, the situation involves a hard-to-analyze and relatively small dataset with a large number of categorical variables, each with

Table 13 Accuracy measures for ISBSG dataset 2

	LS	EbA	Mean(EbA,LS)	LSEbA
MAE	4789.76 (11.47%)	6133.57 (30.86%)	4571.84 (7.25%)	4240.61
MdAE	1637.71 (47.02%)	1183.00 (26.66%)	1276.43 (32.02%)	867.66
MMRE (%)	324.25 (45.17%)	520.95 (65.87%)	399.58 (55.51%)	177.79
MdMRE (%)	86.76 (33.17%)	70.25 (17.47%)	81.83 (29.15%)	57.98
pred25 (%)	4.55 (399.56%)	16.36 (38.94%)	16.36 (38.94%)	22.73
Meanz	3.53 (49.41%[a])	5.61 (72.23%[a])	4.57 (64.15%[a])	2.28
Medianz	0.83 (76.47%[a])	1.10 (60.00%[a])	1.34 (88.24%[a])	1.04

[a] The % improvement for the case of error ratio z is evaluated on the basis of its distance from the optimum value of 1 (i.e. improvement means that a mean or median value is getting closer to 1)

Table 14 Significance of Wilcoxon tests for ISBSG dataset 2 (MdAE)

Comparison	p-value
LSEbA-LS	0.000/Sig.
LSEbA-EbA	0.000/Sig.
LSEbA-Mean(EbA,LS)	0.034/Sig.
EbA-LS	0.181

several levels and missing values. The LSEbA model achieved to deal with these aforementioned issues and to evaluate the cost function in a straightforward manner using only the initial variables, without any preprocessing procedure. The findings showed that the semi-parametric model improved significantly the performance in comparison to LS, EbA and Mean(EbA,LS). The results were verified by both statistical tests and the graphical representation of errors.

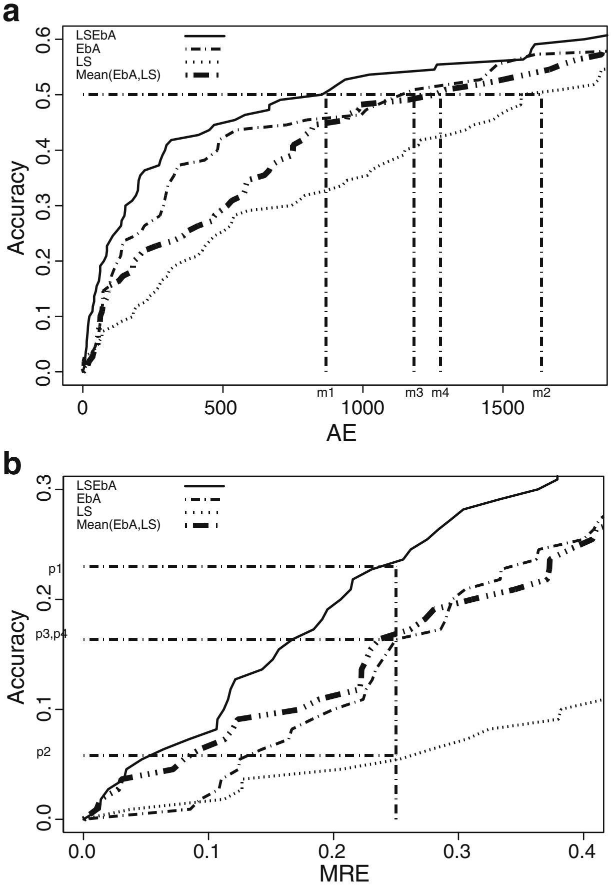

Fig. 4 REC curves over a limited range for **a** AE to visualize the differences between the MdAE values (m_1, m_2, m_3 and m_4 correspond to LSEbA, LS, EbA and Mean(EbA,LS)) and **b** MRE to visualize the differences between the pred25 values (p_1, p_2, p_3 and p_4 correspond to LSEbA, LS, EbA and Mean(EbA,LS))

7 Conclusions and Discussion

In this paper, the problem of constructing a model for predicting the cost of a software project was considered by introducing a new methodology. The method has theoretical foundations on the formal statistical model known as Partially Linear Model and it achieves to incorporate both parametric and non-parametric relationships into a simple semi-parametric model. In this regard, the proposed methodology, called LSEbA, is essentially a hybrid technique which combines two of the most known methodologies in software cost estimation, namely least squares regression analysis and estimation by analogy.

