Bootstrap Confidence Intervals for Regression Error Characteristic Curves Evaluating the Prediction Error of Software Cost Estimation Models

Nikolaos Mittas, Lefteris Angelis

Department of Informatics, Aristotle University of Thessaloniki 54124, Thessaloniki GREECE, e-mail: *{nmittas, lef}@csd.auth.gr}*

Abstract

The importance of Software Cost Estimation at the early stages of the development life cycle is clearly portrayed by the utilization of several algorithmic and artificial intelligence models and methods, appeared so far in the literature. Despite the several comparison studies, there seems to be a discrepancy in choosing the best prediction technique between them. Additionally, the large variation of accuracy measures used in the comparison procedure constitutes an inhibitory factor which complicates the decision-making. In this paper, we further extend the utilization of Regression Error Characteristic analysis, a powerful visualization tool with interesting geometrical properties in order to obtain Confidence Intervals for the entire distribution of error functions. As there are certain limitations due to the small-sized and heavily skewed datasets and error functions, we utilize a simulation technique, namely the bootstrap method in order to evaluate the standard error and bias of the accuracy measures, whereas bootstrap confidence intervals are constructed for the Regression Error Characteristic curves. The tool can be applied to any cost estimation situation in order to study the behavior of comparative statistical or artificial intelligence methods and test the significance of difference between models.

1 Introduction

A crucial issue and an open problem which attracts the interest of researchers in software engineering is the ability to build accurate prediction models in order to estimate the cost of a forthcoming project. Due to this fact, a large amount of studies is towards this direction evaluating the performance of different *Software*

Cost Estimation (SCE) methods and models [1]. Although there is an obligation for a project manager to select the "best" prediction technique, there seems to be no global answer for all kinds of data. Furthermore, the wide variety of the proposed approaches that diversifies from expert judgment techniques to algorithmic and machine learning models renders the task of the selection extremely difficult. According to [2], the main reason for the contradictory results is the lack of standardization in software research methodology which leads to heterogeneous sampling, measurement and reporting techniques and the appropriateness of the prediction techniques on the available data.

The situation becomes much more complicated when we consider the divergent opinions about which accuracy measures are most appropriate in order to compare the predictions obtained by alternative models. Although a lot of accuracy indicators have been proposed in the literature and used in practice so far [3], there is a confusion of what different statistics really measure [4].

From all of the aforementioned, it is clear that the validation of prediction methods and the selection of the most appropriate model is a critical and non-trivial procedure. As we remarked in [5], a single error measure is just a statistic, i.e. a value computed from a sample (mean, median or percentage) and as such contains significant variability. Hence, when we compare models based solely on a single value we take the risk to consider as significant a difference which in fact may be not so significant. For these reasons, the determination of the "best" prediction technique has to be based on a more formal comparison procedure through inferential statistical approaches. On the other hand, in some circumstances, traditional methods might lead to erroneous inference when the dataset is considerably small and skewed or when the parametric assumptions do not hold. Thus, the utilization of resampling techniques is proposed for the selection of the best prediction technique.

In this paper, we extend our previous study [6] in which we presented the *Regression Error Characteristic* (REC) curves and the benefits from using them for the visual comparison of prediction models. Specifically, we propose a bootstrap method for the construction of *Confidence Intervals* (CIs) for REC curves, so as to test graphically the significance of the difference between two prediction techniques. The bootstrap method is the most appropriate, as the sample of errors is non-normally distributed, heavily skewed and usually of small size. By utilizing bootstrap, we illustrate how the selection of the best model can be accomplished with graphical means. Moreover, by providing bootstrap estimates, such as standard error and bias, we show how indicators of accuracy can be affected by the small software samples.

The rest of the paper is organized as follows: In Section 2, we present the bootstrap method. In Section 3, we describe the methodology followed for the construction of bootstrap REC curves. In Section 4, we present the experimental results obtained by the application of bootstrap REC curves on a real dataset. Finally, in Section 5, we conclude by discussing the results and by providing some directions for future research.

