Empirical Software Engineering, 5, 35–68 (2000)

A Simulation Tool for Efficient Analogy Based Cost Estimation

L. ANGELIS lef@csd.auth.gr
Department of Informatics, Aristotle University of Thessaloniki, 54006, Thessaloniki, GREECE

I. STAMELOS stamelos@csd.auth.gr
Department of Informatics, Aristotle University of Thessaloniki, 54006, Thessaloniki, GREECE

Abstract. Estimation of a software project effort, based on project analogies, is a promising method in the area of software cost estimation. Projects in a historical database, that are analogous (similar) to the project under examination, are detected, and their effort data are used to produce estimates. As in all software cost estimation approaches, important decisions must be made regarding certain parameters, in order to calibrate with local data and obtain reliable estimates. In this paper, we present a statistical simulation tool, namely the bootstrap method, which helps the user in tuning the analogy approach before application to real projects. This is an essential step of the method, because if inappropriate values for the parameters are selected in the first place, the estimate will be inevitably wrong. Additionally, we show how measures of accuracy and in particular, confidence intervals, may be computed for the analogy-based estimates, using the bootstrap method with different assumptions about the population distribution of the data set. Estimate confidence intervals are necessary in order to assess point estimate accuracy and assist risk analysis and project planning. Examples of bootstrap confidence intervals and a comparison with regression models are presented on well-known cost data sets.

Keywords: Software cost estimation, bootstrap samples, confidence intervals, distance metrics, estimation by analogy, regression models

1. Introduction

The more software becomes important in almost every human activity, the more it becomes complex and difficult to implement. Even if modern software technologies render easier the development of certain types of software products, increased user demands and new application domains produce additional problems. It is not surprising that software project management activities are becoming increasingly important.

One of the most critical activities during the software life cycle is that of estimating the effort and time involved in the development of the software product under consideration. This task is known as *Software Cost Estimation*. Estimations may be performed before, during and after the development of software.

The cost and time estimates are necessary during the first phases of the software life cycle, in order to decide whether to proceed or not (feasibility study). Accurate estimates are obtained with great difficulty since, at this point, available data may not be precise, wrong assumptions may be made, etc. During the development process, the cost and time estimates are useful for the initial rough validation and the monitoring of the project's progress. After completion, these estimates may be useful for project productivity assessment.

Estimation methods fall in three main categories, namely *expert judgement, estimation by analogy* and *algorithmic cost estimation*. Expert judgement relies purely on the experience of one or more *experts*. Estimation by analogy compares the software project under consideration with few (e.g. two or three) similar historical projects (i.e. projects with known characteristics, effort and schedule). Algorithmic cost estimation involves the application of a *cost model*, i.e. one or more mathematical formulas which, typically, have been derived through statistical data analysis. Well known examples are regression models, COCOMO (Boehm, 1981) and Function Points (Albrecht and Gaffney, 1983).

All of the three approaches have known advantages and disadvantages. Expert judgement is easy to apply and produces fast evaluation but suffers from the difficulty to find real experts and is exposed to wrong subjective assessment. Estimation by analogy concentrates on a concrete, well-defined estimation framework provided that, suitable projects of the past may be easily found and the mechanism applying analogy is correct. Cost models are very useful when they are used correctly after they have been calibrated with historical data reflecting the characteristics of the estimated project.

After years of application of the variety of software cost estimation techniques, certain principles are widely accepted. Applying these concepts should help the generation of more accurate results. The following hold among cost estimation researchers and practitioners:

a. Every "automatic" cost estimation method should be adapted (calibrated) to local data if meaningful estimates are to be generated (DeMarco, 1982). Calibration should be performed very cautiously, since cost data in the software industry are often inaccurate and properly measured software projects are not easily found. Therefore, software cost data sets tend to be relatively small and noisy.

b. It is safer to produce interval estimates, along with a probability distribution over the estimate interval (see for example Kitchenham and Linkman, 1997). Relying blindly on a point estimate, given the error magnitude associated to the cost estimation techniques, may lead easily in wrong managerial decisions and project failure. An interval estimate may be collapsed to provide a point estimate for practical purposes, but still gives invaluable information about the reliability of the estimation process and provides the basis for risk and what-if project analysis.

c. It is preferable not to generate an estimate at all, than to apply a technique that it is known in advance to be inappropriate for the specific project. Therefore, it is important to possess tools that help in making such a decision.

The above suggest strongly that, before application, an estimation method should be thoroughly understood and carefully analyzed using local data. For example, a typical activity for the adoption of an algorithmic cost model by a software development organization is its calibration by estimating values for the model coefficients. As another example, the weights of the various system types in Function Point Analysis are re-estimated to improve the accuracy of the method. Regression techniques are normally used for this task. Innovations in software technology lead software organizations to the use of advanced development

tools and methods. This situation affects the organization's productivity and alters project characteristics, rendering necessary the repetition of the calibration activities after a certain period of time.

The purpose of this paper is to provide means for calibration, when analogy based estimation is considered along with the ability to generate interval estimates. In Section 2, we provide a summary description of the method using analogies for the estimation of the effort of a software project under development. In Section 3, we describe two different versions of the bootstrap method, the non-parametric and the parametric bootstrap, that will be used subsequently as the basis for calibration and validation of the analogy based estimation. In Section 4, we apply the bootstrap method to the calibration of the estimation by analogy procedure. This involves choice of the appropriate parameters for the analogy method (distance metric, number of analogies and location statistic). In Sections 5 and 6, we describe the procedures for obtaining interval estimates by analogy using non-parametric and parametric bootstrap respectively. In Section 7, we use two data sets in order to perform a parallel study of the interval estimates obtained by the proposed bootstrap method and the interval estimates resulting from the traditional method of regression. In Sections 8 and 9 we discuss practical implications and automation issues of the proposed methods. Finally, in Section 10 we conclude with directions for future research.

Throughout the paper, for illustrative examples and comparative studies, we used two well known data sets, namely the Albrecht data set (Albrecht and Gaffney, 1983) and the Abran-Robillard data set (Abran and Robillard, 1996). All statistical and simulation procedures (including graphics) have been implemented with functions and programs developed in the S-Plus® statistical language.

2. Estimation by Analogy

Estimation by analogy is a technique that has been proposed since a long time as a valid alternative to algorithmic cost estimation and expert judgement (Boehm, 1981). However, only recently (Shepperd et al., 1996; Shepperd and Schofield, 1997) it has been presented in the form of a detailed estimation methodology and has been applied uniformly on a number of cost data sets.

The main aspect of the method is the utilization of historical information from completed projects with known effort. Initially, it is necessary to characterize the new active project, with attributes (variables, also called *features*) common to the ones of the completed projects registered in the database. The attribute values are standardized (between 0 and 1) so that they have the same degree of influence and the method is immune to the choice of units. It is useful for our notation to represent such a historical data set, together with a new project, in the form of a matrix as the one in Table 1.

The following step is to calculate how much the new project differs from the other projects in the available database. This can be done by using a "distance" metric between two projects, based on the values of the k variables for these projects. The most known such distance metric (suggested in Shepperd and Schofield, 1997, on an intuitive basis) is the *Euclidean* or *straight-line distance* which has a straightforward geometrical meaning

Table 1. Characterization of an active project and its placement into a historical data set.

	Actual Effort	*Attribute 1*	*Attribute 2*	...	*Attribute k*
Project 1	E_1	X_{11}	X_{12}	...	X_{1k}
Project 2	E_2	X_{21}	X_{22}	...	X_{2k}
Project 3	E_2	X_{31}	X_{32}	...	X_{3k}
...	...	...	...	...	...
Project n	E_n	X_{n1}	X_{n2}	...	X_{nk}
New Project	Unknown	Y_1	Y_2	...	Y_k

as the distance of two points in the k-dimensional Euclidean space:

$$d_{new,i} = \left\{ \sum_{j=1}^{k} (Y_j - X_{ij})^2 \right\}^{1/2}, \qquad i = 1, 2, \ldots, n$$

The estimation of the effort using analogies is based on the completed projects that are similar to the new one. The notion of similarity in this case coincides with the notion of small distance (in any way it can be measured). Therefore, we can consider similar to the new project a fixed number of the completed ones that have the smallest distances from the new among all projects in the database. The number of the "neighbor" projects is determined empirically in advance and for small sets of data is 1, 2 or 3. The estimation of the effort is eventually obtained by calculating a statistic using the efforts of the neighbor projects. Typically, this statistic is the mean (weighted or unweighted) of these effort values.

Estimation by analogy is essentially a form of Case Based Reasoning (Mukhopadhyay et al., 1992). However, as it is argued in Shepperd and Schofield (1997) there are certain advantages in respect with rule based systems, such as the fact that users are more willing to accept solutions from analogy based techniques, rather than solutions derived from awkward chains of rules or neural nets. On the other hand, there are certain limitations in the use of the method of analogy. As pointed out in the description of the method, the new project under estimation must be inserted in a historical data set where the projects are characterized by specific attributes. This implies that the new project has to be described by exactly the same features. The approach is quite realistic for software organizations that undertake projects in the same domain but problems may occur when a firm uses development tools and technology that are changing fast. Problems in the implementation of the method may also arise when the new project is characterized by completely new features that are not present in the older projects or when new projects are built by reuse of existing code. In any case, the method of analogy, as well as other statistical methods (e.g. regression), is based on existing information. Any firm that wants to apply such methods should maintain and update its data sets according to the changes of the development process. In addition, the method itself may require specific adjustments according to the needs and peculiarities of the new projects.

