The Journal of Systems and Software 83 (2010) 621–637

Contents lists available at ScienceDirect

The Journal of Systems and Software

journal homepage: www.elsevier.com/locate/jss

Visual comparison of software cost estimation models by regression error characteristic analysis

Nikolaos Mittas, Lefteris Angelis *

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

ARTICLE INFO

Article history:
Received 22 November 2008
Received in revised form 23 July 2009
Accepted 30 October 2009
Available online 4 November 2009

Keywords:
Estimation by analogy
Regression analysis
Regression error characteristic curves
Software cost estimation

ABSTRACT

The well-balanced management of a software project is a critical task accomplished at the early stages of the development process. Due to this requirement, a wide variety of prediction methods has been introduced in order to identify the best strategy for software cost estimation. The selection of the best technique is usually based on measures of error whereas in more recent studies researchers use formal statistical procedures. The former approach can lead to unstable and erroneous results due to the existence of outlying points whereas the latter cannot be easily presented to non-experts and has to be carried out by an expert with statistical background. In this paper, we introduce the regression error characteristic (REC) analysis, a powerful visualization tool with interesting geometrical properties, in order to validate and compare different prediction models easily, by a simple inspection of a graph. Moreover, we propose a formal framework covering different aspects of the estimation process such as the calibration of the prediction methodology, the identification of factors that affect the error, the investigation of errors on certain ranges of the actual cost and the examination of the distribution of the cost for certain errors. Application of REC analysis to the ISBSG10 dataset for comparing estimation by analogy and linear regression illustrates the benefits and the significant information obtained.

1. Introduction

A vital issue a project manager has to be faced with is the obligation for a wise decision-making in order to obtain the maximum return on investment of a forth-coming project. Due to this fact, a variety of *software cost estimation* (SCE) techniques has been proposed and studied during the last decades (Jorgensen and Shepperd, 2007). The main research question that has to be addressed is the identification of the "best" prediction method since each estimation model attempts to evaluate a function that is able to predict accurately the cost of a new project.

The selection of the "best" prediction method is usually based on accuracy indicators obtained by certain functions of the errors, without taking into account any information of their underlying distribution. This policy for the determination of the superiority of one method against a comparative one can lead to unstable and misleading results, since a few outlying error points can distort the entire error distribution. Moreover, two comparative methods can be identical with respect to the average of errors they produce, but these errors can have entirely different distributions, i.e. the errors can be governed by different laws. In fact, indicators that show a central tendency of the overall predictive performance of a model do not provide enough information about the performances of the predicting methods on specific ranges of the actual cost value, for example projects with small or high actual cost.

An appealing research topic is therefore the study of the error produced by comparative cost estimation methods on specific ranges of values of the actual cost. This corresponds to a realistic situation where for example a project manager wishes to know whether a model provides better predictions for projects of low cost than its comparative model which can be more accurate for projects of high cost. Also, one may be interested in investigating how the errors are distributed for different categories of projects so as to assess the suitability of a method for certain project types (for example application domains). Another interesting issue is to investigate and locate domains of values of the actual cost that are prone to low or high errors.

In more recent studies (see Section 2), the researchers make use of statistical hypothesis tests in order to obtain a more accurate and efficient perspective for the highly sensitive task of comparison and selection between methods. Since the software cost datasets are usually small-sized and the error functions highly skewed, there is a need for more robust procedures. So, apart from the more traditional parametric tests which assume a specific theoretical distribution (like the *t*-test or the analysis of variance), non-parametric tests have also been used (like the Wilcoxon or the Mann–Whitney test) and more recently resampling techniques

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

0164-1212/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jss.2009.10.044

(like the bootstrap and the permutation tests) have been proposed to compare alternative prediction methods. On the other hand, all these well-defined statistical techniques cannot be easily presented to non-experts. An expert with strong statistical experience should be involved in the comparison procedure in order to carry out various hypothesis tests, interpret the results and infer about the superiority of one method against the other.

All of the issues discussed above lead us to conclude that there is an imperative need to use and develop highly readable and easily-interpretable tools which will aid the decision-maker to thoroughly investigate the predictive behavior of different models. And this brings us to the motivation of the present study, which is to suggest a graphical technique for exactly this purpose.

The main goal of this study is to further extend the two previous studies (Mittas and Angelis, 2008b,c) concerning the utilization of *regression error characteristic* (REC) curves in SCE. Analysis by REC curves constitutes a framework based on the visualization of error and facilitates the project managers to compare several different prediction methods by a single inspection of a graph. REC curves have appeared as a recent extension of *Receiver Operating Characteristic* (ROC) curves (Egan, 1975), a well-known graphical procedure utilized in order to evaluate the predictive performance of classification learning algorithms. Due to the limitation of ROC curves in classification problems, Bi and Bennet (2003) generalized the idea of ROC curves to regression problems with similar benefits such as the characterization of accuracy, efficacy and quality of different predicting methods. In simple terms, a REC curve represents the cumulative distribution of the error produced by a cost estimation method, by simply plotting the error values on the *x*-axis and the accuracy of the prediction method on the *y*-axis. Hence, a REC graph consists of one or more monotonically increasing curves, each corresponding to the cumulated errors of a prediction method.

Our aim is therefore to introduce the capabilities of this graphical tool, adjusted for the SCE problem, in a systematic framework that will facilitate the project managers to acquire significant information and guide the management of the software development regarding both the validation of a prediction technique and the comparison of alternative methods. Despite the main scope of the comparisons of different prediction techniques, we show that the knowledge discovery obtained by the REC analysis can be valuable for the calibration of the prediction methodology, the identification of factors affecting the error, the exploration of behavior of alternative models over a certain range of the actual cost and the identification of how errors within a certain range are distributed along the domain of the cost variable. Moreover, we investigate interesting geometrical characteristics and concepts related with the REC curves, such as the *area over the curve* (AOC) that is an estimate of the mean expected error.

The rest of the paper is organized as follows: Section 2 summarizes related work and specifies the contribution of the present work. In Section 3, we present the REC analysis and we give the algorithm in order to plot the REC curves with some interesting topics about their geometrical characteristics. In Section 4, we detail the methodology followed, the description of the comparative models and the measures of errors that are used in the applications on real data. In Section 5, we present the way that REC analysis can be adapted in SCE, whereas we illustrate certain guidelines for the usage of REC curves in different aspects of the estimation process. In Section 6, we demonstrate the results of the application of the analysis to the comparison of two well-known methods over real data. Finally, in Section 7 we conclude by discussing the results and by providing some directions for future research.

2. Related work and contribution

During the past decades a plethora of methods and models for the cost estimation of a software project has been proposed. A systematic review of studies (Jorgensen and Shepperd, 2007) reveals that the most common research topic is the introduction and evaluation of estimation methods. Regression-based approaches significantly dominate since half of all studies deal with the fitting, improvement or comparison of a regression model (Jorgensen and Shepperd, 2007). Some other interesting conclusions of the abovementioned study are that the proportion of analogy-based and expert judgment-based methods is increasing and also that recently other more sophisticated estimation approaches have been proposed (i.e. genetic and linear programming, fuzzy logic modeling and bootstrap-based analogy cost methods) (Jorgensen and Shepperd, 2007).

Despite the evolutionary research concerning the identification of the "best" prediction technique, there seems to be no global answer for all kinds of data. Indeed, the results are widely contradictory, complicating the critical task of project managers to plan, schedule and monitor the development process. A representative example of the controversial findings can be derived by Mair and Shepperd (2005) where researchers combined the results from 20 individual studies, covering the period from 1997 to 2004. These studies used two of the most known cost estimation methods, namely regression and *estimation by analogy* (EbA). Mair and Shepperd found that in 12 out of 20 studies EbA performed better than regression. They also pointed out that only six of the papers (30%) made use of statistical hypothesis testing in order to evaluate the predictive performance of the two comparative models. The researchers discuss that these differences in findings may arise from the lack of standardization in software research methodology which leads to heterogeneous sampling, measurement and reporting techniques.

Recently, Kitchenham and Mendes (2009) discuss the potential sources of the divergent results and underline that there is need to establish guidelines for acceptable comparative cost estimation studies. More analytically, they emphasize that researchers should use appropriate statistical tests when alternative models are compared. Indeed, Stensrud and Myrveit (1998) point out that it is invalid to compare prediction techniques without performing some tests to statistically verify the observed differences in accuracy measures. Although many researchers (Briand et al., 2000; Jeffery et al., 2001; Kitchenham and Mendes, 2004; Mendes et al., 2008; Myrtveit and Stensrud, 1999) take into account the important role of statistical procedures for testing comparative prediction techniques and assessing the superiority of a model against a comparative one, many recent papers still base their findings solely on accuracy statistics without performing any appropriate test (Kitchenham and Mendes, 2009).

Mittas and Angelis (2008a) also remark the divergence in comparative studies indicating that the reason may be the evaluation method utilized in the comparisons of the results obtained by different prediction models. Specifically, they underline that the majority of studies base their inferences on single statistics of errors measured by different functions. Thus, when the researchers compare models based solely on a single value that contains significant variability, they take the risk to consider as significant a difference which in fact may be not so significant. Other important issues that have to be taken into account in the comparison of alternative prediction methods is the fact that these are usually based on small datasets and produce errors skewed, non-normally distributed. In the same paper it is proposed that certain drawbacks of the traditional statistical tests can be addressed by the usage of simulation resampling techniques. In this regard, Kitchen-

ham and Mendes (2009) also point out that the resampling approach has significant potential value.

From what we have already mentioned, it is clear that researchers should use appropriate statistical tests on indicators of error functions in order to justify the selection of the "best" prediction technique. On the other hand, these indicators provide a general tendency of the predictive performance for the entire dataset and do not present enough information about the distribution of errors and the partial performances of the models on specific ranges of interest. The latter constitutes valuable information for a project manager who may prefer to have a closer inspection on the whole distribution of errors for different comparative models on specific ranges of the cost variable in order to detect significantly different performances over different cost values.

Although statistical tests constitute an invaluable procedure for comparing alternative prediction models, they are often overlooked in favor of simplicity, whereas they are sometimes difficult to be interpreted by non-experts. Moreover, a practitioner would like to acquire a much more appealing presentation of the "competitive" models' results than tables of statistics. Inspection of certain characteristics of the errors distributions (like various percentiles) can be facilitated by a graphical comparison method. To the best of our knowledge, the graphical comparison of SCE methods has not been systematically studied yet, although it seems to be of great importance for both the validation and the comparison of the performances of alternative prediction models.

In our previous conference studies (Mittas and Angelis, 2008b,c), we briefly presented the basic concepts and the algorithms for the construction of standard (Mittas and Angelis, 2008b) and partial (Mittas and Angelis, 2008c) REC curves, whereas the prediction methods which were used in the applications, were regression and EbA. The analysis, applied to two small datasets, was based on three local measures of prediction performance; the *magnitude of relative error* (MRE), the *magnitude of relative error to the estimate* (MER) and the *absolute error* (AE).

This paper extends the aforementioned studies and its main contribution can be summarized in the following issues:

- The application of REC curves in SCE is presented in a more systematic fashion by detailing the basic principles, properties and the algorithm for constructing them.
- A formal framework and structured guidelines are introduced for illustrating the benefits of the REC analysis covering two different aspects of the estimation process. In particular, we present the way that REC analysis can be adapted for the validation of a prediction technique and for the comparison of alternative models.
- Interesting geometrical characteristics of REC curves such as the area over the curve (AOC) are investigated and used in order to evaluate graphically the most common location statistics (mean and median) of errors, measured by different functions.
- REC analysis is introduced for providing graphical inspection of predictive performance by visualizing various aspects of error distributions. Three functions of errors, measuring different important aspects of the prediction performance are studied: The absolute error (AE), the residual (or simply the error) and the magnitude of relative error (MRE) that are related to the accuracy, the bias and the spread of errors of a model, respectively. Furthermore, REC curves are utilized for identifying the presence of extreme outlying errors and for inferring whether the prediction model is prone to under- or overestimations.
- The analysis is based on a relatively large and more recently released *International Software Benchmarking Standards Group* (ISBSG) dataset, whereas the REC curves are used for the comparison of regression and EbA methods for the prediction of project's duration.
- The usage of REC curves is extended to the identification of factors that may affect the predictive error measures in various manners and to the calibration of the EbA method.
- Partial REC curves are used to investigate the performance of the comparative models on specific ranges of the dependent variable (for example only small or large values) and more generally, to identify portions of the dependent variable range that are prone to certain degrees of errors.