The method described in this paper is intended to be useful for modeling complicated datasets, a situation quite common in software cost estimation. More specifically, the proposed methodology is a particularly plausible choice if the dataset contains mixed-type data with continuous and categorical attributes which can be used to explain the variability of the cost variable through parametric and non-parametric relationships. As the use of categorical variables in regression analysis is generally problematic inducing many issues to be addressed, and as the estimation by analogy fails to capture parametric relationships causing precision problems, the introduction of a semi-parametric model addresses several problems. The method is simple, accomplished in two steps and it does not require any complicated statistical preprocessing of the data. Furthermore, it is capable to handle unbalanced factors and missing values.

Trial application on three representative case studies showed that certain improvements can be obtained by the proposed model, compared to both the well-established parametric linear regression model and the non-parametric estimation by analogy. Furthermore, LSEbA gives better results in comparison to the mixed model obtained by the average of LS and EbA. The prediction performance of each of the comparative models was closely examined considering three different expressions of error; the absolute error, the error ratio and the magnitude of relative error. The findings that revealed the superiority of the semi-parametric model were verified by both robust statistical tests conducted on the absolute errors and graphical comparisons through the utilization of Regression Error Characteristic curves.

Our scope was not to prove that LSEbA can substitute regression analysis and estimation by analogy. After all, the proposed method can be seen as a generalization of these two. Indeed, since the method requires a splitting of the independent variables into two sets, the EbA-set and the LS-set, we can see estimation by analogy and regression as trivial cases of LSEbA where the LS and the EbA sets (respectively) are empty. For this reason, there is no point to see LSEbA as competitive to the other two methods but rather as a decision procedure on how to partition the predictors' set into subsets containing parametric and non-parametric information. If all independent variables are continuous and there is some kind of prior knowledge or indications about the explicit functional form, all variables are assigned to the LS-set while the EbA set remains empty and then the parametric (usually linear) model seems to be the first choice. On the other hand, when there are only categorical variables in a dataset or when there is no evidence for some parametric relation, the LS-set is empty, all variables are assigned to the EbA-set and non-parametric estimation by analogy should be selected in order to provide predictions based on the most similar projects. However, in practice we know that the cost function is usually linearly depended on a set of predictors (after some functional transformation), whereas there is also a portion of the variability that cannot be explained parametrically, so the combination of the well-established models into a new semi-parametric can resolve the abovementioned problem.

An interesting result of our experimentation with the three representative datasets is that LSEbA is less or equal sensitive to the choice of parameter k (nearest neighbors) for the evaluation of the prediction function compared to the case of EbA methodology. More precisely, we calculated the coefficient of variation (CV=standard deviation/mean) of MdAEs for the range of nearest neighbors (k=1 to k=20) for both EbA and LSEbA models. Generally, the CV shows lower dispersions of MdAEs in LSEbA with respect to the corresponding dispersions of MdAEs in EbA. Except from the first dataset in which the comparative methods present nearly equal performances (CV_{EbA}=0.026 and CV_{LSEbA}=0.025), the results are better for LSEbA than EbA for the second (CV_{EbA}=0.155 and CV_{LSEbA}=0.116) and the third (CV_{EbA}=0.276 and CV_{LSEbA}=0.105) datasets. This practically means that the semi-parametric methodology appeared to be less sensitive to the choice of parameter k than the non-parametric approach.

Another issue that deserves commentary is the time consumption of the proposed methodology. It is clear that since the method combines EbA for the estimation of a set of variables and a subsequent application of LS regression, the calculation of an estimate is generally more time consuming than the simple methods. However, the relatively small number of cases in software project datasets does not causes any practical problems. Indicatively we can mention that for the first dataset the whole leave-one-out cross validation, i.e. the estimation of 759 projects was carried out in about 32 s using the statistical program Splus in a common personal computer. Accordingly, the estimation of all projects of the second dataset (93 projects) took almost 1.2 s while the whole third dataset (110 projects) took almost 1.5 s. Of course the time of a single estimation depends on the number of projects in the training dataset, the number of variables in each subset and the number of missing values but this time is generally reasonable for practical applications.