3. The Bootstrap Method

The method of analogies, as described in Section 2, foresees a method-tuning phase limited to the search of the appropriate project attributes for the estimation procedure. The method also results in a "*point estimation*" for the unknown effort i.e. a single value, which is computed from a particular sample (the historical data set) coming from a practically infinite population with unknown characteristics, i.e. the population of all possible software projects. A critical question is how valid the tuning activity may be and which is the accuracy and reliability of the generated estimation. In these cases, it is common practice to enhance the value of the estimator with various measures of accuracy from its probability distribution. As the underlying probability distribution is unknown too, we are constrained to use the empirical distribution in order to calculate the standard error, the bias, confidence intervals etc. We can also obtain an estimation of the unknown density function of the estimator after the application of a smoothing procedure on the empirical distribution. The crucial point here is to obtain an empirical distribution for the estimator and this is where simulation techniques can contribute.

We will next describe briefly the *bootstrap method*, a simulation technique applicable in two different modes, namely the non-parametric and the parametric bootstrap. The first method is based entirely on the empirical distribution of the data set, without any assumption on the population distribution. The second method is based on the assumption that the data set comes from a theoretical distribution, which is approximated by a parametric model. Bootstrap was invented by Efron in 1979. A detailed account of the principles and the applications of the method can be found in the book of Efron and Tibshirani (1993).

(a) Non-Parametric Bootstrap

The non-parametric bootstrap method is applied when we deal with a random sample $\mathbf{x} = (x_1, \ldots, x_n)$ drawn independently from a family of distributions of the form F_θ where θ is an unknown parameter. Suppose that $\hat{\theta} = T(\mathbf{x})$ is a symmetric function of the sample (independent from the order of the sample). The idea is to take a large number (say B) of samples from $\mathbf{x}$ *with replacement* and to calculate $\hat{\theta}_j,\ j = 1, 2, \ldots, B$ for each of these samples. By "sampling with replacement" we mean the procedure where a random number generator independently selects integers $j_1, j_2, \ldots, j_n$, each of which equals any value between 1 and n with probability $1/n$. These integers determine which members of $\mathbf{x} = (x_1, \ldots, x_n)$ are selected to be in the new random sample. An obvious consequence of this sampling method is that in every new sample there are data points, which appear more than once while others do not appear at all.

(b) Parametric Bootstrap

The difference in this method lies in the way of sampling. Instead of sampling with replacement from the original data points, we draw B samples of size n (same size as in

the original data) from a parametric estimation of the theoretical distribution that generated the data. Therefore, we need first to investigate the possibility that a known distribution (univariate or multivariate, according to our data) fits in a satisfactory way our original sample. Then, applying the appropriate algorithm we can generate random samples from this distribution. It is obvious that the parametric bootstrap is applicable only in cases where there is prior knowledge about the form of the underlying distribution.

The purpose of the method is to measure the variability of $\hat{\theta}$ about θ, studying the distribution of $\hat{\theta}_j$ about $\hat{\theta}$. The most common measures of accuracy that can be obtained from the so-called *bootstrap samples* are the standard error, the bias and the confidence intervals. In our context, we deal mainly with the computation of confidence intervals. The bootstrap sampling (non-parametric or parametric) is applied not only to the univariate data sets but also to the multivariate ones, which is our case. Data points that provide the bootstrap samples are considered the projects in the database. Since the statistical inference based on the non-parametric bootstrap depends solely on the available sample, without any assumptions about the form of the underlying population, the method is particularly useful when our data sets are small and contain non-homogeneous data. As mentioned in the introduction, this is often the case for a software project cost data set. On the other hand, the parametric bootstrap is useful in cases where some knowledge about the form of the underlying population is available.

In the following sections we apply the bootstrap sampling for two purposes: First in order to calibrate the method of analogies choosing the optimal parameters (distance metric, number of analogies, statistic) and second to assess the accuracy of the estimations obtained by the method.

4. Alternative Parameters for the Estimation by Analogy Procedure—Calibration of the Procedure Using Bootstrap

In the previous sections, we pointed out that during the procedure of effort estimation by analogy the user has to decide upon three issues:

(a) The choice of distance metric by which the projects of the database will be sorted according to their similarity to the one under estimation,

(b) the number of closest projects (analogies) and

(c) the statistic that will be computed from the efforts of the closest projects and will serve as estimation for the unknown effort.

As we have already seen, the proposed distance metric is the Euclidean distance, the number of analogies is relatively small (1, 2 or 3) and the estimation statistic is the arithmetic mean. From now on, we refer to these choices as the *parameters of the analogy procedure*. In the following of this section, we review some other alternatives for the above parameters, we discuss measures of performance for them and we illustrate their use by an example.

(a) The Distance Between Projects

It is argued that the Euclidean distance has a disadvantage as far as each co-ordinate contributes equally to its calculation. As the co-ordinate variables represent measurements that are strongly affected by random fluctuations of differing magnitudes, it seems reasonable sometimes to weight the co-ordinates in such a way that attributes with low variability, influence the distance more than those with high variability. Therefore, an alternative metric, which tries to balance the importance of each variable, is the *scaled Euclidean distance*:

$$d_{new,i} = \left\{ \sum_{j=1}^{k} w_j^2 (Y_j - X_{ij})^2 \right\}^{1/2} \qquad i = 1, 2, \ldots, n$$

where the w_j are appropriate weights. Some typical choices are: $w_j = 1/$(standard deviation of j-th variable) or w_j=1/(range of Xj). A very interesting discussion on the usefulness of weighted distances and on the standardization of the variables generally, can be found in Kaufman and Rousseeuw (1990). However, there are some other well known metrics for the distance of two points when the variables are quantitative (see Krzanowski, 1993):

Minkowski distance: $d_{new,i} = \{\sum_{j=1}^{k} |Y_j - X_{ij}^{\lambda}|^{\lambda}\}^{1/\lambda}$ where λ is an integer. When $\lambda = 1$ the distance is known as *city block* or *Manhattan distance* while the case $\lambda = 2$ gives the Euclidean distance.

Canberra distance: $d_{new,i} = \sum_{j=1}^{k} \dfrac{|Y_j - X_{ij}|}{Y_j + X_{ij}}$

Czekanowski coefficient: $d_{new,i} = 1 - \dfrac{2 \sum_{j=1}^{k} \min(Y_j, X_{ij})}{\sum_{j=1}^{k} (Y_j + X_{ij})}$

Chebychev or "maximum" distance: $d_{new,i} = \max_{1 \le j \le k} |Y_j - X_{ij}|$

Although the above metrics refer to quantitative variables, there are some others for the cases where the variables are dichotomous and also others regarding mixed data, i.e. with various types of variables (ordinal, nominal, interval and ratio variables) (Kaufman and Rousseeuw, 1990; Krzanowski, 1993). As it is quite natural for data sets with software projects to have variables of mixed type, we mention a special *dissimilarity coefficient* suggested by Kaufman and Rousseeuw (1990):

$$d_{new,i} = \frac{\sum_{j=1}^{k} \delta_{new,i}^{(j)} d_{new,i}^{(j)}}{\sum_{j=1}^{k} \delta_{new,i}^{(j)}}$$

where:

$$\delta_{new,i}^{(j)} = \begin{cases} 1 & \text{if } X_{ij}, Y_j \text{ nonmissing} \\ 0 & \text{otherwise,} \end{cases}$$

and

$$d_{new,i}^{(j)} = \begin{cases} 1 \text{ if } X_{ij} \neq Y_j \\ \\ 0 \text{ if } X_{ij} = Y_j \end{cases} \text{ when } j\text{-th variable is binary or nominal}$$

while

$$d_{new,i}^{(j)} = \frac{|X_{ij} - Y_j|}{R_j}$$

when j-th variable is ordinal, interval or ratio. R_j is the range of the variable.

(b) The Number of Analogies

When small sets of data are available, it is quite reasonable to consider only a small number of analogies. However, when the available data are comprised of many projects or when the statistic that will provide the effort estimation is unaffected by extreme values, we can increase the number of analogies. The point is that there is no rule of thumb to decide upon this number without experimenting with the available data.

(c) The Calculated Statistic

When the analogy projects are selected, we have to compute a certain location statistic based on their efforts, which may represent an efficient approximation of the unknown effort. Of course, when the analogies have been determined to be only two, the arithmetic mean (or at least a weighted mean) is the only plausible choice. Alternatively, when the number of the closest projects increases, one can replace the mean by another more robust statistic. There is a great number of such statistics available in the statistical literature and the most known ones are the median, the trimmed mean, Tukey's biweight estimator and Huber's estimator. Such statistics (most of them belong to a special class called *M-estimators*) have the advantage of remaining unaffected by the outlying values which usually appear in effort values (see Hoaglin et al., 1983). On the other hand, one may want to take into account the outlying values and, in this case, the mean is the most appropriate statistic. It is always recommended to perform first a descriptive statistical analysis on the entire set of effort values obtained from the neighbor projects and then to decide for the suitable statistic.