3. Regression error characteristic analysis

3.1. Standard regression error characteristic curves

Regression error characteristic (REC) curves, proposed by Bi and Bennet (2003), constitute a powerful technique for the visualization and inspection of several comparative prediction models. REC curves plot simultaneously the *cumulative distribution functions* (CDF) of the prediction errors, obtained by different models, offering a graphical technique to estimate the probability of the error to be less or equal than a certain x-axis value.

Let Y be the real random dependent variable representing in our context the cost of a project and let $\mathbf{X}$ be a p-dimensional random vector with coordinates representing the variables or attributes which characterize the projects in an available historical data set. Our goal is to find a regression function $f(\mathbf{X}) = E(Y|\mathbf{X})$, where by $E(Y|\mathbf{X})$ we denote the conditional mean or expected value of Y given the vector $\mathbf{X}$. This regression function will be used for building a prediction model of the form

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

where Y_i is the ith observation of the dependent variable (i.e. the cost for the ith project) to be explained by variations in the vector of predictors $\mathbf{X}_i$ with dimensions $1 \times p$. We assume that the errors $varepsilon_i$ are independent with zero mean. In the statistical literature, various approaches have been proposed for the estimation of $f(\mathbf{X})$. There are parametric models such as the least squares regression and non-parametric models such as the k-nearest neighbor non-parametric regression, which is essentially the well-known EbA method in SCE.

Suppose now that someone has to evaluate the performance of the predictor function $f(\mathbf{X}_i)$ based on the prediction errors obtained by a validation procedure (such as the *leave-one-out cross-validation*). Various local accuracy measures that evaluate different aspects of the errors (for further details see Section 4) have been introduced, so far. Denote by e_i the prediction error of the ith observation of the dependent variable measured in a certain manner. The REC curve of the model is the CDF of the empirical distribution of prediction errors.

Briefly, a REC curve is a two-dimensional plot where the horizontal (or x-axis) represents the error tolerance of a predefined accuracy measure and the vertical (or y-axis) represents the accuracy of a prediction model. Accuracy is defined as the percentage of projects that are predicted within the error tolerance e

$$accuracy(e) = \frac{\#(\text{projects with error} \leqslant e)}{\#(\text{projects})}. \tag{2}$$

In Fig. 1, we present an example of the REC curves obtained by three different hypothetical comparative prediction models. Generally, a prediction model performs well if the REC curve climbs rapidly towards the upper left corner. Indeed, this behavior shows that the whole distribution of errors is kept below a small tolerance value. On the contrary, curves extending slowly to the upper right corner indicate the existence of very large errors.

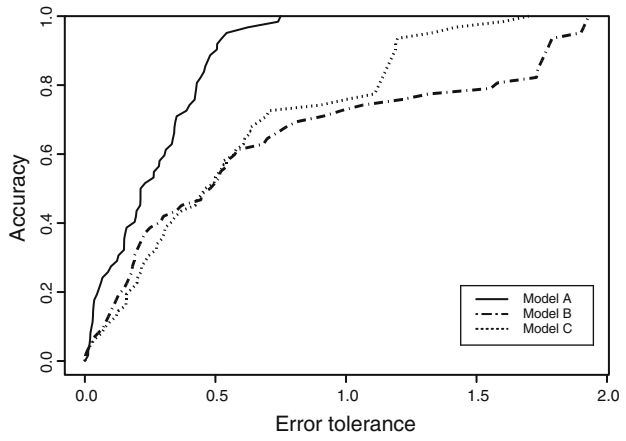

Fig. 1. REC curve example.

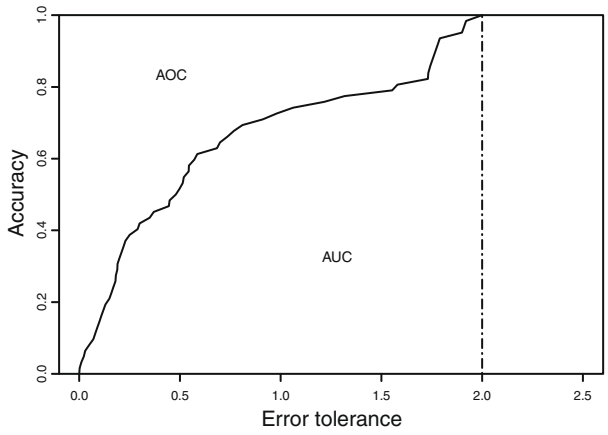

Fig. 2. REC curve example with AOC and AUC.

Obviously, there is a positive association between the error tolerance and the accuracy of a predictor and therefore a REC curve is a monotonically increasing function. It is also clear that there are two extreme conditions concerning the construction of the curve. By setting error tolerance $e = 0$, only the projects that their predictions are identical to their actual values are considered accurate. On the other hand, if we set $e > \max_{\{i=1,\ldots,n\}}(e_i)$, where n the total number of projects, then all the projects are considered accurate.

In the example of Fig. 1, we can conclude that Model A, represented by the solid line, has clearly the best prediction performance since its REC curve always dominates the corresponding curves of both Model B and Model C over the whole range of possible error tolerances. Regarding the comparison between Model B and Model C, there seems to be no global superiority of one model against the other since their REC curves are intersected. Initially, for small error tolerances, Model B shows better performance but as the error tolerance increases and becomes higher than 0.4, Model C dominates.

3.2. Geometrical properties of regression error characteristic curves

As we have already mentioned, a REC curve is an estimate of the predictive error's CDF function and can be used for the comparison of the performance of several prediction models. In addition, there are a few interesting geometrical characteristics that deserve investigation. The most significant one is that commonly used measures of the distribution such as the mean or the median of errors can be estimated by exploiting the geometry of a REC curve.

Analogous to the area under the ROC curve (AUC) that provides an estimate of the expected accuracy of classification methods (Egan, 1975); Bi and Bennet (2003) prove that the area over the curve (AOC) is an estimate of the expected mean error. As the comparison of models in SCE literature is usually based on such statistics of error, we illustrate how these can graphically be obtained from the REC curves, presenting an alternative methodology for the comparison procedure.Suppose that a hypothetical prediction model is constructed on a specific dataset and we evaluate the performance through its error CDF function. In Fig. 2, we plot the REC curve computed by the error CDF function in which the vertical dashed reference line corresponds to the maximum value of observed errors ($e_{\max} = 2$ in the example). The REC curve partitions the rectangle defined by the vertical line and the x and y axes in two plane regions denoted by AOC and AUC. Obviously, their relation is

$$\mathrm{AOC} = e_{\max} - \mathrm{AUC}. \tag{3}$$

Bi and Bennet (2003) proved that AOC is an estimator of the expected mean error $E(e)$ of a prediction model. Concerning the measures appearing in the SCE literature, we have to adjust Eq. (3) in order to evaluate the prediction performance of a model. In practical terms, if e is calculated using a certain measure like AE, residual or MRE (see Section 4, Table 1), then the theory implies that the AOC is an approximation of the corresponding MAE, mean residual and MMRE (see Section 4, first column of Table 2).

On the other hand, as software project cost datasets are usually small and highly skewed (Kitchenham et al., 2001; Mittas and Angelis, 2008a), the researchers make use of more robust statistics, such as the median, in order to evaluate the overall prediction performance of a model. Another interesting property of the REC curve is that we can easily assess the median of the error CDF function in a straightforward manner.

In Fig. 3a, we also present the REC curve of another hypothetical prediction model evaluated through the error CDF. The horizontal dashed reference line from 0.5 intersects the REC curve in a point which corresponds to $e = 0.21$ (vertical dashed line). This means that based on the specific sample of errors, 50% of projects are expected to have an error smaller than 0.21 which is the median value of errors. In this way, we can geometrically evaluate the overall median prediction error. In the case of MRE, a similar approach can be followed for the estimation of the pred25 accuracy measure, which is the percentage of projects having MRE less than or equal to 0.25 (Conte et al., 1986), by drawing a reference vertical line from 0.25 and from the intersecting point of the REC curve, a horizontal line to meet the accuracy axis (Fig. 3b). In our example, the REC curve reveals that 48% of projects are expected to have MRE less than or equal to 0.25.

Summarizing the above discussion, it becomes clear that the REC curve has certain geometrical properties that allow the graphical representation of well-known accuracy measures. It therefore offers a straightforward basis for visual inference of a model's predictive accuracy and for the comparison of different models.

3.3. Partial regression error characteristic curves

The standard REC curve, described in the previous sections, provides a graphical mean for evaluating the overall performance since it portrays the prediction error across the whole range of

Table 1
Local accuracy measures.

$AE_i = \lvert Y_{A_i} - Y_{E_i} \rvert$	$residual_i = Y_{A_i} - Y_{E_i}$	$MRE_i = \frac{\lvert Y_{A_i} - Y_{E_i} \rvert}{Y_{A_i}}$

Table 2
Global accuracy measures.

$MAE = \frac{1}{n}\sum_{i=1}^{n} AE_i$	$MdAE = median\{AE_i\}$
$Mean\ residual = \frac{1}{n}\sum_{i=1}^{n} residual_i$	$Median\ residual = median\{residual_i\}$
$MMRE = \frac{1}{n}\sum_{i=1}^{n} MRE_i$	$MdMRE = median\{MRE_i\}$

all possible values. Very often however, there are cases in which an analyst requires further information on how the errors are distributed across a specific range of the dependent variable. Practically, the problem raises from the fact that even if we have two comparative models with exactly the same accuracy over a certain error tolerance, the observed errors might have been obtained by different actual values of the cost variable.

Although this is not a major problem when someone has to decide upon the overall predictive performance of comparative models, the additional information of the superiority of a certain model against a comparative one over a specific range of interest (i.e. projects with small, medium or high actual cost values) cannot be derived from the construction of standard REC curves. Moreover, two comparative models can have different predictive performances on certain values of the actual cost response. It is therefore beneficial for someone to be able to focus on specific cost ranges in order to get deeper insight into the behavior of error and maybe to ensemble predictions from different models so as to achieve optimal overall performance.

The above issues can be addressed by the utilization of *partial regression error characteristic* curve (Torgo, 2005) that is a special case of standard REC curves where the analysis of the error distribution is limited inside a certain range of the dependent variable. The partial REC curves can be easily constructed in the same way as the standard REC curves by simply estimating the CDF of the prediction errors only for those cases falling within the range of particular interest.

The benefits of using partial REC curves can be illustrated by the following example. Suppose that we wish to evaluate the predictive performance of three hypothetical comparative models separately for small and high values of the actual cost. After the evaluation of errors obtained by the three comparative predictors, the dataset is partitioned into two subsets; the first containing projects with small actual costs and the second, projects with high costs. The subsets were formed by simply dividing the cost values below and above their median.

The partial REC curves for small (Fig. 4) and high (Fig. 5) costs suggest that Model A outperforms both Model B and Model C everywhere, since its curves are always above the corresponding curves of Model B and Model C. On the other hand, Model B seems to outperform Model C for projects with low cost (Fig. 4) but the opposite holds for the projects with high cost (Fig. 5). Now, let us just ignore for the needs of the example Model A which is the best overall. After a rough estimation of the project's cost level (low or high) we would choose Model B for low cost forth-coming projects and Model C for high cost projects, combining in this way the predictions derived from the two comparative models.