2 The bootstrap method

The comparison of prediction models is usually based on a validation procedure where various functions of errors are evaluated from the actual Y_A and the estimated Y_E cost. This results in a "point estimation" for the unknown accuracy i.e. a single value, which is computed from a particular sample coming from an practically infinite and unknown population.

Bootstrap is a simulation technique that can be used in order to extract and explore the sample distribution of a statistic [7]. We use here the most known version, the non-parametric bootstrap, which is based entirely on the empirical distribution of the dataset, without any assumption on the population. In general, the technique is to use a random sample $\mathbf{x} = (x_1, ..., x_n)$ from which we draw a large number (say *B*) of bootstrap samples by sampling with replacement in order to make statistical inference about an unknown population parameter θ (mean, median, percentage, etc.). The sample statistic $\hat{\theta}$ is a point estimator of the parameter θ (for details on the method see [5]).

The approximate distribution obtained by bootstrap can be used for computing the standard error, the bias and the CIs for the population parameter θ. In our case the random sample consists of the prediction errors obtained by a certain method. The goal is to utilize the bootstrap distributions in order to construct CI for REC curves and test whether a prediction technique provides better results than a comparative model for a certain accuracy estimator.

The simplest way to construct a $(1-\alpha)\times 100\%$ CI is the bootstrap empirical percentile method. First, from the empirical distribution containing all the θ^{*i} values ($i = 1,2,...,B$) obtained from the bootstrap samples, we compute the values $\theta^*_{a/2}$ and $\theta^*_{1-a/2}$ corresponding to the $100(a/2)$-th and the $100(1-a/2)$-th percentiles. Then, the bootstrap percentile CI is simply given by

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

Two typical measures of accuracy for $\hat{\theta}$ is the *standard error* (SE) (Eq. 2) and the *bias* (Eq. 3) of the estimator that can be also estimated by the bootstrap samples by

$$SE_{boot} = \sqrt{\frac{\sum_{b=1}^{B}[\hat{\theta}^*(b) - \hat{\theta}^*(\cdot)]^2}{\mathrm{B}-1}} \quad \text{where} \quad \hat{\theta}^*(\cdot) = \frac{\sum_{b=1}^{B}\hat{\theta}^*(b)}{B} \tag{2}$$

$$BIAS_{boot} = \hat{\theta}^*(\cdot) - \hat{\theta} \tag{3}$$

Suppose now that we wish to evaluate the prediction performance of a model (ModelA) on a specific SCE dataset. Suppose also that we obtain predictions using the well-known method of *jackknife* (or *hold-one-out)*, i.e. we estimate the cost of each one of the projects in the dataset using a model constructed by all the other projects. After applying the model on the dataset, we obtain by the jackknife method one sample of error expressions which are values of continuous variables. Based on these samples, we have to draw conclusions concerning their means, medians, percentages or certain percentiles of their distributions, so they can be utilized as the basis for the extraction of bootstrap replicates in order to evaluate the CI, SE and bias of the prediction model.

3 Bootstrapping Regression Error Characteristic curves

REC curves were introduced for comparison purposes of SCE models in [6], where it was pointed out that their utilization can be proved quite beneficial since they reinforce the knowledge of project managers obtained either by single accuracy indicators or by comparisons through formal statistical comparisons. Their most important feature is their ability to present easily accuracy results to non-experts and support the decision-making.

More precisely, a REC curve is a two-dimensional plot where the x-axis represents the error tolerance (i.e. all possible values of) of a predefined error measure and the 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. An important feature is that REC curves are very informative since they take into account the whole error distribution, and not just a single statistic of the errors, providing information about extreme points, bias and other characteristics.

REC curves have interesting geometrical characteristics. The most significant one is that commonly used measures of the distribution such as the median or certain percentiles of errors can be estimated by exploiting the shape of a REC curve. In Fig. 1 (a), we see the REC curve of a hypothetical prediction model. The horizontal reference line from 0.5 intersects the REC curve in a point which corresponds to $e = 0.32$ (vertical reference line). This means that 50% of projects have an error smaller than 0.32 which is the median of errors. Similarly, we can evaluate other measures, as for example the well-known pred25.