It is obvious that the choice of the distance metric, the determination of the number of analogies as well as the choice of the appropriate location statistic are quite empirical aspects and require a great amount of experimentation before the final decision. The *jackknife* method (also known as *leave-one-out cross validation*) is a useful tool to validate the error of the prediction procedure employed, in order to estimate the effort of a new project. In each step, a completed project (say the i-project) is removed from the data set and the remaining projects are used as a basis for the estimation $\hat{E}_i$ of its effort (the parameters of the procedure are fixed throughout this process). Two common estimates

of the prediction error, computed from the jackknife process, are the *mean magnitude of relative error* (MMRE) defined as

$$MMRE = \frac{100}{n} \sum_{i=1}^{n} \frac{|E_i - \hat{E}_i|}{E_i}.$$

and the percentage of estimations that fall within 25% of the actual effort denoted by PRED(25) (see Conte et al., 1986). The following example indicates how the parameters affect the two measures of prediction error in the case of a well-known data set.

Example 1. We use the known as Albrecht data set to test how the error of the predictions based on analogies is influenced by the choice of the parameters of the analogy procedure. The (explanatory) variables for the calculation of the distances are IN (number of external inputs), OUT (number of external outputs), FILE (number of external and internal files) and INQ (number of user inquiries), according to Function Point terminology. The dependent variable is the variable EFFORT measured in man months. Here we apply the jackknife technique for various combinations of the above parameters and each time we calculate MMRE and PRED(25). The results in Table 2 have been obtained by considering three distance metrics (Euclidean, Manhattan and Maximum), three choices for the number of analogies (2, 5 and 10), two choices for the location parameter (mean and median—coincident for 2 analogies) and two alternatives regarding the standardization of the variables. The standardization is depicted in the third column of Table 2, where "Yes" means that before the calculations of the distances, the i-th value of the X_j variable has been standardized in the interval [0, 1] according to the rule:

$$X_{ij} \leftarrow \frac{X_{ij} - \min_{1 \le m \le n}(X_{mj})}{\max_{1 \le m \le n}(X_{mj}) - \min_{1 \le m \le n}(X_{mj})}$$

As can be seen from Table 2, for this particular small data set (it contains 24 completed projects) the smallest error is obtained when the estimation is based on a small number of analogies and when the variables are standardized. An interesting result is that the "maximum" distance improves the prediction error when the number of analogies is 2 while it provides the best overall MMRE (68%) and a relatively high PRED(25) (33%). Another fact that is also worth mentioning is that the best PRED(25) (42%) was achieved with 10 analogies and the "maximum" distance.

The question here is whether we can rely on the MMRE or PRED(25) in order to decide for the parameters of the estimation by analogies procedure when we have to predict the effort of a new project. The reason for being cautious towards a positive answer is this: The values of MMRE and PRED(25) that estimate the prediction error of the procedure, are derived from only a small sample of projects drawn from an unknown infinite population. Hence, the prediction of a new project's effort is subject to an error of intangible variability. Ideally, we should be able to have repeated samples from the original population and for each of them to compute the MMRE and PRED(25) after using exactly the same parameters. Then, their average values would produce efficient estimates of the prediction error with the specific parameters. Additionally, we would have the ability to measure the accuracy

Table 2. Comparison of the various parameters in estimation by analogy.

Distance	Analogies	Std	Statistic	MMRE	PRED(25)
Euclidean	2	Yes	Mean	73%	33%
		No		81%	29%
	5	Yes	Mean	80%	33%
		No		81%	33%
		Yes	Median	82%	21%
		No		85%	13%
	10	Yes	Mean	114%	25%
		No		98%	25%
		Yes	Median	108%	25%
		No		98%	21%
Manhattan	2	Yes	Mean	74%	29%
		No		85%	33%
	5	Yes	Mean	81%	33%
		No		82%	29%
		Yes	Median	85%	17%
		No		86%	17%
	10	Yes	Mean	106%	25%
		No		100%	25%
		Yes	Median	102%	17%
		No		98%	21%
Maximum	2	Yes	Mean	68%	33%
		No		75%	33%
	5	Yes	Mean	87%	25%
		No		83%	33%
		Yes	Median	86%	25%
		No		86%	17%
	10	Yes	Mean	116%	42%
		No		106%	21%
		Yes	Median	112%	25%
		No		102%	17%

of the estimated error taking confidence intervals for the error. Comparing these values for a range of parameters, we could point out the parameters that optimize the performance of the prediction procedure. Of course the availability of new data from the population is impossible, but instead we can draw bootstrap samples (non-parametric or parametric) from the available data set. This kind of search for optimal parameters is sometimes called *calibration of the estimation procedure*. Example 2 is a demonstration of the bootstrap samples utilization, for the calibration of the estimation by analogies method in the Albrecht data set.

Example 2. We used the same data set as in Example 1 (Albrecht data set) to draw 1000 bootstrap samples with replacement. For each one of these samples we computed MMRE and PRED(25) for fixed distance metric and location statistic and for the complete range of possible values for the number of analogies. Actually, the number of analogies can range from only 1 to 23 (the number of projects in the data set minus one). In Figure 1, there are

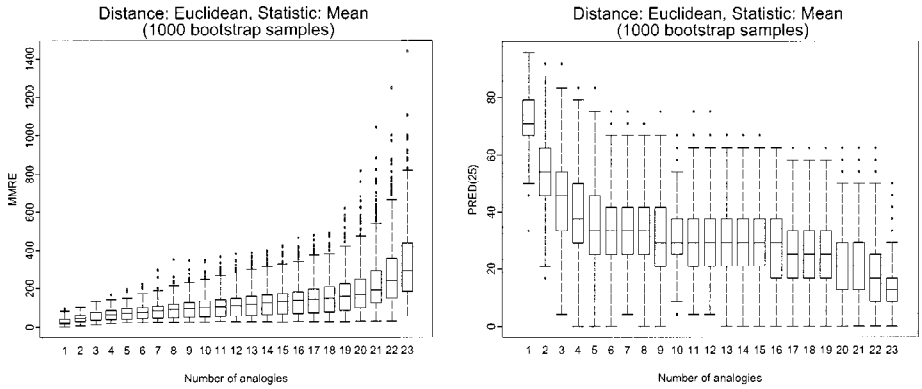

Figure 1. Distributions of MMRE and PRED(25) for various numbers of analogies.

boxplots, each one for a certain number of analogies, showing the distribution of MMRE and PRED(25) when the distance is Euclidean and the statistic is the arithmetic mean. The variables have been standardized as described in Example 1. Both panels of Figure 1 show that the best choice for the number of analogies, for the specific data set and the particular choice of the other fixed parameters, is one project. The empirical distributions of MMRE and PRED(25) for every number of analogies give us invaluable information. Table 3 shows for every case the mean values of MMRE and PRED(25) together with the standard error (SE) and a 95% confidence interval (CI). The standard error is simply the sample standard deviation computed from all the samples while the bounds of every confidence interval are the 2.5% and 97.5% percentiles of the empirical distribution. For example, for the case of only one analogy, the mean value of the MMRE is 31% while a 95% confidence interval (CI) is [9%, 74%]. For the same case, the mean value of PRED(25) is 72% and a 95% CI is [54%, 88%]. An interesting point is that the increase of the number of analogies causes not only the increase of MMRE's mean but also its variability towards large extreme values. This means in practice, that when we use a large number of analogies for our prediction, not only we have to expect an increased value of MMRE, but we also have to face the risk of having an extremely large value of prediction error, that is to make an unreasonably optimistic or pessimistic prediction.

In the same data set, we also fixed the distance (Euclidean) and the number of analogies. We chose the number of analogies to be 23 i.e. each project's effort in the data base is estimated by a statistic computed from all the other 23 projects' efforts. Of course, the maximum number of analogies is used here only for the purpose of illustration, as it is expected that the prediction error will be at extremely high levels. However, the two statistics that are compared, the mean and the median, seem to differ in performance as we can see in Figure 2. The median decreases the mean value of MMRE (331% for the mean, 165% for the median) and increases the mean value of PRED(25) (14% for the mean and 27% for the median).Similar simulation techniques were applied by fixing the number of analogies and the statistic in order to investigate any differences between three distance

Table 3. Mean, standard error and confidence intervals for MMRE and PRED(25) from the application of bootstrap on the Albrecht data set.

	MMRE (%)			PRED(25) (%)		
Analogies	Mean	SE	95% CI	Mean	SE	95% CI
1	31	18	9–74	72	9	54–88
2	47	20	17–86	54	13	25–79
3	58	23	24–100	45	14	17–71
4	66	26	28–114	39	14	13–67
5	72	29	30–130	36	13	13–58
6	77	33	32–142	34	13	12–58
7	82	37	33–158	33	12	13–63
8	88	42	35–177	33	12	8–58
9	94	47	36–198	32	12	13–58
10	100	51	38–221	32	12	13–58
11	106	55	39–230	31	12	8–58
12	112	59	41–244	31	12	8–54
13	119	63	43–263	30	11	8–54
14	126	67	45–272	30	11	8–50
15	133	71	47–288	28	12	8–50
16	139	74	48–298	28	12	8–50
17	146	78	50–315	27	12	4–50
18	153	82	53–333	25	12	4–50
19	165	92	54–371	24	12	4–50
20	187	109	56–451	22	12	4–50
21	221	133	63–555	20	12	0–46
22	270	159	74–658	17	11	0–42
23	327	182	95–756	14	9	0–33

metrics (Euclidean, Manhattan, Maximum). The results from 1000 bootstrap samples showed that for the particular data set the three distances produce almost the same results for MMRE and PRED(25). However, in other data sets, different distance metrics may produce different results.