We can see in both figures that the accuracy of all the models attains the value 0.5 since the initial dataset is partitioned into two subsets that contain half (50%) of the projects.

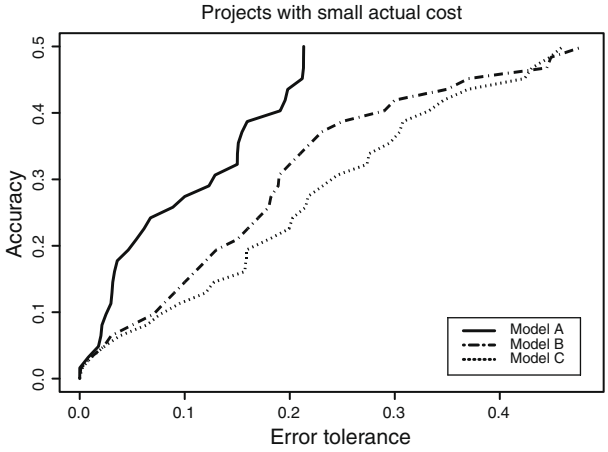

Fig. 4. Partial REC curve example for small actual costs.

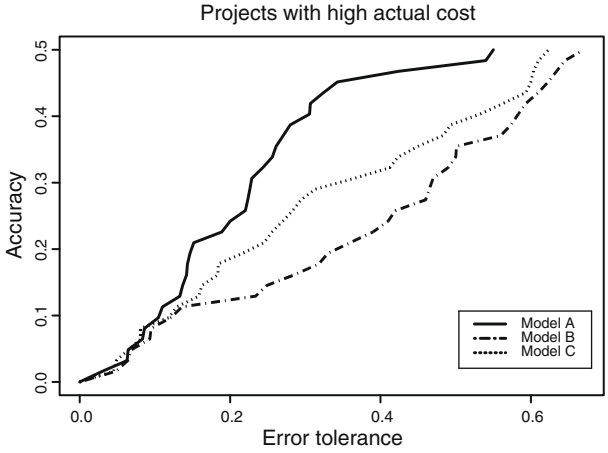

Fig. 5. Partial REC curve example for high actual costs.

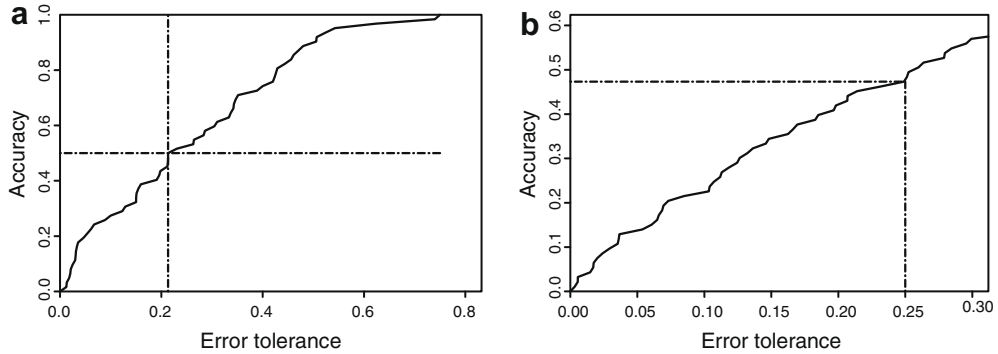

Fig. 3. REC curve example for the evaluation of (a) median error and (b) pred25 measure.

3.4. Algorithm description

In this section, we analytically present the algorithm for the construction of REC curves. The algorithm given by Bi and Bennet (2003) can be easily adjusted to the needs of SCE data and can be described by the following steps:

1. Fit a prediction model and use a validation procedure in order to evaluate the predictions of the dependent variable (some expression of cost).
2. Use a "local" accuracy measure in order to estimate the errors e_i of each project.
3. Define the range of interest for the actual values of the response cost variable.
4. Sort the errors e_i that fall into the range of interest in ascending order and evaluate the *accuracy*(e) of the prediction model at error tolerance e which ranges from min(e_i) to max(e_i).
5. Plot error tolerance e versus *accuracy*(e) and draw interpolating straight lines between the plotted points.

Note that Step 3 refers to the construction of partial REC curves (Section 3.3). The standard REC curves which are based on the whole set of errors can be evaluated by omitting Step 3 of the aforementioned algorithm.

The monotonically nature of the REC curves implies that as e increases the *accuracy*(e) increases. The *accuracy*(e) reaches 1 (or 100%), when the e value becomes larger or equal to the maximum value of error obtained by the prediction model.

In the plots constructed by the aforementioned algorithm, apart from the REC curves of the models we want to evaluate or compare, there is also another curve which plays the role of a reference model for assessing the suitability and appropriateness of the specific prediction techniques on a certain dataset. This curve represents the "*Mean model*" for each dataset.

The Mean model basically ignores the effects of predictor attributes and estimates the cost of each project by the mean cost of all the others. It is essentially an expression of the "worst model" that someone can invent, without taking into account any information from the project characteristics. A prediction technique should be therefore considered meaningful, if it significantly outperforms the Mean model obtained by the same data (Bi and Bennet, 2003). In our applications, we often employ the Mean model for a common reference of the other comparative models.

4. Methodology

As we have already mentioned, REC curves offer a handy tool for making visual comparisons between different prediction techniques. Moreover, as we saw in Section 2, both regression analysis and EbA are approaches that have been extensively used and studied in SCE, whereas the findings of different studies reveal a lack of convergence when attempting to determine the "best" prediction technique. Therefore, we decided to deal with these specific methods. As the goal of this paper is to further extend the research regarding the means of comparison between cost models, we have to emphasize that it is not our intention to compare the aforementioned methods and suggest that one of them is better than the other, but rather to further contribute in the systematic comparisons of cost prediction methods in general.

Ordinary Least Squares Regression (the acronym LS is used in the rest of this study) is the statistical technique whereby a linear parametric model is built in an attempt to explain the relationship between a numerical dependent variable and a set of independent variables. On the other hand, EbA is a type of non-parametric regression procedure (Mittas et al., 2008) where the unknown values of the dependent variable are estimated by the known values, of the same variable, corresponding to neighbors (analogies) of the estimated case (Shepperd and Schofield, 1997). Analogies are found through the evaluation of a prefixed similarity (or dissimilarity) criterion of cases, based on the independent variables. So, this method tries also to explain the relationship between the dependent variable and the independent variables, but this relationship is not expressed in the form of an explicit function.

The validation method used in order to evaluate the predictive accuracy and perform the REC analysis was the leave-one-out cross-validation technique: At each step of the procedure a completed project i was removed from the dataset and the remaining projects were used as the basis for the estimation $\hat{f}_i(\mathbf{x})$ of the unknown predictor function which next provides the estimated cost Y_{E_i} of the removed project. The method allows us to obtain a sample of the prediction error, since the actual cost Y_{A_i} of the removed project is known.

The next issue to be addressed is the selection of measures that describe the accuracy of a prediction model. To this regard, there is also a lack of agreement about which accuracy measure is the most appropriate in order to compare the predictions obtained by alternative models (Kitchenham et al., 2001). Several indicators have been used in the SCE literature, describing different aspects of the predictive performance of models (Foss et al., 2003). In our application, we decided to use three meaningful measures of local error (Table 1) in order to obtain a comprehensive view of the predictive accuracy:

1. The *absolute error* (AE).
2. The *residual or error.*
3. The *magnitude of relative error* (MRE).

These local measures (the term "local" refers to the individual error of each project) are the basis for the estimation of the overall prediction performance of each model through the global measures given in Table 2.

The AE is used in order to measure the *accuracy* whereas the residual is a measure of *bias* for the predictions obtained by a model. On the other hand, the most commonly used measure is the MRE. However MMRE has been criticized (see for example the paper by Foss et al. (2003)) that it does not always select the "best" model. Moreover, Kitchenham et al. (2001) suggest that MMRE essentially measures the *spread* of the z-values obtained by dividing the estimated cost by the actual cost. This ratio is clearly related to the distribution of the residuals and that is why MMRE and MdMRE are considered as measures of the spread of the residuals.

In order to examine the effectiveness of the REC curves to detect differences between comparative models, we decided to use non-parametric statistical tests due to the fact that these measures are usually non-normal (Kitchenham et al., 2001; Kitchenham and Mendes, 2004; Mendes and Kitchenham, 2004; Mittas and Angelis, 2008a). Specifically, the *Wilcoxon signed rank test* was used for related (paired) samples, whereas the *Mann–Whitney* for independent samples.

5. Formal framework and guidelines for REC analysis

As the main scope of this study is the introduction of a visualization framework that utilizes both standard REC curves (Section 3.1) and partial REC curves (Section 3.3) for the better management of the SCE procedure, in this section we analytically present systematic guidelines that a practitioner can follow in order to exploit the potentialities of REC analysis. We have to emphasize that REC

analysis can be proved beneficial for two different aspects of the estimation process:

• *Validation* of prediction method. The proposed analysis can be used in practice in order to validate a certain prediction technique, covering different aspects such as the calibration process, the applicability and the appropriateness of the method to a specific dataset and the thorough study of the behavior of the derived prediction model (Section 5.1 – Fig. 6).

• Comparison of alternative prediction models. A practitioner can acquire significant information through REC analysis in the comparison procedure for both the overall and partial prediction performances of alternative models (Section 5.2 – Fig. 7).

5.1. Validation of prediction model

As most of the proposed estimation techniques are usually based on historical data, these methods involve many design decisions and require specific adjustments that have to be examined in order to calibrate the methodology on each available dataset. For example, we can consider the calibration of the EbA technique in which the practitioner should decide about the distance metric which determines the closest projects (neighbors), the number of neighbors to be used for prediction and the statistic for the computation of the cost estimation. REC analysis through the construction of different curves for each value of the parameter under consideration can provide an easy way for graphical selection of the best parameters in order to produce accurate predictions (Fig. 6).

Moreover, REC curves provide a mechanism to assess the appropriateness of an estimation model on a specific dataset. This constitutes another innovation of REC analysis since most of the prediction techniques do not provide a typical way to signify whether a model can be used in order to predict the cost of a new project. This specialized issue can be resolved through the comparison of a model's curve with the curve evaluated by the Mean model (see Section 3.4). Such graphical information allow project managers to make decisions based on the available information of a specific dataset since different dataset characteristics may favor different prediction systems (Mair and Shepperd, 2005).