Based on the bootstrap distributions of error functions, we can easily construct a 95% CI using the bootstrap empirical percentile method in order to draw conclusions regarding the predictive performance of a model. For example, in Fig. 1 (b), we can see the 95% CI of the hypothetical model for the entire distribution of errors. Suppose now that one wishes to know how confident should feel about the accuracy of the constructed model which results in a median error 0.32, that is to estimate a lower and upper bound for this median. Utilizing the bootstrap REC curves, the practitioner should be 95% confident that the unknown parameter for

the median population error varies within the interval [0.22, 0.43]. The same procedure can be followed in order to graphically compare alternative prediction models by constructing the REC 95% CI curves for each model. When the 95% CIs of models do not have an overlapping point, this means that there is a statistically significant different between the predictive performance of the two comparative models. Hence, REC curves provide an easily interpretable visualization technique for the complicated task of the selection of the "best" prediction model on a specific dataset.

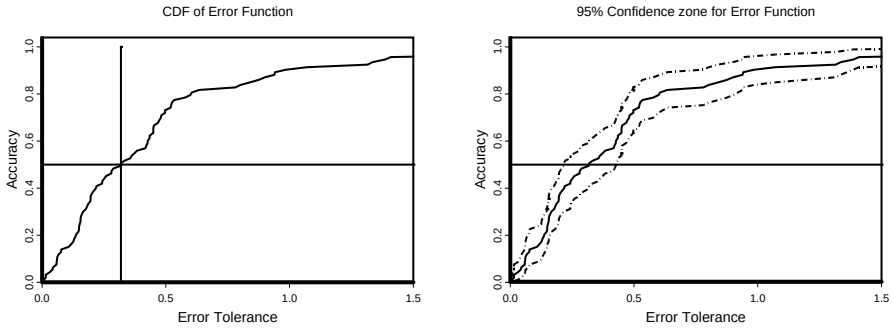

Fig. 1 (a) REC curve example for the evaluation of median error (b) 95% CI of error distribution

4 Experimentation

As the scope of this study is the investigation of bootstrap approach for the construction of REC 95% CIs, we have to choose the prediction techniques to deal with. There are two approaches that have attracted the research interest and have been extensively studied [2], namely the Regression Analysis and the Estimation by Analogy (EbA). The predictive accuracy of the models is usually based on two measures of "local" errors. More specifically, we use the *Magnitude of Relative Error* ($MRE = |\text{actual effort} - \text{estimated effort}| / \text{actual effort}$) and the *Absolute Error* ($AE = |\text{actual effort} - \text{estimated effort}|$) obtained by the jackknife validation of each model. These samples of errors can be utilized for the evaluation of the overall predictive accuracy of each model through well-known statistics (such as the mean and the median) [5]. Furthermore, the sample of errors constitutes the basis for the construction of REC curves and the extraction of bootstrap replicates of the proposed bootstrap method.

The dataset used for experimentation contains 63 projects from a commercial Finnish bank [8]. In order to fit the Regression model, we had to follow all the preliminary analysis for the dataset concerning the transformations and the concatenation of the variables [9]. A Stepwise Regression procedure was then applied to determine the variables having a significant impact on the response variable.

As EbA is free of assumptions, we used all the original variables for building the model, whereas the analogue projects were found through a special dissimilarity coefficient suggested by [10] that takes into account the mixed-type variables. The statistic for the combination of the efforts of neighbor projects was the arithmetic mean, whereas the number of analogies was decided by a calibration procedure, was three.

The REC curves for each of the local accuracy measures obtained by the two comparative models are presented in Fig. 2 (a) and (b). As the REC curves (MRE and AE) for the LS model are always above the corresponding REC curves of EbA, we can infer that LS dominates. Generally, a prediction model performs well if the REC curve climbs rapidly towards the upper left corner. REC curves can also identify extreme errors. When these outliers are present, the top of the REC curve is flat and does not reach 1 until the error tolerance is high. In our plots, we limit the range of the x-axis not to include the extremely high error values for better illustration of the figures. For example, in Figure 2, we can see that both the MRE and AE REC curves for EbA do not reach 1. This fact is a consequence of the presence of few projects producing extremely high values of errors.