5. Measurement of Accuracy for the Effort Estimation with Non-Parametric Bootstrap

In this section, we deal with the generation of interval estimates. We follow the non-parametric bootstrap method described in Section 3, but we consider now the set of the completed projects as a sample of multivariate observations coming from an unknown population. The population's parameter we have to estimate is the effort of the new project, i.e. the particular value of one variable (effort) of a point whose values for the other variables (dependent) are known. The function used by the method of analogies for estimation, is the mean value of the efforts obtained by projects with minimum Euclidean distance from the new one. In order to obtain confidence intervals for this estimation, we apply an algorithm, which generates a large number of independent bootstrap samples drawn randomly with replacement from the available data set. Each sample contains exactly n projects (as many

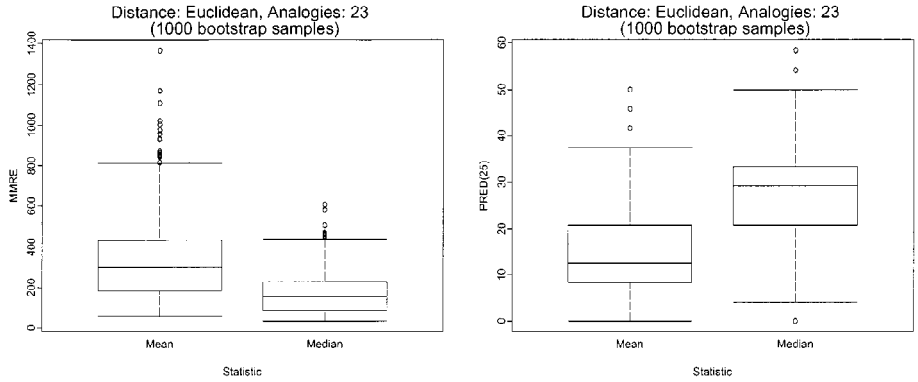

Figure 2. Distributions of MMRE and PRED(25) for two alternatives of the statistic.

as the original data set) but some of the original projects appear more than once while others do not appear at all. Each time a sample is drawn, we find the closest projects to the new one and we calculate the mean of their efforts. Of course, as we saw in Section 4, we can always employ alternative measures for the distance or for the location statistic instead of the mean. It must be noted that the estimations from the bootstrap samples are derived in exactly the same way as from the original data. The estimations from each sample are recorded and will be used for the computation of the appropriate variability measures.

Let us denote by $\hat{E}_{new}$ the estimation for the effort of the new project obtained by the procedure using analogies when applied to the original data set. Let also $E^*_{new}(b)$ denote the estimation obtained by the same procedure applied to the b-th bootstrap sample where $b = 1, 2, \ldots, B$ and B is the total number of the samples. A typical measure of accuracy for $\hat{E}_{new}$ is the standard error, which in turn can be estimated by

$$SE_{boot} = \sqrt{\frac{\sum_{b=1}^{B}[E^*_{new}(b) - E^*_{new}(\cdot)]^2}{B-1}} \quad \text{where} \quad E^*_{new}(\cdot) = \frac{\sum_{b=1}^{B} E^*_{new}(b)}{B}.$$

Another well-known measure of accuracy is the bias of the estimator. This can also be estimated by the bootstrap samples by

$$BIAS_{boot} = E^*_{new}(\cdot) - \hat{E}_{new}.$$

A simple rule for the accuracy of an estimator is that $|\, BIAS_{boot} \,/\, SE_{boot} \,| \leq .25$. The most useful of all the measures is the confidence interval. The point estimation $\hat{E}_{new}$ is just a guess for the true value of the new project's effort. This estimation should always be accompanied with a lower and an upper bound together with a probability level. A *(1-a)% confidence interval* for $\hat{E}_{new}$ can be estimated from the bootstrap samples by

$$(1-a)\%\, CI_{boot} = \left[E^*_{new;a/2},\, E^*_{new;(1-a/2)}\right]$$

where a is a small probability level while $E^*_{new;a/2}$ and $E^*_{new;(1-a/2)}$ are the $100 \cdot a/2$ and the $100 \cdot (1 - a/2)$ empirical percentile points of all the $E^*_{new}(b)$ values. A confidence interval

of this type is called *percentile interval*. The goal of confidence intervals is to provide us with an interval estimation consisting of an "optimistic" and a "pessimistic" guess for the true magnitude of the effort.

Beside the measures of accuracy we just mentioned, it is often very useful to represent graphically the entire distribution of all the estimations $E^*_{new}(b)$, $b = 1, 2, \ldots, B$ computed from the B bootstrap samples. For this purpose, we can employ several known graphic methods such as boxplots, histograms etc. An advanced statistical analysis requires the fitting of a smooth density function to the empirical distribution of the bootstrap estimations. In a simple case, a known theoretical distribution such as the Normal, the Lognormal or the Gamma distribution may be fitted satisfactorily to the estimations. In the most usual case where no such distribution can be found, we estimate the unknown density function $f(x)$ by a smoothing procedure such as the *kernel density estimation* (see Silverman, 1986). The estimation is based on an expression of the form

$$\hat{f}(x) = \frac{1}{w} \sum_{b=1}^{B} K\left(\frac{x - E^*_{new}(b)}{w}\right)$$

where we need to predefine a *kernel function* $K(\,)$ and a constant w called the *bandwidth*. The kernel is, usually, chosen to be a known probability density function such as the Gaussian function, while the bandwidth determines the level of smoothing. Of course, someone has to try several kernels and bandwidths in order to represent efficiently the distribution of the data (see Venables and Ripley, 1994).

Example 3. For an illustration of the non-parametric bootstrap confidence intervals, we use again the Albrecht data set to estimate the effort of a new hypothetical project. Suppose that the new project has the following known values for the (explanatory) variables: IN = 27, OUT = 36, FILE = 20, and INQ = 10. The estimation based on 2 analogies from the existing data set (Euclidean distance, standardized variables and mean of the closest projects efforts) is 5.55 man months. In order to figure out the accuracy of this estimation by the measures discussed, we perform a bootstrap procedure taking with replacement $B = 1000$ samples and estimating each time the effort using 2 analogies exactly as for the point estimation. Figure 3(a) shows a histogram of the empirical distribution of these estimations. The histogram shows that the underlying probability density function is far from being normal. In (b) we can see three estimations of the density function, with different smoothness, based on the kernels' method. In both panels, the vertical dashed line indicates the estimation by the method of analogies (5.55).

At this point, from the 1000 bootstrap estimations we can compute the standard error, which is $SE_{boot} = 2.29$. The mean of all estimations is $E^*_{new}(\cdot) = 6.29$ and hence an approximation to the bias of the estimation is derived by the difference $BIAS_{boot} = 6.29 - 5.55 = .74$. According to the empirical rule, the absolute ratio of the bias to the standard error is $.32 > .25$, so the bias is considered large. This simply means that in the average we tend to overestimate the unknown effort. Finally, we can easily compute a confidence interval for our estimation utilizing the percentiles of the empirical distribution derived by the bootstrap estimations. For example, the 2.5% and 97.5% percentiles (the $1000 \cdot .025 = 25$th and the $1000 \cdot .975 = 975$th values respectively when we sort the estimations in ascending

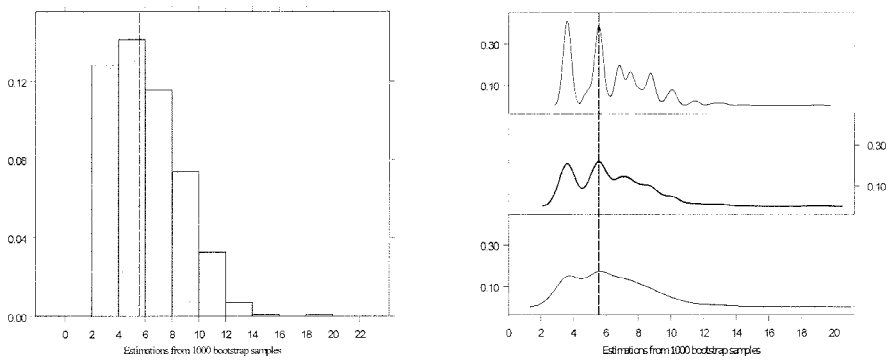

Figure 3. (a) Histogram of the estimations obtained from 1000 bootstrap samples. (b) Kernel density estimation of a smooth density function.

order) produce the 95% $CI_{boot} = [3.6, 11.45]$. A narrower confidence interval, is 50% $CI_{boot} = [3.6, 7.5]$ which provides more reasonable predictions but with less probability.