At the second level of the validation for a prediction model (Fig. 6), one of the potential usages of REC analysis is the identification of factors affecting the distribution of errors. In most of

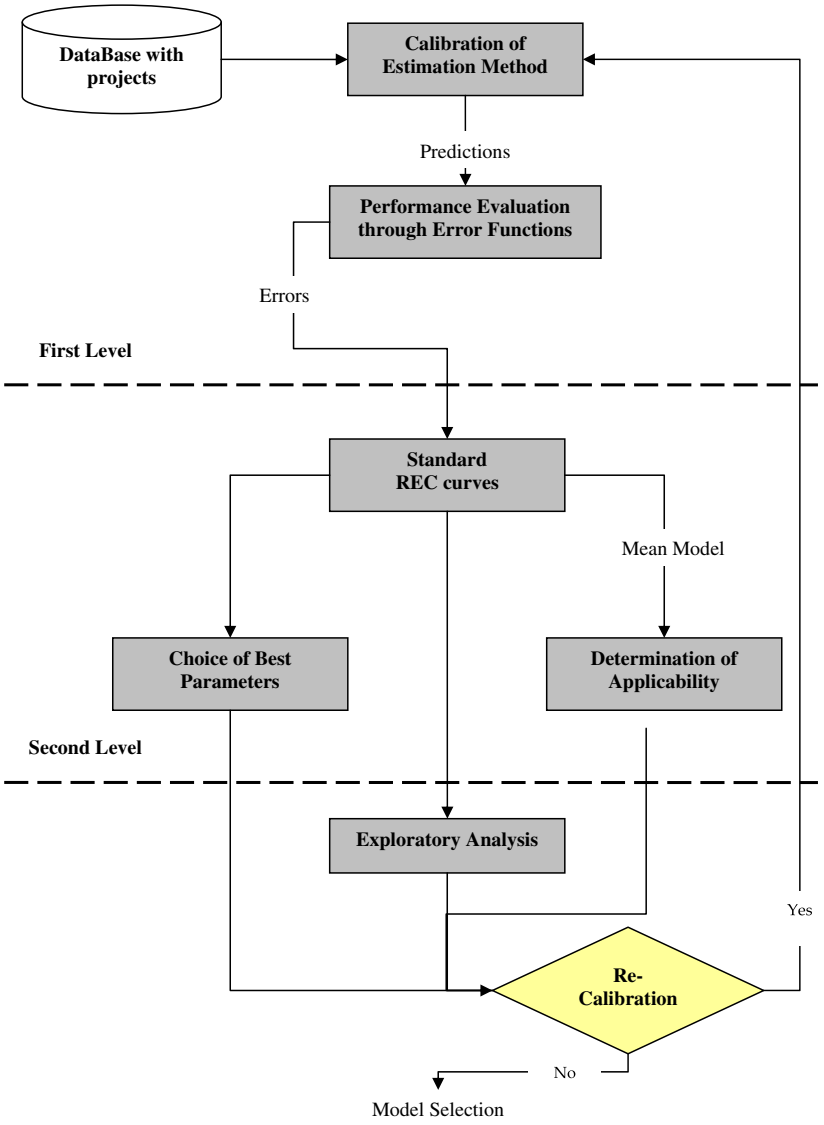

Fig. 6. Validation of a prediction model.

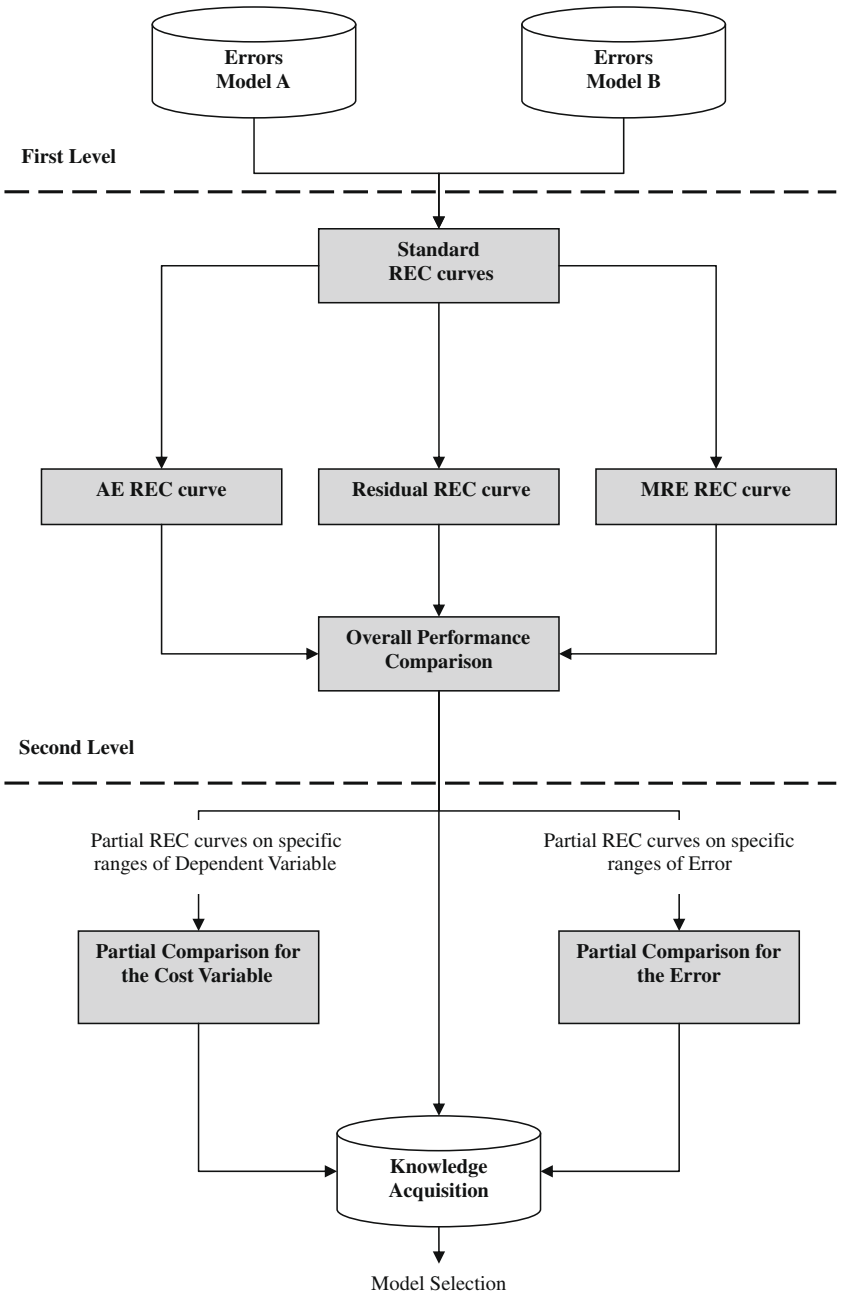

Fig. 7. Comparison of alternative prediction models.

the available SCE datasets there are various categorical variables (or factors) that characterize each project, such as the development type, the organization type, and the language type. The values of each factor are called levels and they partition the dataset into separate subsets of projects. It is therefore very useful for a practitioner to investigate whether there are differences among the REC curves representing these subsets, since these differences may indicate appearance of systematic error. Hence, REC analysis facilitates the practitioner to gain insight into the prediction capability of a model, explore and reveal potential sources of errors.

Summarizing, a project manager can exploit the additional information contained in the REC curve beyond simple statistics, whereas the investigation of the error distributions can guide the modeling process in the following issues:

- Choice of the best parameters.
- Applicability of the estimation method to a specific dataset.
- Exploratory analysis and identification of factors affecting the predictive power.

5.2. Model comparison

One of the most important decisions for a project manager is the selection of the "best" prediction technique and this fact is depicted from the various studies appeared so far in the literature (see Section 2). The necessity for the introduction of procedures that provide easily-interpretable results for the selection of the "best" prediction method becomes more and more compulsory. As we have already mentioned (Section 1), the determination of the "best" model should not be based solely on indicators derived from the distribution of errors. On the other hand, REC analysis provides a systematic framework for the difficult task of the model selection, whereas it can be taken place at two different levels.

At the first level of the analysis, standard REC curves can be utilized in order to compare the overall prediction performance of the competitive models (Fig. 7). Due to the fact that alternative error functions measure different aspects of the performance (see Section 4), a practitioner is able to obtain a global perspective of the whole error distributions and select the "best" model through a conceptually simple graphing method.

In a more complicated task, the practitioner may require additional information about the performances of the alternative models on specific ranges of the cost variable. This practically means that there are cases in which the selection of the best prediction technique is vital when the cost of any prediction error has not the same impact for the organization. For example, the projects with high actual cost are of particular interest since the consequences of an inaccurate prediction can lead to catastrophic results for an organization due to high overruns. On the other hand, an inaccurate prediction for a project with small actual cost does not give the chance to bid other contracts in order to increase the profits of the organization. The abovementioned issues cannot be resolved and studied through the examination of the overall measures of errors since these statistics provide limited information about the general predictive power of the models. In this regard, partial REC curves give the opportunity to obtain a further degree of detail by plotting at the second level of the analysis (Fig. 7) the error distributions on specific ranges of the cost or the error variables.

6. Illustrative application of REC analysis

In this section, we present the results of the application of REC analysis based on the formal approach that has been described in Section 5. Initially, we illustrate the way that REC analysis can be applied for the validation of EbA technique. In particular, we display how the optimum parameter for the nearest neighbors can be explored, whereas we also investigate the appropriateness of the methodology on the dataset of our application. Finally, an exploratory analysis is carried out in order to identify factors that may affect the predictive power of the derived model.

After the tuning and the determination of the best EbA model through REC analysis, we compare the performances of EbA, and regression analysis in order to select the "best" prediction technique. More precisely, we explore the distributions of the competitive models across the range of all possible errors (overall performances) through standard REC curves, whereas through partial REC curves, we also compare their distributions of errors over certain ranges of the cost variable and examine for which domains of values of the cost variable certain errors are more frequent. .

6.1. Dataset description

In this study, we use the *International Software Benchmarking Standards Group* (ISBSG) dataset (release 10) (ISBSG, 2007) which is a well-known, multi-organizational and international repository of completed software projects that can be used for estimating the required cost for new projects.

The initial dataset contains 4106 software projects from over 25 countries but most of the variables have a large amount of missing values. As the ISBSG guidelines suggest, we have to select a suitable subset of projects, in order to make our analysis meaningful. Due to this fact, we finally choose to work with the proposed dataset that is also utilized by the ISBSG Early Estimate Checker V5.0 (ISBSG, 2007).

Hence, the dataset contains 1530 completed projects with four independent predictors (three nominal and one continuous), whereas the dependent variable is the *Duration* (Table 3) which is closely related to the cost. Following the recommendation of ISBSG, we ignored the projects rated with *C* and *D* in variables *Data Quality Rating* and *Ufp Rating* (i.e. unadjusted function points rating) and we worked with the remaining projects rated with *A* and *B*. From variable *fp Standards* (i.e. function points standards), we selected all variants of IFPUG (International Function Point Users' Group) 4 or NESMA (Netherlands Software Metrieken Gebruikers Associatie). After the deletion of cases with missing values, the dataset was restricted to 759 projects.

Table 3
Variables of the dataset.

Variable	Scale	Description
Duration	Ratio	Total elapsed time for the project in calendar months
Ufp	Ratio	Application size in unadjusted function points
DevType	Nominal	Development type
DevPlat	Nominal	Development platform
LangType	Nominal	Language type used for the project

Summary statistics of the ratio-scaled and nominal variables are presented in Tables 4 and 5, respectively. The mean value of duration is 8.75 months with 6.64 standard deviation, whereas the median (7 months) is quite different from the mean, indicating the presence of outliers and highly-skewed data (Table 4). This is also true for the ratio-scaled predictor *Ufp* that presents a large divergence between the mean and the median (425 and 234 respectively) with very high standard deviation (804). Table 5 provides the mean duration and the number of projects in each value (level) of all the nominal variables.

6.2. Validation of estimation by analogy

In this section, we analytically present the way that REC analysis can guide the model specification of EbA.

The EbA methodology is a procedure based on finding few neighbors (or analogies) of the project under estimation among completed projects of the dataset. As the methodology is free of assumptions, we decided to use the initial variables of Table 3. The new project is first characterized with attributes common to the ones of the historical dataset and then a distance metric is used to calculate the distance of the new project from all the others. The dataset contains mixed-type variables so we use a dissimilarity coefficient suggested by Kaufman and Rousseeuw (1990) that

Table 4
Summary statistics for the ratio-scaled variables.

Variable	Mean	Median	St. Dev.	Min.	Max.
Duration	8.75	7.00	6.64	0.20	48.00
Ufp	425	234	804	4	13580

Table 5
Summary statistics for the nominal variables.