Conclusively, the main points summarizing the contribution of the paper are:

- The basic principles, the algorithm and the properties of LSEbA were presented systematically and illustrated by examples and extensive experimentation.
- The experimentation was based on representative datasets with structures that usually cause problems in the modeling procedure, (e.g. categorical variables, missing values, nonlinear relationships). The goal of the experimentation was to compare LSEbA with LS, EbA and a simple mixture of them in order to show that the combination of the two methods can address the problems of each one of them and take advantage from their best characteristics, so as to improve accuracy.
- The prediction accuracy of all models was validated by the leave-one-out cross validation procedure and measured by three different error functions.
- The comparison of LSEbA with LS, EbA and the mixed model resulting by averaging LS and EbA was based on summary statistics of prediction errors, on statistical tests and on visualization of the error distributions.

Apart from the encouraging results obtained by the evaluation of LSEbA on the datasets, there are also some challenging issues which deserve further research. First, the proposed semi-parametric method has to be applied to other software cost datasets. A more appealing research topic is the evaluation of the linear and non-linear components of the proposed methodology. In this study, we indicatively applied the ordinary least squares regression and the estimation by analogy technique for the computation of the parametric and non-parametric part, respectively. However, there are lots of other well-known methods that can substitute the abovementioned methods and can be used for evaluation of these components. For example, practitioners may use a robust regression in the computation of the parametric portion of the proposed model in order to have a model less vulnerable to outliers. Additionally, other non-parametric methods such as Neural Networks can replace estimation by analogy for the evaluation of the non-linear part.

Moreover, a very important question needing systematic treatment is the choice of the predictive variables that have to be assigned in the LS and EbA subsets. The method for separating the variables in the experiments of this paper used a simple criterion: all the continuous variables were assigned to the LS subset whereas all categorical variables to the EbA subset. Although this seems quite reasonable, having in mind the mixed-type nature of data, there is a need for the automation of the process. A future research goal is the introduction of an optimization algorithm that automates the procedure for variable partition in cases where there are a large number of potential explanatory variables and no underlying theory to base this partition.

Finally, various studies in SCE area suggest that it is desirable for all methods resulting in a "point estimation" to be accompanied by a prediction interval for this estimation. The main goal of prediction intervals is to provide an interval estimation consisting of an "optimistic" and a "pessimistic" guess for the true magnitude of the cost. Prediction intervals constitute an invaluable tool for project managers. Since there is not a way to compute confidence intervals through the utilization of the theory for semi-parametric models, bootstrap techniques that generate a large number of independent samples drawn with replacement from the original sample can be very helpful for evaluation of confidence intervals.

Acknowledgement We would like to thank the reviewers and the editor for their valuable comments which helped us to improve the paper.

References

Angelis L, Stamelos I, Morisio M (2001) Building a software cost estimation model based on categorical data. Proceedings of the IEEE 8th International Symposium on Software Metrics, pp. 4–15

Anglin P, Gencay R (1996) Semiparametric estimation of a hedonic price function. J Appl Econ 11(6):633–648

Bi J, Bennet K-P (2003) Regression error characteristics curves. Proceedings of the AIII 20th International Conference on Machine Learning, pp. 43–50

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

Cartwright MH, Shepperd MJ, Song Q (2003) Dealing with missing software project data Proceedings of the METRICS, pp. 154–165

Foss T, Stensrud E, Kitchenham B, Myrtveit I (2003) A simulation study of the model evaluation criterion MMRE. IEEE Trans Softw Eng 29(11):985–995

Hardle W (1990) Applied non-parametric regression. Economics Society Monographs, Cambridge University Press

Hardle W, Liang H, Gao J (2000) Partially linear models. Physica-Verlag, Heidelberg

ISBSG Dataset 10 (2007), http://www.isbsg.org

Jorgensen M, Shepperd M (2007) A systematic review of software development cost estimation studies. IEEE Trans Softw Eng 33(1):33–53

Kaufman L, Rousseeuw P (1990) Finding groups in data: an introduction to cluster analysis. John Wiley, New York

Kitchenham B (1998) A procedure for analyzing unbalanced datasets. IEEE Trans Softw Eng 24(4):278–301