6. Measurement of Accuracy for the Effort Estimation with Parametric Bootstrap

The bootstrap method we discuss in this section is characterized by the initial assumption that the historical data set, or some transformation of it, is a sample from a theoretical multivariate distribution which can be consider as normal. Experience shows, that in many practical problems the assumption of normality is quite realistic while the mathematical modeling of the data as sample from a normal distribution has many theoretical advantages (see Johnson and Wichern, 1988). In general, the method may as well, presuppose some other multivariate distribution other than the normal. In this case, the verification of the assumption and the generation of random numbers are extremely complicated tasks. Additionally, in most practical situations, an appropriate transformation on the original data can produce variables following a normal distribution. For the above reasons, we deal only with situations where we can assume normality.

The method we are going to describe, generates a large number of samples, each one of size n (the same as the size of the original data set) with *artificial projects* from a hypothetical distribution, which is assumed to fit the original data. This method is different from the non-parametric bootstrap, which is based on samples containing real projects drawn from the available data set. The purposes of the parametric bootstrap are the same as those of the non-parametric one. In other words, we want to estimate measures of accuracy for the estimation by analogy procedure and especially confidence intervals. The complete procedure of the estimation can be summarized in the following steps:

Step 1: Assess the Assumption of Normality. We investigate the possibility that the set of our data, in the form of Table 1, is derived from a $k + 1$-dimensional normal distribution.

However, this seems unrealistic for software cost data sets as all the variables can take only positive values while often large positive outliers appear. The assumption of normality is not only misleading but also without practical purpose, since the sampling we are about to perform later will inevitably generate negative values. A common way to bring the data in a "more normal" framework, is their transformation, i.e. their re-expression in different units. Experience and theory have shown that logarithmic (or *logit*) and square root transformations are quite effective in producing quantities normally distributed. Moreover, these transformations ensure the generation of only positive values in the subsequent sampling.

The next question we have to answer is whether the transformed variables really show normal behavior or it is necessary to proceed with the application of an alternative transformation, which will be more effective. As the known overall goodness-of-fit tests assessing the joint normality in more than one dimensions are very complex, it is a common practice to concentrate on the behavior of the variables in one and two dimensions (univariate and bivariate cases). This is justified by the property of the multivariate normal distribution stating that every subset of its variables also comes from a normal distribution. Of course, the opposite is not true, but for practical situations it is usually enough to ensure that each one of the variables comes from a univariate normal distribution and that every pair of the variables comes from a bivariate normal. We refer to Johnson and Wichern (1988) for a complete discussion on assessing normality in the univariate and bivariate case. A very useful graphical tool in order to detect normality in one dimension is the *Q-Q plot*, the plot of order data against the percentiles of a normal distribution. A straight line here shows normal behavior. A more objective way to check normality in the univariate case is the Kolmogorov-Smirnov goodness-of-fit test, which tests the null hypothesis that a normal distribution equals the real distribution of the data. In the bivariate case we just form scatterplots for all possible pairs of the variables. If the points exhibit an almost elliptical pattern, we have some evidence of normality.

Step 2: Estimate the Parameters of the Multivariate Normal Distribution. For the random vector $\mathbf{X} = (X_1, X_2, \ldots, X_p)'$, the *p-dimensional normal* density function, which is denoted by $Np(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ has the form

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-(1/2)(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

where the $p \times 1$ vector $\boldsymbol{\mu}$ represents the expected value of the random vector $\boldsymbol{X}$ and the $p \times p$ matrix $\boldsymbol{\Sigma}$ the variance–covariance matrix. The vector $\boldsymbol{\mu}$ and the matrix $\boldsymbol{\Sigma}$ are the *parameters* of the multivariate normal distribution, which have to be evaluated from the sample.

In our setup, we have originally n data points (projects) $\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_n$, and each one of them is a $(k+1)$-vector $\mathbf{P}_i = (E_i, X_{i1}, \ldots, X_{ik})'$, $i = 1, 2, \ldots, n$ (see Table 1). Suppose that under a certain set of transformation functions $g_j(x)$, $j = 1, 2, \ldots, k+1$ the new vectors

$$\mathbf{T}_i = [g_1(E_i), g_2(X_{i1}), \ldots, g_{k+1}(X_{ik})]' = [Y_i, Z_{i1}, \ldots, Z_{ik}]$$

have been found in Step 1 to represent a sample coming from a $k+1$-dimensional normal distribution $N_{k+1}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then, the parameters of this distribution are evaluated from the sample by the sample mean and the sample covariance matrix

$$\hat{\boldsymbol{\mu}} = \overline{\mathbf{T}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{T}_i \quad \text{and} \quad \hat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{T}_i - \overline{\mathbf{T}})(\mathbf{T}_i - \overline{\mathbf{T}})'.$$

In simple words, we estimate the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ of the multivariate normal distribution that generated the (transformed) data by computing the arithmetic mean and the variance for each one of the (transformed) variables, and the covariances between all pairs.

At this point, we have to remark that the procedure of approximating the population distribution by a parametric model, estimated from the sample, is typical in statistical practice. However, we have to bear in mind that the accuracy of such approximations depends on the quality of the sample. In general, we have better approximations when the sample is large and homogeneous while small samples or data generated from more than one populations may produce bad approximations and eventually misleading bootstrap samples.

Step 3: Sample from the Assumed Normal Distribution. We extract a large number of samples each of size n, from the multivariate normal distribution that was found in Steps 1 and 2 to fit to our transformed data. The sampling is accomplished by one of the algorithms that are available in the literature for the generation of random numbers (see for example Law and Kelton, 1991). Each sample must then be re-transformed in the original units, applying the inverse transformation, in order to produce the artificial projects which in turn will be used to estimate the effort of the new project.

Step 4: Estimate by Analogy Using the Parametric Bootstrap Samples. For each one of the samples drawn in Step 3, we use a standard estimation procedure by analogies to estimate the new projects' effort. The distance metric, the number of analogies and the statistic are kept fixed throughout the process. The estimation obtained from each sample is stored in a vector, which will provide the parametric measures of accuracy for the estimation by analogies.

Step 5: Calculate Confidence Intervals and Other Measures of Accuracy. The estimations obtained from each bootstrap sample, are used for the calculation of confidence intervals (CI) and the other measures of accuracy (SE and BIAS) described in Section 5.

Example 4. We illustrate the method of parametric bootstrap using the Albrecht data set to estimate the effort of a new hypothetical project with values for the explanatory variables as in Example 3. The point estimation based on 2 analogies from the existing data set is 5.55 man months. In the first step, the tests for normality showed that none of the original variables is normally distributed. After trying various transformations on the variables, the logarithmic transformation was found suitable for variables EFFORT, IN, OUT and FILE while for the last variable, INQ, the square root transformation was found to accomplish normality. For example, in Figure 4 we can see Q-Q plots for the

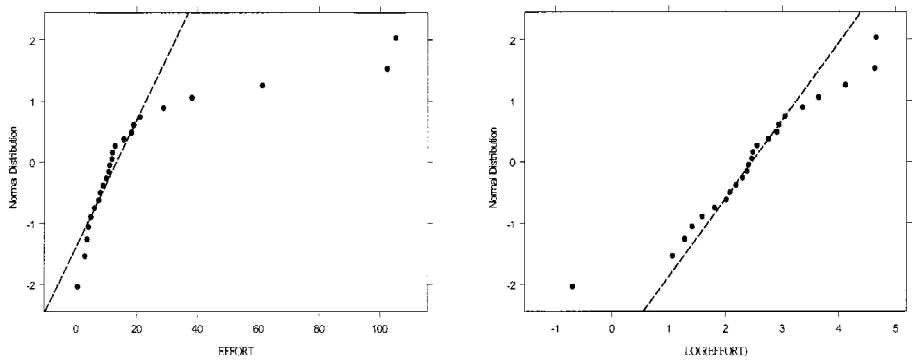

Figure 4. Q-Q plots for the variable EFFORT and its logarithmic transformation.

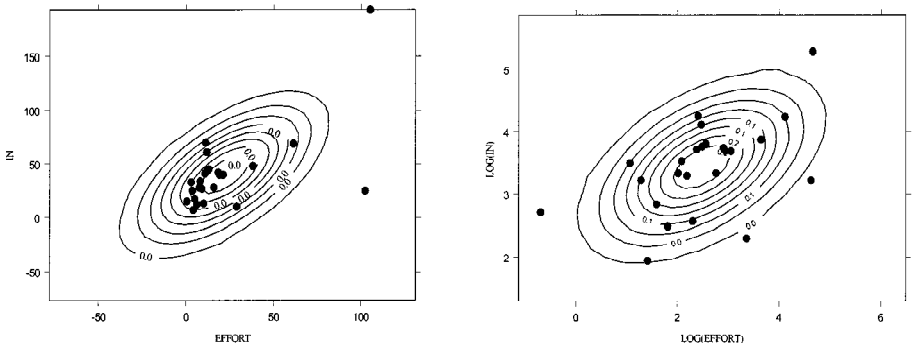

Figure 5. Scatterplots for the variables EFFORT and IN before and after the logarithmic transformation together with contours of a bivariate normal distribution.

variable EFFORT and its logarithmic transformation LOG(EFFORT). Specifically, the large deviation from the straight line in left panel is an indication of non-normality while the right panel suggests that the logarithmic transformation removes a great deal of the curvature and produces values normally distributed. The Kolmogorov–Smirnov test verifies the above indications.