Variable	Levels	Mean duration	No. of projects
DevType	Enhancement	7.67	454
	New development	10.01	281
	Re-development	14.41	24
DevPlat	MF	8.72	456
	MR	9.51	91
	Multi	10.79	77
	PC	7.15	135
LangType	2GL	12.45	6
	3GL	9.24	482
	4GL	8.11	200
	ApG	6.88	71

takes into account this generic nature of data. The duration values of the analogue projects are then combined in order to obtain an estimation of the unknown duration. The statistic chosen for this combination is the arithmetic mean since this choice essentially makes EbA to coincide with the k-nearest neighbor non-parametric regression (Mittas et al., 2008).

6.2.1. Choice of the best parameters

An important design issue that has to be addressed is the number of analogies that will be used for the estimation. The critical value of the nearest neighbors has been the subject of debate (Angelis and Stamelos, 2000; Shepperd and Schofield, 1997) and there is no rule of thumb for the determination of the best parameter without empirical experimentation. In order to select a sort of optimal number of analogies, we used as criterion the MdAE and through the leave-one-out cross-validation procedure we evaluated the MdAE for a range of possible neighbors ($k = 1$ to $k = 10$ in our example). The number that minimizes MdAE gives the best model.

The experimentation on the specific dataset showed that the best MdAE value (=3) was obtained by $k = 1$ and $k = 7$ analogies. Based only on this value, someone would believe that both models (i.e. with $k = 1$ and $k = 7$ analogies) have similar prediction performances and therefore that there is no essential difference in choosing one of them against the other for future predictions. As we have already mentioned in Section 1, the policy of determining the "most accurate" model based on a single indicator constitutes a common misconception in SCE that can lead to unstable and erroneous decision-making.

A more appropriate approach should consider information from the whole distribution of AEs and not just a single indicator, since a value computed from a sample contains significant variability. The utilization of REC can be an aid towards this direction providing a valuable calibration tool that can reinforce the knowledge of the project manager for this specific parameter decision.

The AE REC curves for the two competitive EbA models (with $k = 1$ and $k = 7$ analogies) are presented in Fig. 8. Despite the fact that both models have similar prediction performance for small values of AE (smaller than 2.5), the accuracy of the model with $k = 7$ becomes higher for higher AE values since its curve is always on the top and climbs more rapidly. Hence, there is evidence that the $k = 7$ analogies model is generally better than the corresponding model with $k = 1$ analogy. In order to support this evidence and statistically signify the findings obtained by the visual inspection of the REC curves, we decide to perform the non-parametric Wilcoxon sign rank test for matched pairs. Indeed, the test shows that there is a statistically significant difference between the two comparative models ($p = 0.000 < 0.05$) and confirms that the EbA model with seven analogies has the best prediction performance.

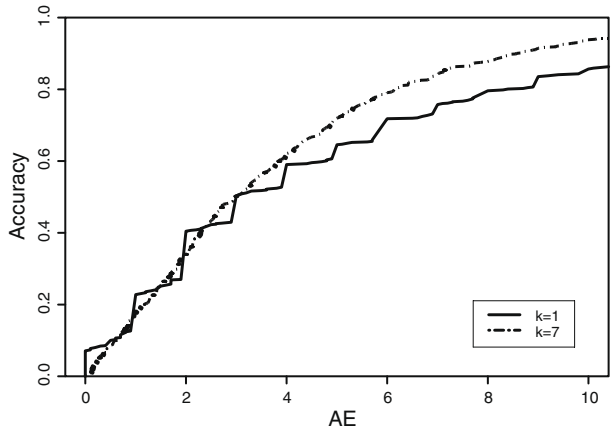

Fig. 8. AE REC curves for different number of neighbors (one and seven).

6.2.2. Applicability of estimation by analogy to ISBSG dataset

After the selection of the best parameter for the combination of the nearest projects, a manager has to decide, whether EbA with $k = 7$ analogies is an appropriate estimation technique for the ISBSG dataset. In Fig. 9a, we can see that the EbA model, with $k = 7$ analogies clearly, dominates the Mean model over a wide range of possible error values. In addition, we also evaluate the curve for EbA model with $k = 1$ analogy in order to better present the utilization of the Mean model for the assessment of the appropriateness of a model to a certain dataset. The curve for EbA model with $k = 1$ analogy (Fig. 9b) dominates the Mean model for small values of error tolerance (smaller than 50%) but there is also a high percentage of projects that present worse prediction performance compared with those of the Mean model. This fact is also depicted from the shape of the EbA curve since it is always below the corresponding curve of the Mean model and climbs slowly to the maximum value of 1 of the accuracy axis. The Wilcoxon tests also verify these results since there is a statistically significant difference only for the case of the EbA model with $k = 7$ analogies and the Mean model comparison.Summarizing, in this section we saw how standard REC curves can be utilized for the calibration of the EbA method. Furthermore, we assess the suitability of the model, with specific choices of parameters, to the ISBSG dataset. The findings showed that the choice of one ($k = 1$) and seven ($k = 7$) analogies gave identical error indicators (MdAEs) but the distributions of AEs produced quite different shapes. Moreover, the comparisons of the abovementioned models with the Mean model reveal that only EbA with seven analogies seems to be an appropriate calibration choice for the prediction of a forth-coming project.

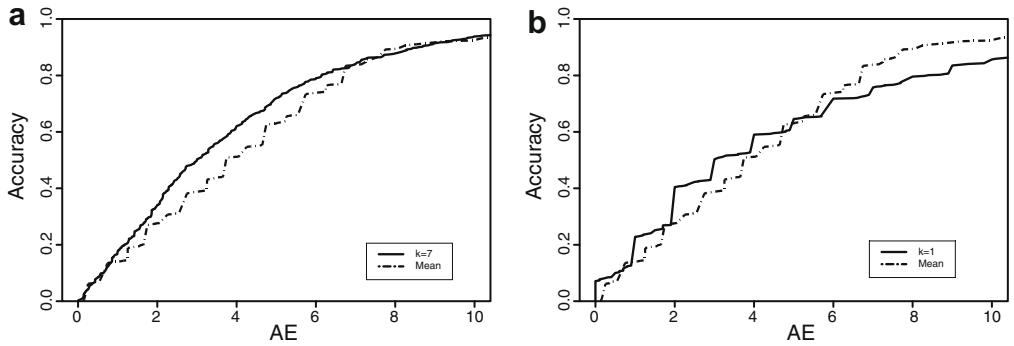

Fig. 9. AE REC curves for applicability of models with (a) seven analogies and (b) one analogy.

Obviously, the graphical calibration can be extended to other estimation methods, especially those which require from the practitioner to make decisions on some initialization parameters. The benefit again is that we can work here with the entire distribution of errors and not just a single indicator and decide whether a specific parameter choice is suitable for a certain dataset.

6.2.3. Exploratory analysis and identification of factors affecting the predictive power

At the second level of the graphical validation of EbA model (Fig. 6), a project manager would like to gain a better insight into the prediction performance of the model. Hence, after the selection of the best parameters for the EbA model, REC analysis constitutes a multi-tool that enables practitioners to acquire significant knowledge concerning the identification of factors affecting the predictive error. For such an illustration, suppose that the manager would like to investigate the distribution of the prediction errors obtained by the EbA model with $k = 7$ analogies for the three levels of the variable *DevType*.

Errors measured by AE, MRE and residual are plotted in Figs. 10–12 respectively, one curve for each level of *DevType*. As the AE REC curves (Fig. 10) for Enhancement and New Development projects are always above the corresponding curve of Re-development projects, we can infer that the EbA model shows a tendency to produce poor predictions especially for Re-development projects. Furthermore, 12.5% of Re-development projects seem to have extremely large AE values since the corresponding REC curve does not reach 1 before the AE tolerance becomes extremely high (higher than 15).

These visual observations can be confirmed by other statistical tools. In Table 6, we evaluated the MAE and MdAE measures for each level of *DevType* and we further performed the non-parametric Mann–Whitney statistical test for independent samples. The findings of the measures and the *p*-values of the test (less than 0.05) statistically signify the results derived from the visual inspection of AE REC curves.

The MRE REC curves (Fig. 11) also show that the prediction errors of Re-development projects have the highest spread compared to those of Enhancement and New Development projects for MRE values smaller than 100% but as the error tolerance increases, all projects seem to have similar performances. Indeed, all pair-wise comparisons do not indicate a statistically significant difference between these levels (all *p*-values are greater than 0.05).

In order to investigate the bias of errors due to *DevType*, we constructed the REC curves representing the residuals for each level of the nominal variable (Fig. 12). This graph has interesting properties, since the errors are allowed to have negative values.

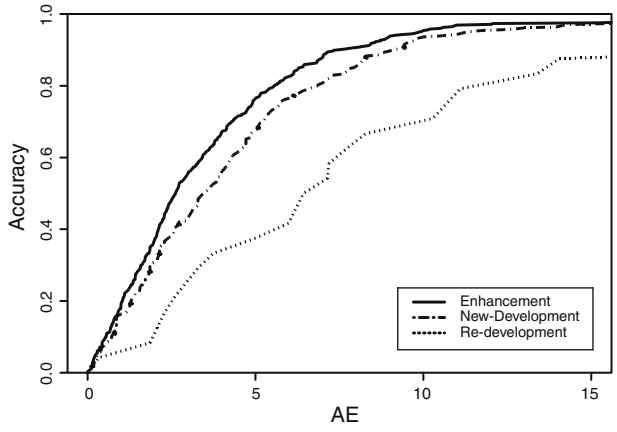

Fig. 10. AE REC curves for DevType.

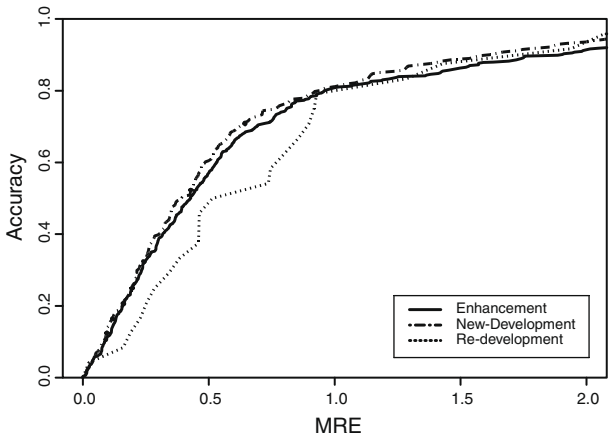

Fig. 11. MRE REC curves for DevType.

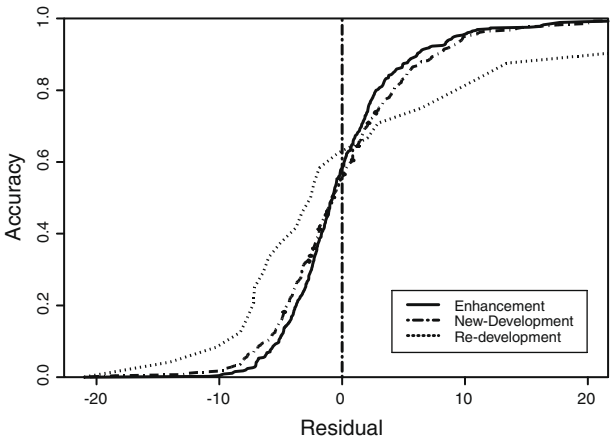

Fig. 12. Residual REC curves for DevType.

Table 6
Prediction measures for DevType.

Measure	Enhancement	New development	Re-development
MAE	3.72	4.52	9.13
MdAE	2.63	3.57	6.79
Mean residual	0.11	0.22	1.95
Median residual	−0.79	−0.71	−2.43
MMRE (%)	77.58	72.79	97.47
MdMRE (%)	42.85	39.29	62.58

Obviously, negative residuals show overestimation while the positive show underestimation.