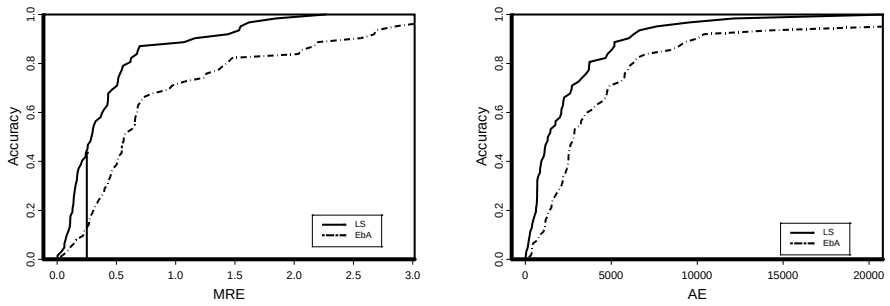

Fig. 2 (a) MRE and (b) AE REC curves for the comparative models

Method	LS		EbA	
Measure	MMRE (%)	MdMRE (%)	MMRE (%)	MdMRE (%)
Actual	45.37	29.38	99.41	56.98
$Estimate_{boot}$	45.41	29.47	98.73	58.88
SE_{boot}	6.09	5.26	12.70	5.79
$Bias_{boot}$	0.04	0.09	-0.68	1.90
95% CI	[34.46, 57.93]	[19.98, 42.56]	[75.65, 124.08]	[50.28, 67.95]
Measure	MAE	MdAE	MAE	MdAE
Actual	2624.16	1373.77	5410.47	2834.33
$Estimate_{boot}$	2624.73	1454.91	5424.04	2952.89
SE_{boot}	430.22	345.81	1006.20	444.48
$Bias_{boot}$	0.57	81.14	13.57	118.56
95% CI	[1908.07, 3558.92]	[922.69, 2150.84]	[3759.62, 7592.91]	[2462, 4193.67]

Table 1 Accuracy measures for the comparative models

The vertical line that crosses the x-axis at 0.25 can be used for the estimation of the pred25 accuracy measures that are based on the MREs of a model (Fig. 2a). More precisely, the pred25 is defined as the percentage of projects with $MRE \leq 0.25$. The aforementioned accuracy measure can be easily evaluated by getting the accuracy of a model at 0.25 error tolerance. It is clear that LS also dominates in terms of pred25 since its value is very close to 0.44 (or 44%), whereas the corresponding value for EbA model is not higher than 0.12 (or 12%).

The general results of the predictive accuracy of the two comparative models are presented in Table 1. It is obvious that LS outperforms EbA in terms of all the accuracy measures. Hence, the conclusions derived from the inspection of REC curves are verified by the accuracy measures that evaluate the prediction performance of the comparative models through certain statistics.

At this point from the bootstrap replicates of MREs (or AEs), we can construct a lower and upper bound (dash line) for each point of the REC 95% CIs (Fig. 3). Moreover, we also report the SE and bias evaluated through the bootstrap technique for each of the comparative models (Table 1). For example, we can observe that the mean (or $\text{Estimate}_{\text{boot}}$) for all the MMRE (Mean MRE) replicates is 45.41%, which is a value very close to the estimated MMRE through the jackknife procedure on the initial dataset and for this reason the bias can be considered low (0.04%). Another interesting issue arisen from the evaluation of the bootstrap accuracy measures is that MMRE and MAE (Mean AE) for EbA present extremely high values of SE compared to the corresponding estimates of the other indicators of error.

Although the REC 95% CIs are very informative, since we can assess a confidence zone for each percentile of the distribution of errors, we cannot draw conclusions for the predictive performance of the alternative models because they do not have one common basis for the comparison procedure. Indeed, the x-axis for the LS model (Fig. 3a) varies from 0 to 1.5, whereas EbA seems to have extremely higher values of error with the maximum value to be higher than 3 (Fig. 3b). This fact is also verified by the inspection of the initial REC curves (Fig. 2), where the LS model climbs more rapidly on the left corner of the graph meaning lower values of errors. The findings are also similar for the case of AEs (Fig. 3c and 3d).