Kitchenham B, Mendes E (2004) A comparison of cross-company and within-company effort estimation models for web applications. Proceedings of the Empirical Assessment in Software Engineering, pp. 47–55

Kitchenham B, Pickard L, MacDonell S, Shepperd M (2001) What accuracy statistics really measure. IEE Proc Software 148(3):81–85

Kitchenham B, Pfleeger L, McColl B, Eagan S (2002) A case study of maintenance estimation accuracy. J Syst Softw 64(1):57–77

Korte M, Port D (2008) Confidence in software cost estimation results based on mmre and pred. Proceedings of the 4th ACM International Workshop on Predictor Models in Software Engineering, pp. 63–70

Liebchen G, Shepperd M (2008) Data sets and data quality in software engineering. Proceedings of the 4th ACM International Workshop on Predictor Models in Software Engineering, pp. 39–44

Lokan C, Mendes E (2006) Cross-company and single-company effort models using the ISBSG database: a further replicated study. Proceedings of the ACM-IEEE International Symposium on Empirical Software Engineering, pp. 75–84

MacDonell S, Shepperd M (2003) Combining techniques to optimize effort predictions in software project management. J Syst Softw 66(2):91–98

Mair C, Shepperd M (2005) The consistency of empirical comparisons of regression and analogy-based software project cost prediction. Proceedings of the International Symposium on Empirical Software Engineering, pp. 509–518

Mendes E, Kitchenham BA (2004) Further comparison of cross-company and within company effort estimation models for web applications. Proceedings of the 10th IEEE International Symposium on Software Metrics, pp. 348–357

Mendes E, Lokan C (2008) Replicating studies on cross—vs single-company effort models using the ISBSG database. Emp Softw Eng 13(1):3–37

Mendes E, Lokan C, Harrison R, Triggs C (2005) A replicated comparison of cross-company and within-company effort estimation models using the ISBSG database. Proceedings of the IEEE 11th International Software Metrics Symposium, pp. 36–46

Mittas N, Athanasiades M, Angelis L (2008) Improving analogy-based software cost estimation by a resampling method. Inform Softw Technol 50(3):221–230

Mittas N, Angelis L (2008a) Combining regression and estimation by analogy in a semi-parametric model for software cost estimation. Proceedings of the ACM-IEEE 2nd International Symposium on Empirical Software Engineering and Management, pp. 70–79

Mittas N, Angelis L (2008b) Comparing cost prediction models by resampling techniques. J Syst Softw 81 (5):616–632

Mittas N, Angelis L (2008c) Comparing software cost prediction models by a visualization tool. Proceedings of the IEEE 34th Euromicro Conference on Software Engineering and Advanced Applications, pp. 433–440

Myrtveit I, Stensrud E, Olsson U (2001) Analyzing data sets with missing data: an empirical evaluation of imputation methods and likelihood-based methods. IEEE Trans Softw Eng 27(11):999–1013

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

NASA93 (2007) Dataset, http://promisedata.org/repository/#nasa93. (NASA93 2007)

Port D, Korte M (2008) Comparative studies of the model evaluation criterions mmre and pred in software cost estimation research. Proceedings of the ACM-IEEE 2nd International Symposium on Empirical Software Engineering and Management, pp. 51–60

Robinson P (1988) Root-n-consistent semiparametric regression. Econometrica 56(4):931–954

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

Shepperd M, Schofield C (1997) Estimating software project effort using analogies. IEEE Trans Softw Eng 23(11):736–743

Sheskin DJ (2004) Handbook of parametric and nonparametric statistical procedures (Third Edition) Chapman & Hall/CRC

Strike K, Emam KE, Madhavji N (2001) Software cost estimation with incomplete data. IEEE Trans Softw Eng 27(10):890–908

Wissmann M, Toutenburg H, Shalabh (2007) Role of categorical variables in multicollinearity in the linear regression model. Technical Report, Number 008, Department of Statistics, University of Munich

Nikolaos Mittas received his B.Sc. degree in Mathematics from University of Crete, his M.Sc. and Ph.D. degree in Informatics from Aristotle University of Thessaloniki (A.U.Th). His research interests involve application of statistics, especially computational statistics, to cost estimation of software projects and generally to data from software projects.

Lefteris Angelis received his BSc 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.