The residual REC curve is therefore a visual tool for inspecting the balance between overestimation and underestimation. If the balance is distorted, there is a bias in the model's predictions, i.e. a systematic trend towards larger or smaller values than the actual estimations. This balance is detected easily by the median of the residuals (i.e. the error tolerance corresponding to accuracy 0.5), which should be close to zero, and the symmetry of the curve. That is why in residual REC plots we draw a vertical dashed reference line through the critical value of 0. Generally, we can claim that the model favors underestimation when the accuracy value at error tolerance 0 is smaller than 0.50 (or the median is positive) while it favors overestimation when this value is higher than 0.50 (or the median is negative). It is also worth mentioning that if the residual REC curve is symmetrical and S-shaped, we can

assess that the residuals are normally distributed, since the CDF of any normal distribution has exactly this shape.

In our example, the residual REC curves (Fig. 12) show that the prediction of Re-development projects is generally the worst, but also that there is a bias, i.e. the model tends to overestimate the specific projects. Note that the Median residual for Re-development projects is −2.43 (Table 6), a value negative and smaller than those of the other two levels. On the other hand, the EbA model has similar predictive performance for Enhancement and New Development projects. The curves are more symmetrical and the medians are closer to 0 (Table 6). It is interesting to note here that all these differences in the behavior of the LS model were not detected as significant by the Mann–Whitney tests (all p-values are greater than 0.05).

Summarizing the results, we presented a potential usage of REC curves regarding the identification of factors affecting the predictive error in terms of accuracy, bias and spread. We emphasize the fact that these errors are prediction errors and not fitting errors, i.e. they were obtained when the model was trying to predict unknown values and not to fit known ones. The analysis conducted with the EbA model and the *DevType* variable, showed clearly the problem of poor prediction of a certain class of projects. A similar exploratory analysis can be performed for all the nominal variables contained in a dataset, so as to lead us to very useful conclusions regarding the predictive power of a model and certainly to further improvements.

6.3. Comparison of estimation by analogy and regression models

In the previous section, we analytically present the way that REC curves can guide the modeling process of an estimation technique. Besides this purpose of the proposed analysis, we have already mentioned (Section 5.2) that the main scope of REC curves is the comparison of alternative prediction models.

Let us assume that a manager calibrates the EbA method through the analysis of the previous section and also aims to fit a LS model in order to decide which model appears to have the "best" performance on the ISBSG dataset. Before the comparison of the alternative models, we have to describe in detail the construction of LS model, so as to make our analysis meaningful.

LS methodology estimates the parameters of a linear model through minimization of the overall sum of squared errors. As there are several assumptions that need to be accounted for the fitting of the LS model, it is important to make some tests and probably transformations in order to ensure the reliability of the model. Initially, graphical means for testing normality like histograms and Q–Q plots showed that the two ratio-scaled variables (*Duration* and *Ufp*) were not normally distributed. Hence, we transformed the initial variables to the natural logarithmic scale in order to achieve better approximation of the normal distribution for both of them.

The dataset also contains three nominal variables that have to be replaced by binary (or dummy) variables in order to be included in the LS model. The final set of variables used for the construction of the LS model is presented in Table 7. Note that for a nominal variable with c possible values, we always need $c-1$ dummy variables.

A Stepwise Regression procedure was then applied to include in the model only the variables having a significant impact on the response variable. In a multiple linear regression model, *adjusted* r^2 measures the proportion of the variation in the dependent variable accounted for by the explanatory variables. It is considered as more accurate goodness-of-fit measure than the simple coefficient of determination r^2 since it takes into account the degrees of freedom associated with the sums of the squares. The adjusted r^2 was 32.6% and the equation describing the best fitting model was:

Table 7
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 the other types as 0
DevType2	Dummy variable where "New Development" type is coded as 1 and all the other types as 0
DevPlat1	Dummy variable where "MF" platform is coded as 1 and all the other platforms as 0
DevPlat2	Dummy variable where "MR" platform is coded as 1 and all the other platforms as 0
DevPlat3	Dummy variable where "Multi" platform is coded as 1 and all the other platforms as 0
LangType1	Dummy variable where "2GL"language is coded as 1 and all the other languages as 0
LangType2	Dummy variable where "3GL"language is coded as 1 and all the other languages as 0
LangType3	Dummy variable where "4GL"language is coded as 1 and all the other languages as 0

Table 8
Overall prediction measures.

Measure	LS	EbA	Mean
AE			
MAE	3.9	4.19	4.81
AOC	3.88	4.17	4.75
MdAE	2.53	3	3.75
Residual			
Mean residual	1.39	0.21	0
AOC	1.37	0.19	−0.1
Median residual	0.15	−0.79	−1.75
MRE			
MMRE (%)	60.11	76.44	110.67
AOC (%)	59.16	75.25	105.92
MdMRE (%)	39.4	42.86	45.75

$$\begin{aligned}\ln duration = {} & -0.225 + 0.331 \times \ln ufp - 0.180 \times DevType1 \\ & + 0.311 \times DevPlat1 + 0.279 \times DevPlat2 \\ & + 0.480 \times DevPlat3 + 0.795 \times LangType1 \\ & + 0.283 \times LangType2 \end{aligned} \quad (4)$$

6.3.1. Overall comparison of models

At the first level of the comparison procedure, our goal is to examine the overall prediction performances of the alternative models (Fig. 7). The general results concerning the accuracy, bias and spread of error for the two comparative models are presented in Table 8. Apart from the evaluation of LS and EbA, we also employ the Mean model in order to assess the suitability of the prediction techniques on ISBSG dataset (Section 3.4).

It is obvious that LS and EbA outperform in accuracy the Mean model in terms of MAE and MdAE. This means in practice that both models are suitable predictors for this specific dataset. Regarding the comparison between LS and EbA, LS seems to be the most accurate technique since it presents lower mean and median values than AE. Having in mind the main scope of REC analysis, we construct the REC curves in order to graphically examine the whole distribution of absolute errors (Fig. 13).

The REC curve corresponding to AE obtained by LS is always above the curves of EbA and Mean model, whereas the EbA curve dominates the Mean model curve. Thus, the evaluation of REC curves verifies the indicators of Table 8 and shows that LS is the "best" technique in terms of accuracy. In order to verify the

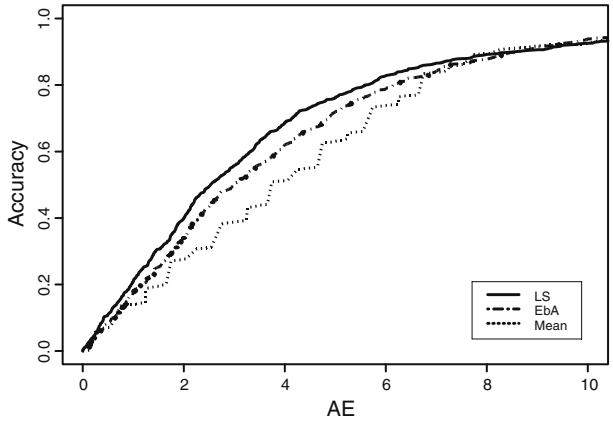

Fig. 13. AE REC curves for the three comparative models.

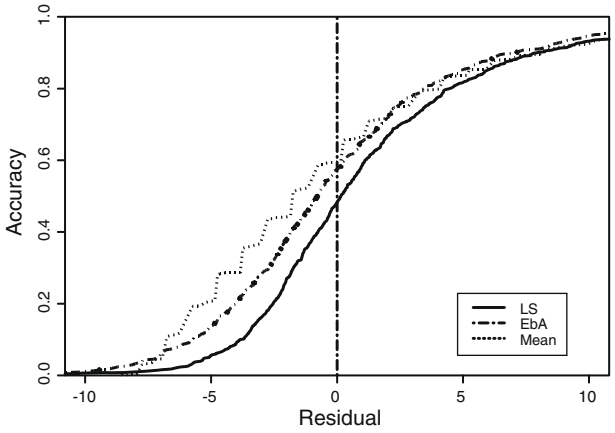

Fig. 15. Residual REC curves for the three comparative models.

effectiveness of the REC curves to detect the differences between the comparative models, we also make use of the Wilcoxon sign rank test for matched pairs to detect significant differences. All pair-wise tests have *p*-values smaller than 0.05 revealing that the differences observed in Fig. 13 are also statistically significant.

The MMRE and MdMRE as measures of the residuals' spread also show clearly better results for LS. The closer examination of their distributions in Fig. 14 shows that the three comparative models have similar performances for small values of MRE (smaller than 30%) but as the error tolerance increases, the LS model obviously dominates. Moreover, the accuracy of EbA is also better than the corresponding accuracy of the Mean model. Another interesting issue coming up from the visual inspection of their performances is that the top of the curves is extremely flat and does not reach 1 until the error tolerance becomes high, indicating the presence of outliers. Again, the Wilcoxon tests statistically signify the aforementioned findings obtained by the REC curves (all *p*-values are less than 0.05).

The behavior of the comparative models in terms of bias deserves some investigation since the mean and the median of residuals, presented in Table 8, appear quite contradictory. The mean value of the residuals of the Mean model is nearly 0, whereas LS seems to have the worst performance in terms of bias. On the other hand, the results should be carefully examined since the distributions of residuals seem to be non-normal and heavily skewed. As we can figure out from Fig. 15, the shapes of the distributions are not symmetric and the median value should be considered a more robust measure of bias. Indeed, the non-parametric Kolmogorov–Smirnov test for normality showed that the residuals of all three comparative models are not normally distributed. Moreover, as we can see from Fig. 15, and verify from Table 8, the median of the LS residuals (i.e. the error tolerance at 0.50 accuracy) is positive and quite near 0. On the other hand, it is clear that both EbA and the Mean model tend to overestimate the cost since their REC curves achieve 0.50 accuracy at the negative values of error tolerance (negative median). Concluding, the REC analysis of residuals shows the skewness of the distributions and in terms of median, LS appears to produce the least bias. Finally, the pair-wise comparisons show statistically significant differences among all distributions (all *p*-values are less than 0.05).

As we have already mentioned in Section 3.2, the AOC for each of the comparative models is an estimate of the expected mean error and for this reason, we indicatively present the AOC estimates for all models in Table 8. These were computed using the trapezoid formula for numerical integration as it is known that the areas under and above a curve are directly computed from the integral of the curve. It is worth noting that AOC is known theoretically as a biased estimator of the mean of error (measured in any way), since it always underestimates the actual mean. However, as we can see from Table 8, its values are very close to the corresponding mean. Furthermore, theory proves that as the amount of data tends to infinity (practically for very large datasets), AOC convergences to the mean of error.

Summarizing our findings, we can conclude that the REC curves for all the expressions of error we studied show for this specific dataset that LS outperforms and is the most plausible choice for predicting the duration of a forth-coming project. The most important here is that the conclusions obtained by a simple visual comparison can be verified by the utilization of formal statistical comparisons.

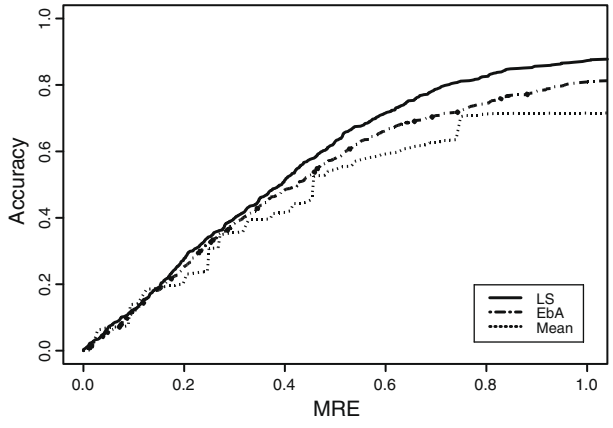

Fig. 14. MRE REC curves for the three comparative models.

6.3.2. Partial comparison of models

In the previous section, we presented in detail how REC curves can lead a project manager to select the "best" model. Although this is one of the most crucial issue that has to be addressed for the best management of a new software project, the aforementioned analysis concerns the overall prediction performances of comparative models.