Usually, in practice we can assume joint normality when approximate marginal normality is achieved for each one of the variables. However, we can detect bivariate normality for all the pairs of the variables by scatterplots as in Figure 5. The scatterplots are accompanied by contour plots of the most fitting bivariate normal distribution. In the left panel, we see that the pairs for the values of the original variables EFFORT and IN do not exhibit an elliptical pattern, as they are concentrated in a region of the plot. In the right panel, the transformed values are distributed randomly forming an elliptical configuration.

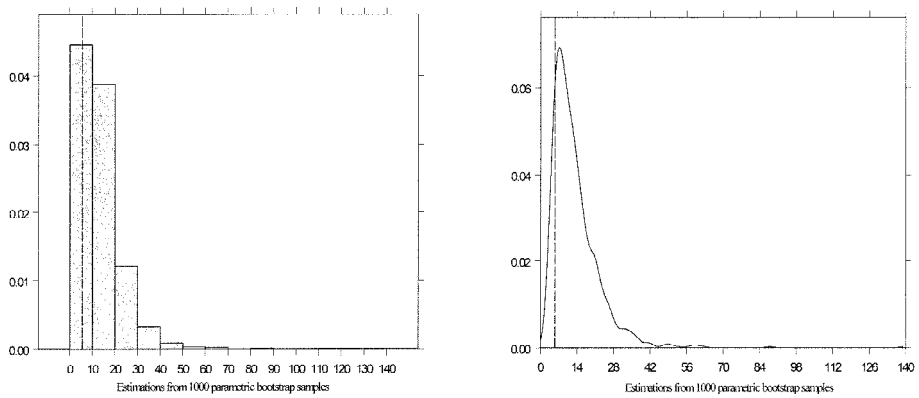

Figure 6. Histogram of the estimations obtained from 1000 parametric bootstrap samples.

Next, we estimate the parameters of the multivariate normal distribution fitting the transformed data obtaining

$$\hat{\boldsymbol{\mu}} = [2.48, 3.43, 3.60, 2.52, 3.37]$$

and

$$\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} 1.363 & .438 & .712 & .670 & 1.401 \\ .438 & .523 & .296 & .206 & .611 \\ .712 & .296 & .537 & .391 & .902 \\ .670 & .206 & .391 & .691 & .664 \\ 1.401 & .611 & .902 & .664 & 5.788 \end{bmatrix}$$

Finally, we are ready to draw samples from the $N_5(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ normal distribution we found fitting our transformed data. A number of $B = 1000$ samples of size $n = 24$ (as the original data) were taken. For each one of them, the inverse transformations (exponential and square power) were applied to bring the samples into the original units. Each time the effort of the new projects was estimated using 2 analogies. Figure 6 shows a histogram of the empirical distribution and a smooth density function for these estimations. The vertical dashed line is used to mark the position of estimation by the method of analogies (5.55).

The standard error is $SE_{boot} = 9.61$. The mean of all estimations is $E^*_{new}(\cdot) = 13.30$ and bias of the estimation is approximated by $BIAS_{boot} = 13.30 - 5.55 = 7.75$. The absolute ratio of the bias to the standard error is $.81 > .25$. A 95% confidence interval for our estimation is 95% $CI_{boot} = [3.46, 35.12]$.

Comparing now these results with the ones from the non-parametric bootstrap, we notice that the parametric bootstrap reveals lower accuracy for our predictions. It is remarkable that while the two methods provide almost identical lower bounds for their confidence intervals, the upper bound is much larger in the parametric case. This is reasonable considering that this last estimation was derived under the assumption of a theoretical distribution (some

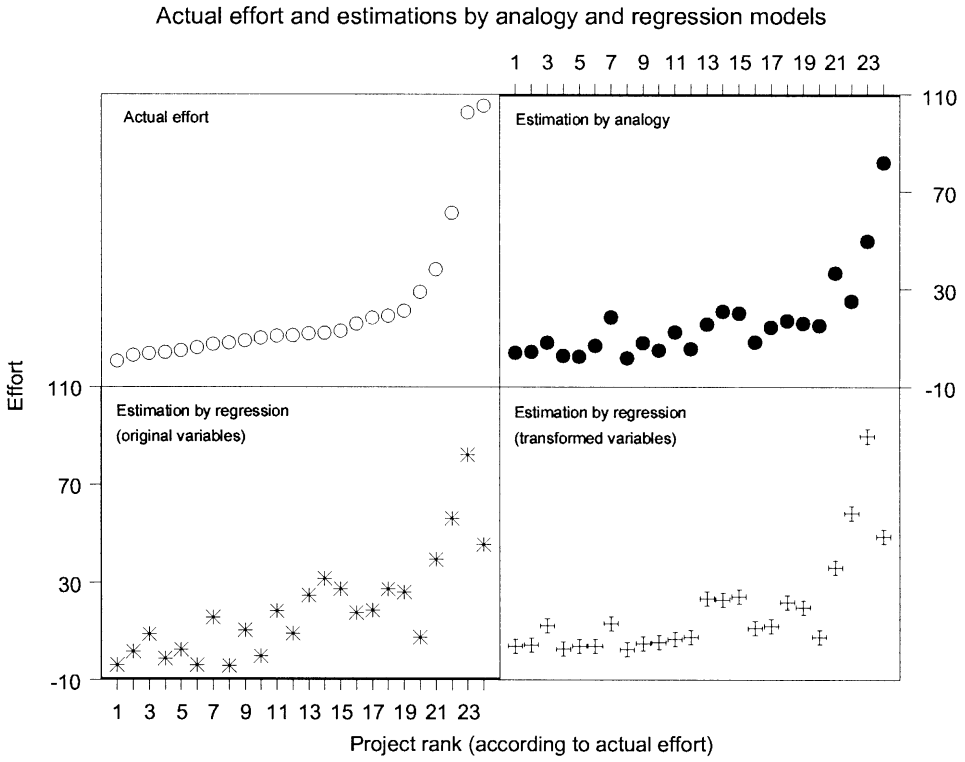

Figure 7. The actual efforts of the Albrecht data set with estimations by analogy and regression models.

transformation of the normal) whose values can be infinitely large, although with probability nearly zero. The samples drawn from this distribution may contain very large values for the effort. This does not happen in the non-parametric case where the samples come from a certain data set with bounded values. However, in cases where we suspect that the effort of a project may take an unreasonably large value, it is preferable to use the parametric bootstrap for interval estimation.

7. Comparison of Bootstrap Confidence Intervals for the Method of Analogies and Regression Confidence Intervals

In this section, we present a comparison of confidence intervals obtained by the non-parametric and the parametric bootstrap procedures applied to estimation by analogies method and the confidence intervals derived by a traditional algorithmic method, namely the linear regression model. Specifically, two linear regression models were built for each data set, the first based on the original variables and the second one based on transformations of the variables.

Table 4. Description of the attributes of the Abran-Robillard data set.

Attribute	Description
DETIF	Number of Data Element Type for the Internal Logical Files
GREIF	Number of Record Element Type for the Internal Logical Files
DETEF	Number of Data Element Type for the External Logical Files
GREEF	Number of Record Element Type for the External Logical Files
DETIP	Number of Data Element Type for the Inputs
GREIP	Number of File Type Referenced for the Inputs
DETOP	Number of Data Element Type for the Outputs
GREOP	Number of File Type Referenced for the Outputs
DETIQ	Number of Data Element Type for the Inquiries
GREIQ	Number of File Type Referenced of the Inquiries

At this point, we must emphasize that estimation by analogy was compared by Shepperd and Schofield (1997) against regression models on the basis of the MMRE and Pred(25) values. In this study, nine industrial data sets were used and, in general, analogy appeared to give better results than regression.

In the present study, we used two data sets: (a) the Albrecht data set with 24 projects and 4 attributes (independent variables) as described in Example1 (Section 4) and (b) a data set of 21 projects from a major Canadian financial organization with 10 attributes described in Table 4. The complete data set was given by Abran and Robillard (1996) where the specific attributes were used to build a linear regression model. The dependent variable is denoted by WE and represents actual effort days.

The results of the present study were obtained by using the jackknife procedure for all estimation methods, including the regression models. Therefore, after removing a project from the data set, a regression model was generated using all the remaining projects and this in turn was used to provide an estimation along with a confidence interval for the work effort of the removed project. The confidence intervals for the regression predictions are given by certain formulas and they are available from every known statistical package. In the estimation by analogy procedure, we used in all cases standardized variables of the data sets, while the estimations were based on 2 analogies. In the parametric bootstrap procedure, where we need to sample from a multivariate normal distribution, we used transformations of the original data sets. Specifically, for the Albrecht data set we used the logarithmic transformation for the EFFORT and the independent variables IN, OUT and FILE and the square root transformation for variable INQ. For the Abran-Robillard data set, we used the square root transformation for the dependent variable WE and all the independent variables of Table 4. The same transformations were used in order to build the second regression model.

The results of the study are presented in Tables 5 and 6 and in Figures 7–16. In the last row of Table 5, we can see from the values of MMRE and Pred(25) that estimation by analogy gives better results compared to both regression models in the case of Albrecht data set. In the case of the Abran–Robillard data set, we can see from Table 6, that the analogy method performs better than the regression model with the original variables, while the regression model built on the transformed variables gives the best values of MMRE and Pred(25). The percent denoted by INC, shows the percentage of confidence intervals containing the

Table 5. Estimation and confidence intervals for the Albrecht data set.