In order to compare the overall predictive performance of the alternative models, we can use the Wilcoxon signed rank test, which constitutes a non-parametric procedure testing whether there is a significant difference between the medians of two paired samples. Alternatively, we propose the utilization of bootstrap REC curves for the identification of significant differences between the medians of the models. Having in mind that we wish to compare the medians of LS and EbA models, we can easily exploit the geometry of REC 95% CIs for 0.50 accuracy value.

As we can observe from Fig. 4a, the 95% CI for the MdMRE (Median MRE) of LS varies within the interval [19.98%, 42.56%], whereas for EbA (dash line) the corresponding interval diversifies within the interval [50.28%, 67.95%]. More importantly, it is obvious from the inspection of the geometry that the two CIs do not

have an overlapping point which means that there is a statistically significant difference between the alternative models. This is also the case regarding the comparison of MdAE (Median AE) (Fig. 4b). More specifically, the 95% CI for LS constructed through the bootstrap replicates of MdAE varies within the interval [922.69, 2150.84], whereas for EbA model within the interval [2462, 4193.67], indicating a statistical significant difference. In order to verify the effectiveness of bootstrap REC curves to graphically detect the differences between the comparative models, we also make use of the Wilcoxon sign rank test for matched pairs. All pair-wise tests have p-values smaller than 0.05 revealing that the differences observed in Fig. 4 are in fact statistically significant.

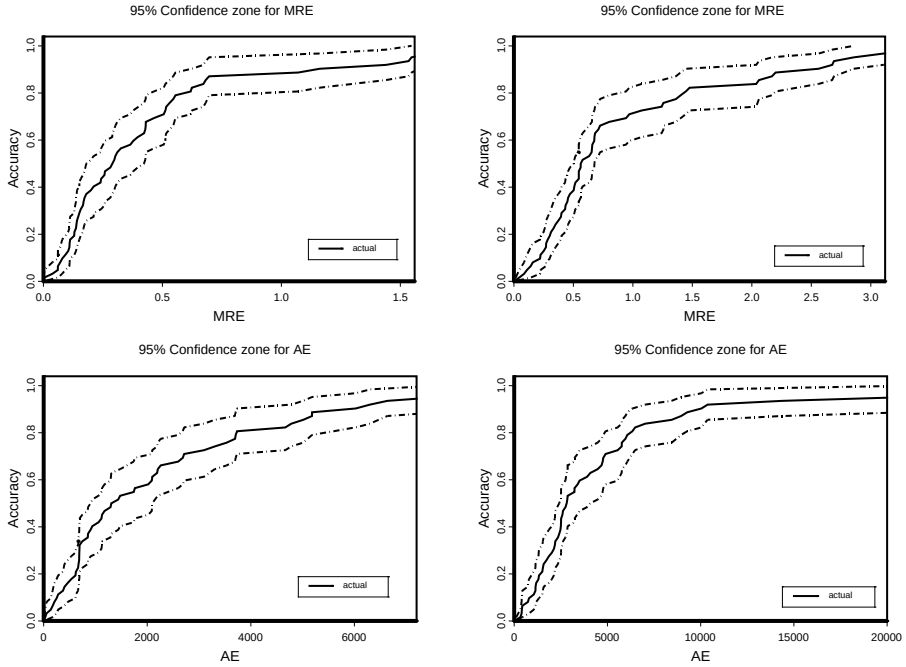

Fig. 3 (a-c) LS and (b-d) EbA REC 95% CI curves for the comparative models

The bootstrap REC curves can also be utilized for the construction of CIs for the pred25 accuracy measure, whereas a hypothesis test can also be conducted for the comparison purposes. Contrary to the MdAE, the pred25 CIs are evaluated by drawing a reference vertical line from 0.25 and from the intersecting points of the REC curve's lower and upper bound, a horizontal line to meet the accuracy axis. Hence, the graphical tool for performing statistical tests for the pred-measures that are essentially percentages and have not been considered yet in formal comparisons, constitutes an easily interpretable manner to assess the predictive power of different models. In Fig. 5, we can notice that the 95% CI for LS varies within the

interval [30.65%, 56.45%] and does not present an overlapping point compared to the EbA model that diversifies within the interval [4.84%, 19.35%], so there is a statistically significant difference between the alternative models regarding the pred25 accuracy measure.