As we have already pointed out in Section 5.2, there are certain cases in which a project manager would like to acquire information about the performances of comparative models on specific ranges of interest. This specialized analysis (partial comparison of models) can be proved beneficial in the development process and can be

conducted through the construction of partial REC curves that have extensively been described in Section 3.3.

6.3.2.1. Partial comparison on specific ranges of duration. Let us suppose that someone has to investigate how the prediction models perform separately for small and for large values of the dependent variable (*Duration*). Following the algorithm for the construction of partial REC curves (see Section 3.3), we have to define the range of interest (step 3 of the main algorithm). For this reason, we calculated the quartiles of the empirical distribution of the dependent variable and we constructed the partial REC curves for each one of the following subsets:

- Projects that have small actual duration, that is actual duration $\leqslant Q_1$,
- Projects that have large actual duration that is actual duration $\geqslant Q_3$, where Q_1 and Q_3 are the first and third quartiles, respectively.

Due to the fact that there is more than a project with duration value equal to Q_1, the size of the subsets is not fairly equal. The first subset with small duration projects contains 213 (28.06%) cases and the second (high duration projects) contains 190 (25%) cases. In Table 9, we present the accuracy measures computed by the errors from the first subset with small duration projects from which we can clearly observe that LS model has the "best" performance in terms of accuracy, bias and spread.

The partial REC curves concerning the three aspects of the prediction performance are presented in Figs. 16–18. Regarding the accuracy of the comparative models, the superiority of LS model is unambiguous (Fig. 16), and the Wilcoxon tests also signify the results since all the pair-wise comparisons have *p*-values smaller than 0.05.

Table 9
Prediction measures for small and high duration projects.

Measure	LS		EbA		Mean	
Duration	Small	High	Small	High	Small	High
MAE	2.52	8.51	3.66	7.39	6.01	8.94
MdAE	2.12	6.69	3.41	5.54	5.75	6.25
Mean residual	−2.44	8.00	−3.58	6.88	−6.01	8.94
Median residual	−2.13	6.66	−3.41	5.48	−5.75	6.25
MMRE (%)	123.83	44.41	173.10	37.67	303.80	45.29
MdMRE (%)	83.78	44.86	128.57	36.80	191.50	41.70

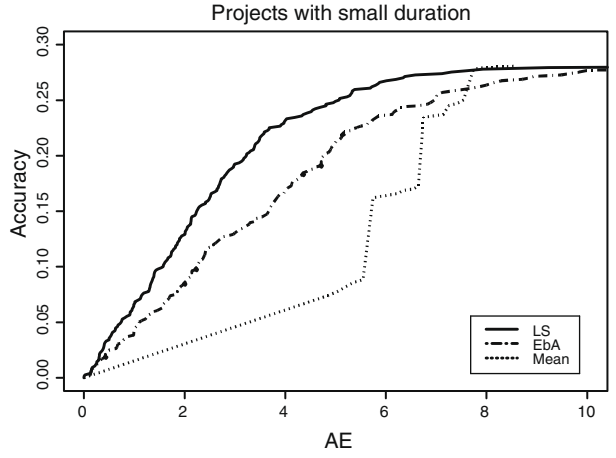

Fig. 16. Partial AE REC curves for the small durations.

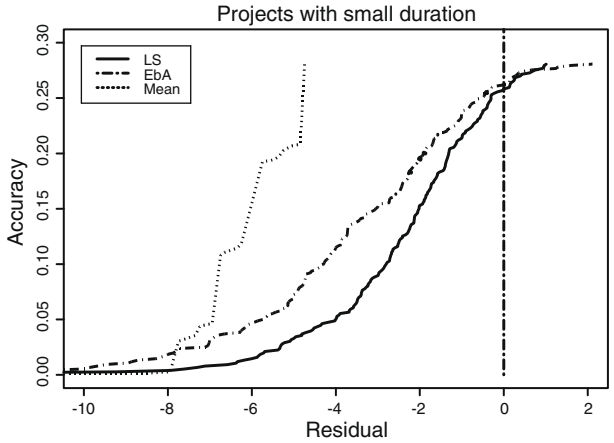

Fig. 17. Partial residual REC curves for the small durations.

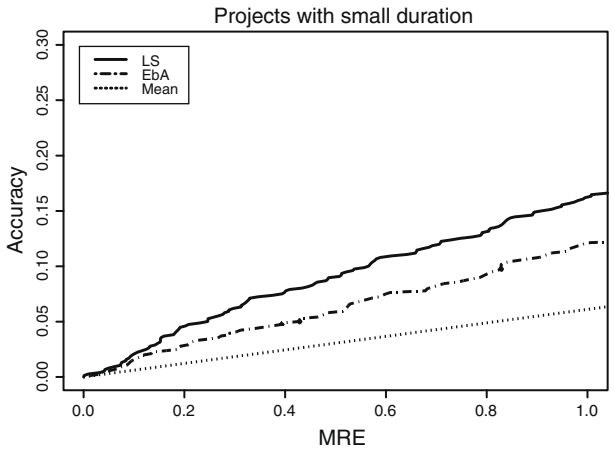

Fig. 18. Partial MRE REC curves for the small durations.

As far as the bias is concerned, all models are prone to overestimation since the relative position of their residual partial REC curves are at the left of 0 (Fig. 17). Furthermore, the shape of the LS and EbA models seem to be more symmetric indicating closeness to a normal distribution. The Kolmogorov–Smirnov test for both models confirms this conjecture.

This is also the case for the MRE accuracy measures indicating the superiority of LS model. However, we have to point out that all models appear large spread in their predictions since their partial REC curves climb slowly and not reach 0.28 (maximum accuracy value) until the error tolerance becomes very high (Fig. 18). So, we can infer the presence of a few outliers that affect their performances.

The partial REC analysis is also conducted for the second subset of large duration projects (Figs. 19–21). The measures of Table 9 show that the findings are different for projects with high duration. More precisely, EbA has the "best" prediction performance, whereas the comparison of LS and Mean models reveals a similar behavior in terms of accuracy, bias and spread.

The partial AE REC curves (Fig. 19) graphically reinforce the inference obtained by the examination of Table 9. Moreover, Wilcoxon tests statistically signify the superiority of EbA compared to both LS and Mean models, whereas it is found that there is no difference between LS and the Mean model ($p = 0.196$).

Contrary to the first subset, all models favor underestimation since their partial residual REC curves are at the right of 0, whereas their shapes are not symmetric (Fig. 20). Indeed, the tests for

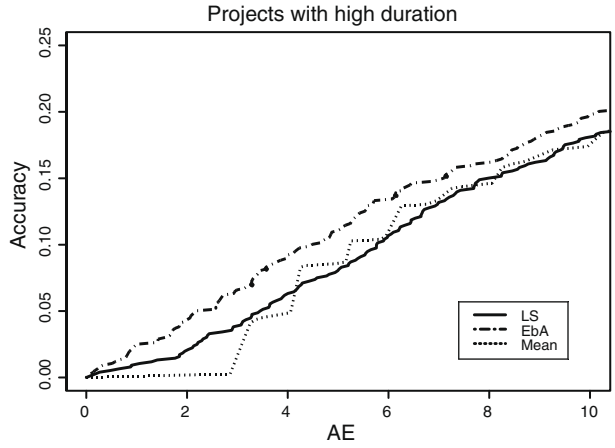

Fig. 19. Partial AE REC curves for the high durations.

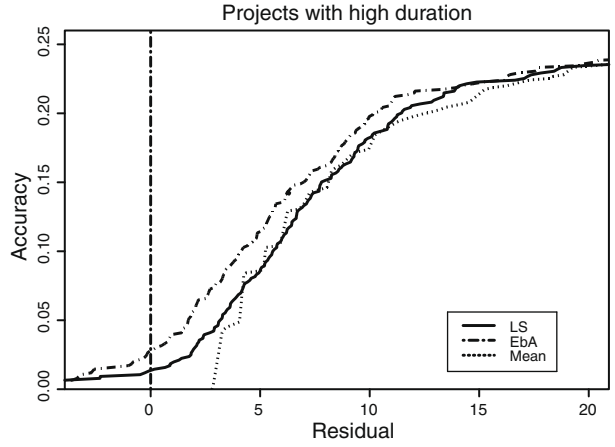

Fig. 20. Partial residual REC curves for the high durations.

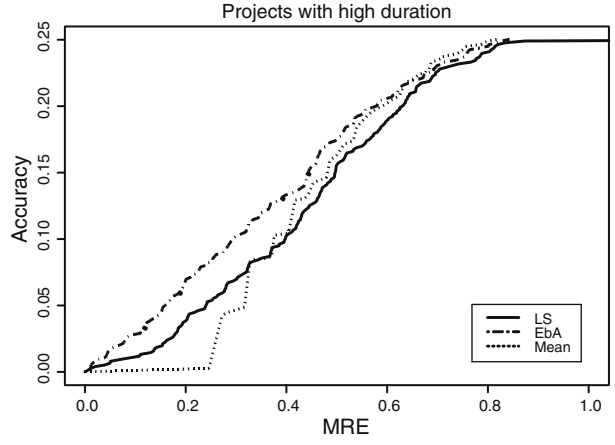

Fig. 21. Partial MRE REC curves for the high durations.

normality for the three distributions show that none of them is normal (all p-values are less than 0.05). Again, EbA presents the most unbiased predictions and statistically outperforms both LS and the Mean model. Regarding the comparison between LS and the Mean model, no inference can be derived for the superiority of one method against the other ($p = 0.056 > 0.05$).

Finally, the partial MRE REC curves show that EbA has the smallest residual spread and statistically dominates the other models (Fig. 21). Again, no difference can be inferred from both the inspection of the curves and the statistical comparison ($p = 0.395 > 0.05$) between the LS and the Mean model.

Summarizing the findings of this section, we presented a potential usage of partial REC curves concerning the evaluation of the prediction power of alternative models over specific ranges of the dependent variable. The findings of such an analysis can lead a project manager to acquire meaningful information for the performances of comparative techniques that can guide the selection of alternative models to different ranges of interest.

In our dataset, we found that LS significantly outperforms the comparative models for small duration projects in terms of all the accuracy measures and also that EbA seems to be an alternative choice since it outperforms the Mean model. On the contrary, EbA appears to have the "best" performance for high duration projects, whereas LS does not significantly outperform the predictions derived from the Mean model, so the suitability of LS for high duration projects is doubtful for the specific dataset. Moreover, the analysis of the residual REC curves showed that all models are prone to overestimation and underestimation for small and high duration projects, respectively. Finally, the shapes of the residual curves reveals normally distributed residuals for the LS and EbA models but only for the first subset and not for the second. All these results were also verified from the formal statistical procedures.

It is important to emphasize here that although this type of analysis is not common in comparative studies, it is essential and very useful since it can uncover sources of systematic error that need to be considered in the choice of the suitable model.

Although the abovementioned procedure analytically described in Section 3.3 can be used in order to investigate the performances of comparative models in certain ranges of the response variable, it seems that a more realistic approach is to construct the partial REC curves over the cases that fall in a range of actual size values. Indeed, the size is a main cost predictor and is usually known in advance, so it can be alternatively utilized for this specific comparison purposes. Next, we indicatively illustrate how the partial REC curves can be drawn only for projects with small size.

Initially, we have to define the range of interest (i.e. projects with small size) by computing the first quartile of *Ufp* and evaluate partial REC curves for these 193 (25.43%) small-sized projects with $Ufp \leqslant Q_1$. Note that Stensrud et al. (2002) follow a very similar approach in order to investigate the relationship between the MRE and project size. Fig. 19 shows the three partial AE REC curves for each of the comparative models. It is clear that the patterns of all curves are very similar to those obtained from the first subset of small duration projects (Fig. 22), verifying once again the superiority of LS model.

6.3.2.2. Partial comparison on specific ranges of error. In the previous section, we illustrated the application of partial REC curves to the problem of comparing the performances of comparative models on specific ranges of the dependent variable.