Rank	*Actual effort*	*Estimation by analogy*	*95%CI non-parametric bootstrap*	*95%CI parametric bootstrap*	*Estimation by regression (original variables)*	*95%CI regression (original variables)*	*Estimation by regression (transformed variables)*	*95%CI by regression (transformed variables)*
1	0.5	3.9	2.9–8.0	1.6–10.9	−3.8	−8.6–1.0	4.0	2.6–6.0
2	2.9	4.3	0.5–8.0	1.2–14.3	1.7	−2.7–6.2	4.4	2.8–6.9
3	3.6	8.1	5.1–10.6	3.4–34.9	8.8	3.4–14.2	12.2	8.3–18.1
4	4.1	2.7	0.5–7.5	1.1–14.3	−1.1	−6.9–4.8	2.7	1.3–5.8
5	4.9	2.3	0.5–8.9	1.2–14.6	2.6	−2.8–7.9	3.9	2.4–6.2
6	6.1	6.8	2.3–10.0	1.5–16.6	−3.8	−9.2–1.5	3.9	2.1–7.4
7	7.5	18.5	13.6–21.1	3.2–34.4	15.6	12.0–19.3	13.2	10.1–17.2
8	8.0	1.7	0.5–7.0	1.0–11.2	−4.2	−8.8–0.4	2.4	1.4–4.1
9	8.9	7.9	4.1–15.8	1.6–21.3	10.6	4.2–16.9	5.0	3.1–8.1
10	10.0	4.9	3.6–7.1	2.0–23.5	−0.2	−7.0–6.7	5.3	2.6–10.8
11	10.8	12.4	4.9–15.8	2.2–25.5	18.3	11.6–25.0	6.6	3.9–11.4
12	11.1	5.5	2.9–11.8	2.4–25.7	9.1	2.2–16.1	7.6	3.8–15.5
13	11.8	15.6	12.9–28.8	4.8–52.8	24.8	16.0–33.5	23.4	10.6–51.5
14	12.0	20.9	3.6–38.1	5.6–65.0	31.7	24.9–38.4	23.0	14.0–37.9
15	12.9	20.1	9.7–21.1	5.6–54.5	27.5	23.2–31.7	24.2	17.6–33.2
16	15.8	8.2	7.5–19.0	2.6–33.6	17.7	12.6–22.7	11.2	8.0–15.7
17	18.3	14.3	7.5–21.1	3.3–38.3	18.6	12.3–24.9	12.0	6.5–22.2
18	19.0	17.0	7.5–21.1	4.9–55.1	27.4	23.4–31.3	21.9	16.1–29.7
19	21.1	16.0	7.5–19.0	5.0–48.4	26.2	21.6–30.8	19.8	14.2–27.8
20	28.8	15.1	9.7–18.3	2.6–29.5	7.6	−0.5–15.7	7.5	2.5–22.4
21	38.1	36.6	3.6–61.2	8.1–93.6	39.4	27.1–51.8	36.1	21.1–61.9
22	61.2	25.1	12.0–38.1	12.2–129.4	56.2	46.6–65.9	58.3	34.9–97.3
23	102.4	49.7	16.0–105.2	12.9–184.4	82.1	54.3–109.9	89.7	31.1–258.9
24	105.2	81.8	12.0–102.4	10.6–144.1	45.6	16.8–74.4	48.8	18.5–129.2
		MMRE:73% Pred(25):33%	INC:58%	INC:96%	MMRE:103% Pred(25):33%	INC:50%	MMRE:79% Pred(25):25%	INC: 63%

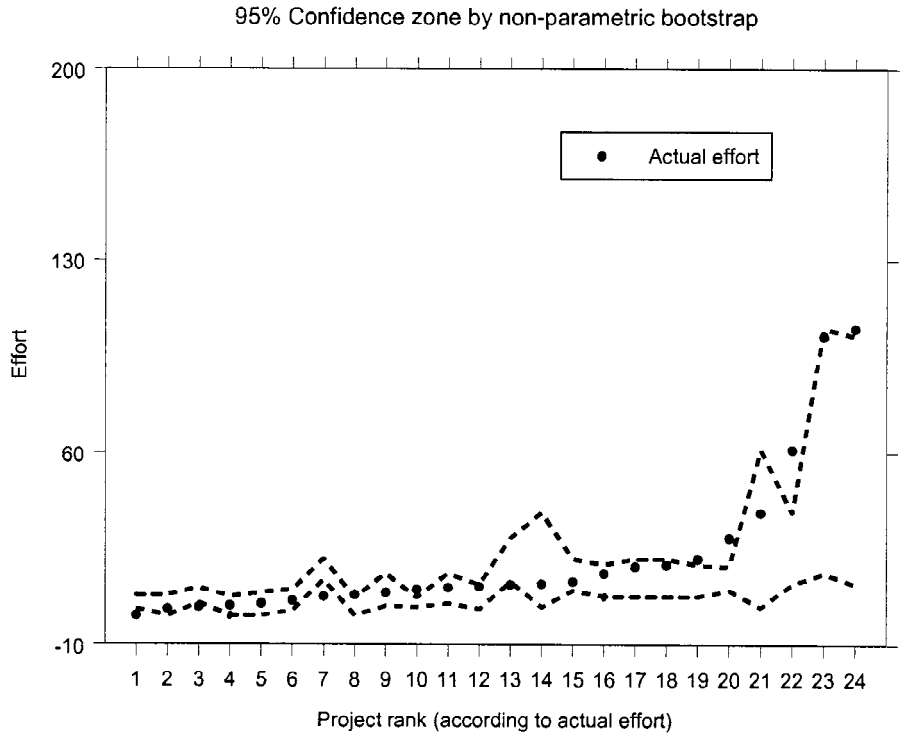

Figure 8. A 95% confidence zone for estimation by analogy using non-parametric bootstrap (Albrecht data set).

actual effort and is an estimation of the probability that the actual effort will fall in a 95% C.I. using the specific method of prediction. From both tables, we can also see an obvious disadvantage of linear regression based on the original variables, i.e. the negative values in the estimates and in the lower bound of confidence intervals. We note that the projects in both tables have been sorted in ascending order according to their actual efforts.

Figure 7, shows in four different panels the actual efforts of the Albrecht data set arranged in ascending order together with the estimations obtained by analogy and regression models. Figures 8, 9, 10 and 11 present the actual efforts of the same data set along with the 95% confidence intervals by non-parametric bootstrap, parametric bootstrap and the two regression models respectively. For better interpretation, we present all the lower and all the upper bounds connected with a dashed line forming in every graph a 95% *confidence zone*. Figures 12, 13, 14, 15 and 16 are the same as before but for the Abran–Robillard data set.

As we can see from the figures, in the case of the Albrecht data set, the parametric bootstrap and the regression model based on the transformed variables provide in general wide confidence intervals. On the other hand, the non-parametric bootstrap intervals are quite similar to the ones obtained by regression based on the original variables. In the Abran–Robillard data set, both regression models give wide confidence intervals.

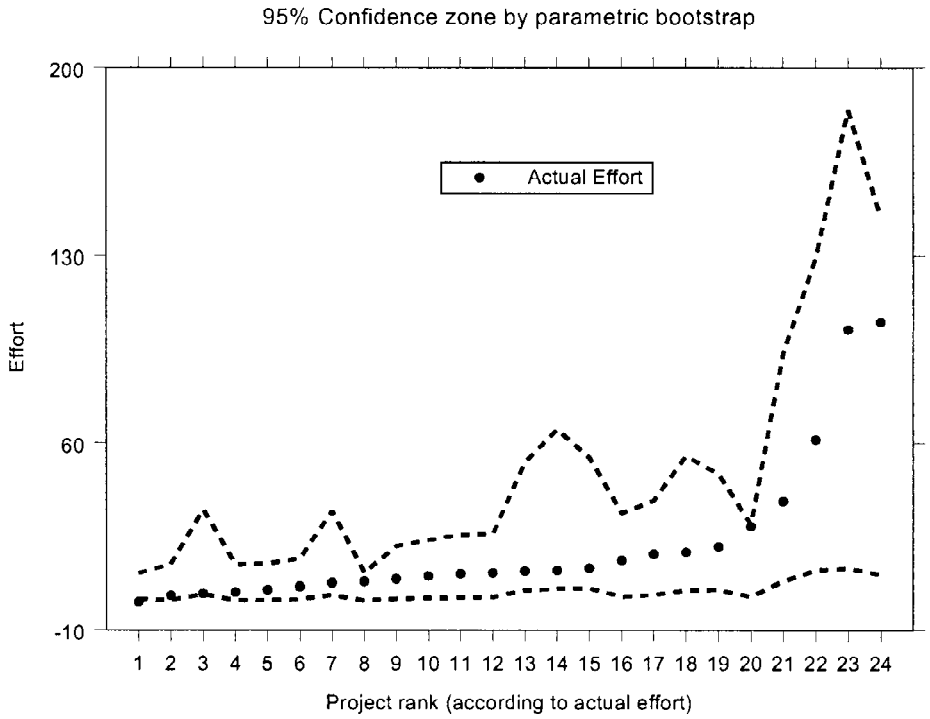

Figure 9. A 95% confidence zone for estimation by analogy using parametric bootstrap (Albrecht data set).