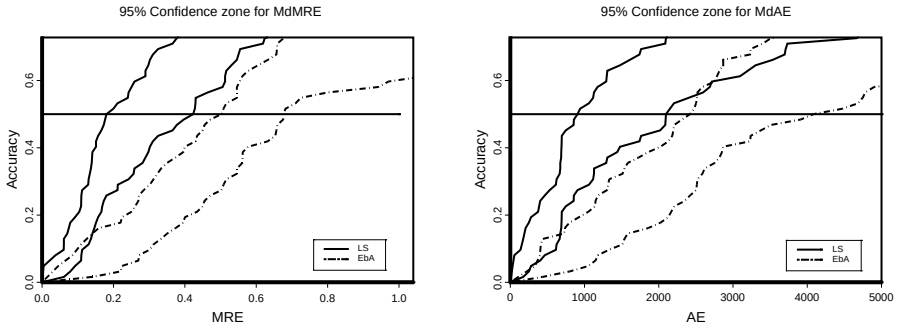

Fig. 4 (a) MdMRE and (b) MdAE comparison for the comparative models

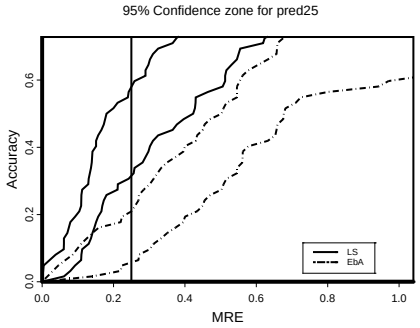

Fig. 5 pred25 comparison for the comparative models

5 Conclusions

In this paper, we deal with the critical task of the selection of the "best" model for a specific Software Cost Estimation dataset with completed projects. More specifically, we extend the utilization of Regression Error Characteristic curves that constitutes an easily interpretable tool, by the construction of bootstrap Confidence Intervals for different error functions.

As the plethora of comparative studies concerning the selection of the "best" model reveals contradictory results, the goal of this paper is to further extend the research on this area. Our intention is not to determine the superiority of either Regression analysis or Estimation by Analogy methods, but rather to facilitate the project managers with a visualization tool that contributes to the systematic com-

parisons of any kind of prediction methods either statistical or artificial intelligence. Moreover, the most important problem that a practitioner has to be faced with is the small-sized and heavily skewed samples of projects and the unavailability of new data in Software Cost Estimation area. This limitation can be resolved by the utilization of bootstrap resampling techniques.

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 EbA and is the most plausible choice for predicting the effort of a forth-coming project. The most important here is that the conclusions obtained by a simple visual comparison through REC curves constructed by the jackknife samples of errors. In addition, the most essential advantage provided by this study, is that we enhance the comparison procedure through the construction of bootstrap CIs for the entire distributions of error functions. In this way, a practitioner is able to assess the benefits for each of the comparative models through the examination of certain percentiles of errors. Furthermore, we also provide a graphical tool to test the statistical significance of the differences between the comparative models for common accuracy measures, like MRE, pred25 and AE through geometrical characteristics and properties of the bootstrap REC curves. Finally, as shown in our experiments, the statistical tests comparing the 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 bootstrap 95% CIs of REC curves.

References

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

[2] Mair, C.M., Shepperd, M.J. (2005). The consistency of empirical comparisons of regression and analogy-based software project cost prediction. *IEEE Proceedings International Symposium on Empirical Software Engineering, (ISESE)*, 509-518.

[3] 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.

[4] Kitchenham, B.A., Pickard, L., MacDonell, S., Shepperd, M. (2001). What accuracy statistics really measure. *IEE Proceedings-Software*, 148(3), 81-85.

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

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

[7] Efron, B., Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall.

[8] Maxwell, K. (2002) *Applied statistics for software managers*. Prentice-Hall, PTR.

[9] Sentas, P., Angelis, L., Stamelos, I., Bleris, G. (2005). Software productivity and effort prediction with ordinal regression. *Information and Software Technology*, 47, 17-29.

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