An alternative analysis that can be performed through the partial REC curves aims to compare how a specific range of errors is distributed across the whole domain of the response variable. By the examination of these particular curves, a project manager can focus on specific magnitudes of error and therefore gain additional information for a better understanding of the benefits and drawbacks of comparative prediction models.

As Torgo (2005) propose, the abovementioned analysis can be attained through the construction of a partial CDF of the dependent variable only for those projects for which the prediction error falls within a range of interest.

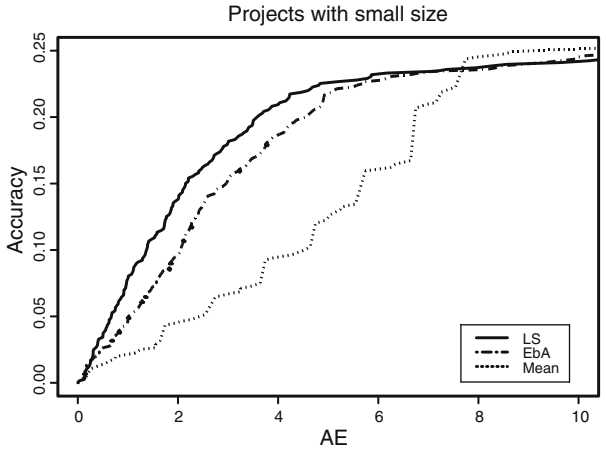

Fig. 22. Partial AE REC curves for the small sizes.

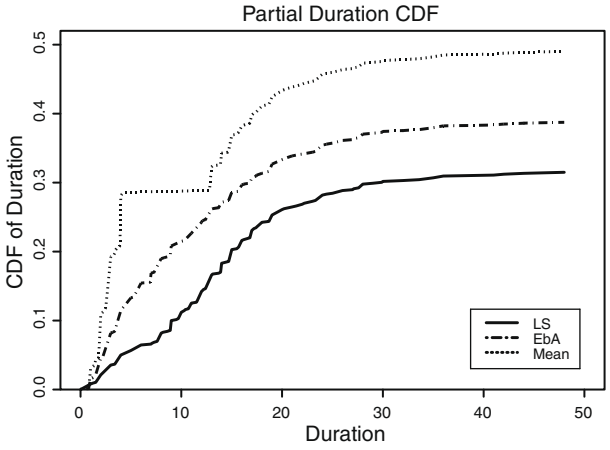

Fig. 24. Partial CDF of the Duration for high AEs.

For this purpose, we have to define the ranges that we would like to examine the behavior of our prediction models. The analysis is indicatively performed for two specific ranges of AE values:

- Projects with AE < 4,
- Projects with AE $\geqslant$ 4, where the discriminating value (4) of the two subsets is selected to be a value very close to the MAEs of the two main comparative models (LS and EbA). We have to make clear that this choice was made only for illustration purposes and that anyone can select any error range according to the practical needs.

Fig. 23 depicts the partial CDFs of Duration for the projects with AE < 4, that is relatively small AEs. The *x*-axis represents the actual values whereas the *y*-axis the CDF of the dependent variable. We can see that LS achieves more frequently such values of errors, which is reflected in a higher final value of the CDF. More precisely, 68.5% of the projects achieve an AE smaller than 4 for LS, whereas the percentages for EbA and Mean model are almost 61.26% and 50.99%, respectively. Moreover, these values of AE are concentrated for both LS and EbA models in the duration range [0, 12] indicated by the flat shape of the curves for duration values higher than 12. This means that small AEs are more likely to be achieved in small values of the dependent variable. On the other hand, the Mean model concentrates small values of AEs in the range [4.9, 12] showing that this model cannot predict with high accuracy projects with small actual durations.

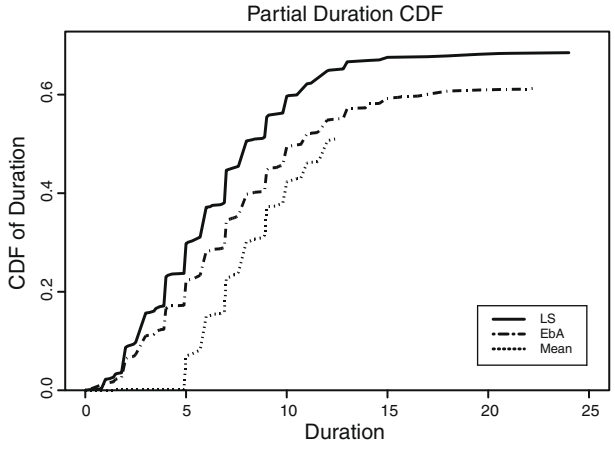

Fig. 23. Partial CDF of the Duration for small AEs.

On the contrary, Fig. 24 shows that Mean curve dominates both LS and EbA for projects with AE $\geqslant$ 4 meaning that it achieves more frequently such high values of error. In addition, the Mean curve seems to be flat for durations that belong in the range [4.3, 12.8], meaning that there is no project with AE $\geqslant$ 4 in this specific range of the response variable. Furthermore, the range of the duration for all models begins from a duration value very close to 0, whereas the maximum value of the *x*-axis is 48 that is the maximum dependent value for the whole dataset. This means that high values of AEs are likely to be found over the whole range of the dependent variable.

7. Conclusions and discussion

In this paper, we deal with an important research question in SCE, concerning the identification and selection of the "best" prediction model for a specific dataset. A various number of studies that appeared so far in the related literature have contributed towards this direction, since the decision of using a certain technique against a comparative one can play a critical role in the well-balanced management of the development process.

In this study we introduced the regression error characteristics analysis that is a class of visualization techniques for the evaluation of the predictive power and comparison of different models. The basis of the abovementioned analysis is the construction of standard and partial regression error characteristics curves that comprise a visualization tool analogous to the well-established receiver operating characteristic curves used in comparisons of classification models. A REC curve is an estimation of the cumulative distribution function of the prediction error in any way it is measured, whereas a Partial REC curve expands the advantages of REC curves by taking into account the actual values of the dependent cost variable.

The benefits of using such an analysis can be proved invaluable for a project manager, since REC curves are very readable and easily interpretable, especially for non-experts, facilitating the strategy that has to be followed for a better management of a new software project. As the usual strategy of giving advance to a prediction model based solely on a single accuracy measure can lead to erroneous decision-making due to the existence of few outliers, a decision-maker would like to easily obtain significant knowledge for the performances of different models.

The most important advantage provided by the REC analysis, is that all common accuracy measures, like MMRE, MdMRE and Pred(p) are directly represented by geometrical characteristics and properties of the REC curves. Furthermore, as shown in all of

our applications, the statistical tests comparing the whole samples of errors, confirm the visual results, in the sense that each time the difference between two prediction error samples is significant, this is clearly shown by the REC curves. Another interesting advantage is that the analysis is able to focus on special subsets of the prediction error. This is very useful since it can be revealed that a prediction model is best only for a special type of projects (for example the small ones) while for other types is not so efficient. Finally, the REC curves provide valuable information about the distributional characteristics of the prediction errors, for example about the symmetry of the distribution and the existence of outliers. All this information is provided in a straightforward and easy-to-construct manner.

As the various comparative studies concerning the selection of the "best" model give in general contradictory results, the goal of this paper is to further extend the research on this area. Although in our applications we used two of the most known methods in SCE (regression analysis or estimation by analogy), we emphasize that it is not our intention to determine the superiority of either prediction methods, but rather to propose an additional tool that contributes to the systematic research of the performances of any kind of prediction technique. Furthermore, we introduce a formal graphical framework that is able to cover different aspects of the comparison procedure such as the calibration of the prediction methodology, the identification of factors that affect the error obtained by a model, the investigation of errors on certain ranges of the actual cost variable and the examination of how a certain range of errors is distributed across the domain of the response variable. Using the most recently released version of the International Software Benchmarking Standards Group (ISBSG) dataset, we presented different examples attempting to cover the aforementioned views of the potential utilization of REC analysis. The error functions selected in this study measure three different aspects of predictive performance, namely the accuracy, the bias and the spread.

Some interesting issues that came up from our study deserve further research. At first, it is essential to investigate various software cost datasets and study their behavior in the comparisons of several other prediction techniques (for example Robust Regression, Classification and Regression Trees, Bayesian Networks, etc.), simultaneously, in order to contribute to the systematic identification of differences between the aforementioned methods.

Moreover, we currently investigate the utilization of bootstrap simulation procedures in order to construct confidence intervals for the REC curves so as to test graphically the significance of the difference among several prediction techniques. A much more appealing research topic is the ensemble of predictions from different models by the investigation of the best performances derived by the REC analysis.

References

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

Bi, J., Bennet, K.P., 2003. Regression error characteristics curves. In: Proceedings of the AIII 20th International Conference on Machine Learning (ICML'03), August, pp. 43–50.

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

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

Egan, J.P., 1975. Signal Detection Theory and ROC Analysis. Series in Cognition and Perception. Academic Press.

Foss, T., Stensrud, E., Kitchenham, B., Myrtveit, I., 2003. A simulation study of the model evaluation criterion MMRE. IEEE Transactions on Software Engineering 29 (11), 985–995.

ISBSG Dataset 10, 2007. <http://www.isbsg.org>.

Jeffery, R., Ruhe, M., Wieczorek, I., 2001. Using public domain metrics to estimate software development effort. In: Proceedings of the IEEE 7th International Software Metrics Symposium (METRICS 2001), April, pp. 16–27.

Jorgensen, M., Shepperd, M., 2007. A systematic review of software development cost estimation studies. IEEE Transactions on Software Engineering 33 (1), 33–53.

Kaufman, L., Rousseeuw, P.J., 1990. Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley, New York.

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

Kitchenham, B., Mendes, E., 2009. Why comparative effort prediction studies may be invalid. In: Proceedings of the ACM 5th International Conference on Predictor Models in Software Engineering, May.

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

Mair, C., Shepperd, M., 2005. The consistency of empirical comparisons of regression and analogy-based software project cost prediction. In: Proceedings of the International Symposium on Empirical Software Engineering (ISESE'05), November, pp. 509–518.

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

Mendes, E., Di Martino, S., Ferrucci, F., Gravino, C., 2008. Cross-company vs. single-company web effort models using the Tukutuku database: an extended study. Journal of Systems and Software 81 (5), 673–690.

Mittas, N., Angelis, L., 2008a. Comparing cost prediction models by resampling techniques. Journal of Systems and Software 81 (5), 616–632.

Mittas, N., Angelis, L., 2008b. Comparing software cost prediction models by a visualization tool. In: Proceedings of the IEEE 34th Euromicro Conference on Software Engineering and Advanced Applications (SEAA'08), September, pp. 433–440.

Mittas, N., Angelis, L., 2008c. Partial regression error characteristic curves for the comparison of software cost prediction models. In: Proceedings of the Artificial Intelligence Techniques in Software Engineering (AISEW'08), July, pp. 6–10.

Mittas, N., Athanasiades, M., Angelis, L., 2008. Improving analogy-based software cost estimation by a resampling method. Information and Software Technology 50 (3), 221–230.

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

Shepperd, M., Schofield, C., 1997. Estimating software project effort using analogies. IEEE Transactions on Software Engineering 23 (11), 736–743.

Stensrud, E., Myrveit, I., 1998. Human performance estimating with analogy and regression models: An empirical validation. In: Proceedings of the IEEE 5th International Software Metrics Symposium (Metrics'98). November, pp. 205–213.

Stensrud, E., Foss, T., Kitchenham, B., Myrtveit, I., 2002. An empirical validation of the relationship between the magnitude of relative error and project size. In: Proceedings of the 8th IEEE Symposium on Software Metrics, July, pp. 3–12.

Torgo, L., 2005. Regression error characteristic surfaces. In: Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'05), August, pp. 697–702.

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