We would like to remind the reader that the obtained results serve only for the comparison between the above mentioned models and are not meant to demonstrate the validity of analogy based cost estimation.

8. Applying Statistical Simulation within the Analogy Based Estimation Process

In this section, we discuss certain aspects of the practical application of estimation by analogy combined with bootstrap. In Shepperd and Schofield (1997) the authors describe the following steps, needed to introduce estimation by analogy in a software development organization:

1. *identify the data or features (attributes) to collect*
2. *agree data definition and collection mechanisms*
3. *populate the case base*
4. *tune the estimation method*
5. *estimate the effort for a new project*

Table 6. Estimation and confidence intervals for the Abran–Robillard data set.

Rank	*Actual effort*	*Estimation by analogy*	*95%CI non-parametric bootstrap*	*95%CI parametric bootstrap*	*Estimation by regression (original variables)*	*95%CI regression (original variables)*	*Estimation by regression (transformed variables)*	*95%CI by regression (transformed variables)*
1	52	206	106–225	88–357	127	60–195	76	54–102
2	69	98	52–298	85–333	165	71–259	103	74–137
3	143	147	52–278	67–331	176	114–238	126	99–155
4	187	124	52–313	68–342	140	58–222	193	146–246
5	195	294	69–400	74–338	199	144–254	172	145–201
6	198	298	169–400	111–401	180	−1–361	198	133–276
7	225	98	52–191	61–332	220	128–311	217	160–282
8	229	397	286–417	113–459	317	160–473	225	152–312
9	360	272	132–400	114–426	260	183–337	303	256–353
10	363	286	195–400	114–444	419	255–584	418	312–539
11	369	380	134–400	158–539	775	118–1432	603	331–954
12	377	323	195–416	85–409	315	198–432	318	234–414
13	400	278	165–360	88–357	254	186–321	301	258–347
14	416	350	229–471	125–489	575	403–746	584	482–696
15	418	446	143–531	130–487	631	387–875	434	276–627
16	428	501	388–531	185–609	523	269–778	378	240–547
17	468	365	198–531	124–498	682	302–1063	359	144–670
18	471	388	251–531	118–441	339	240–438	382	317–453
19	525	445	307–471	140–486	866	−91–1824	834	460–1260
20	531	444	359–471	170–563	434	364–504	566	510–625
21	544	417	198–501	192–600	778	536–1020	716	526–935
		MMRE:40% Pred(25):62%	INC:67%	INC:76%	MMRE:43% Pred(25):38%	INC:71%	MMRE:22% Pred(25):71.4%	INC:76%

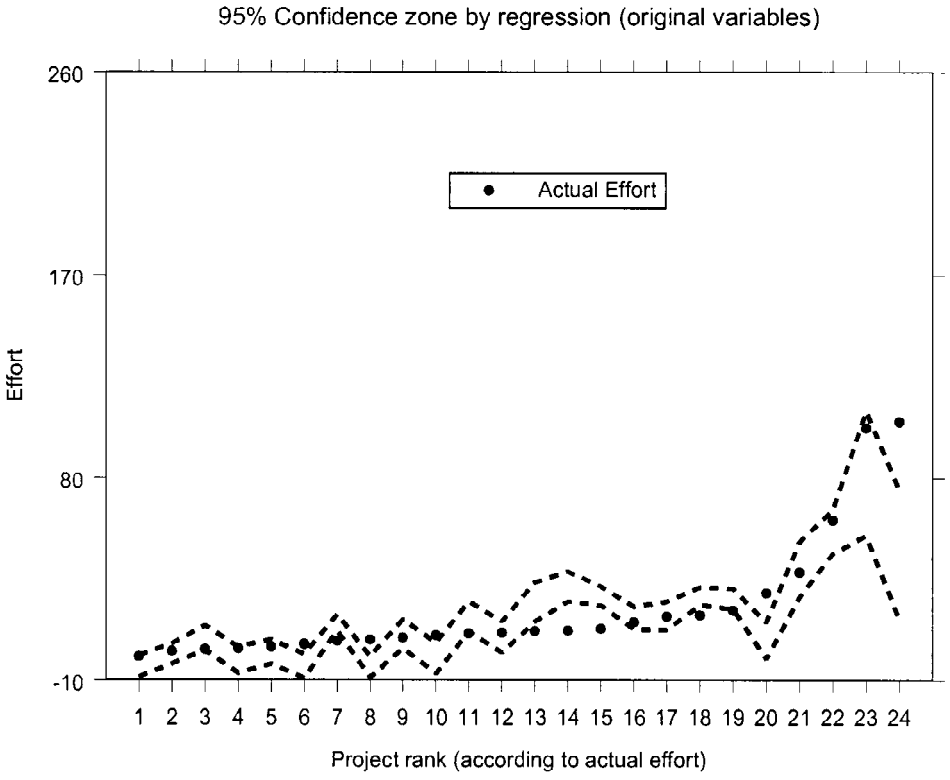

Figure 10. A 95% confidence zone for estimation by regression with the original variables (Albrecht data set).

According to this sequence of actions, statistical simulation is involved mainly in the last two steps. During the fourth step, the parameters of the analogy procedure must be decided. Subsequently, using the approach described in this paper, the estimation accuracy is thoroughly assessed. Poor prediction indicators suggest the revision of the selected parameters and may lead in the repetition of the previous steps. Finally, the calculation of interval estimates using bootstrap enhances the last stage of the above process. Software project managers may take advantage of interval estimates in order to reduce the risks related to project cost and schedule overruns.

Standard practical rules apply in the first three steps concerning attribute identification and collection (attributes should affect development effort, should be measurable, etc.). However, these initial steps may present serious problems, such as the appearance of significant attributes that can not be collected for all cases, or failure to apply coherent collection mechanisms across the organization. Problems may also arise in firms where the development tools and technology are changing fast, since such development environment render obsolete the previously collected data. Particular attention is needed when new projects

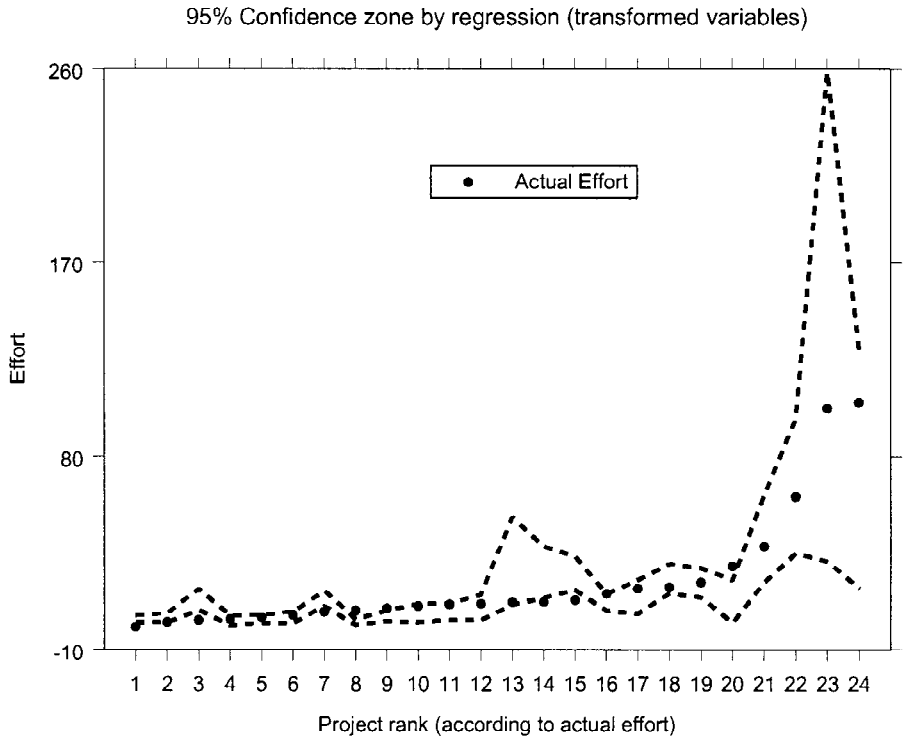

Figure 11. A 95% confidence zone for estimation by regression with transformed variables (Albrecht data set).

include the reuse of existing code, since this modern approach will affect dramatically new code size and project productivity.

The growth of the historical database is another issue. Shepperd and Schofield recognize that there are tradeoffs between the size of the data set and homogeneity. They suggest alternative strategies to cope with this problem. For example, they suggest dividing highly distinct projects into separate data sets. According to their experience, data sets of 10–12 projects already provide a stable basis for estimation.

Our approach helps in dealing with these problems, since it provides the means for tuning and monitoring the accuracy of the predictions through the analysis of the case data set. In practice, the software organization may assess the accuracy of the method each time a major technological change is experienced. If the results are satisfactory, the organization may decide to continue estimating on the basis of the existing data set. Otherwise, the five set-up steps must be repeated and data that are more suitable to the new technology must be collected again while on going projects are gradually completed. The organization may also repeat the analysis to assess the impact of the growth of the dataset on the accuracy of the predictions.

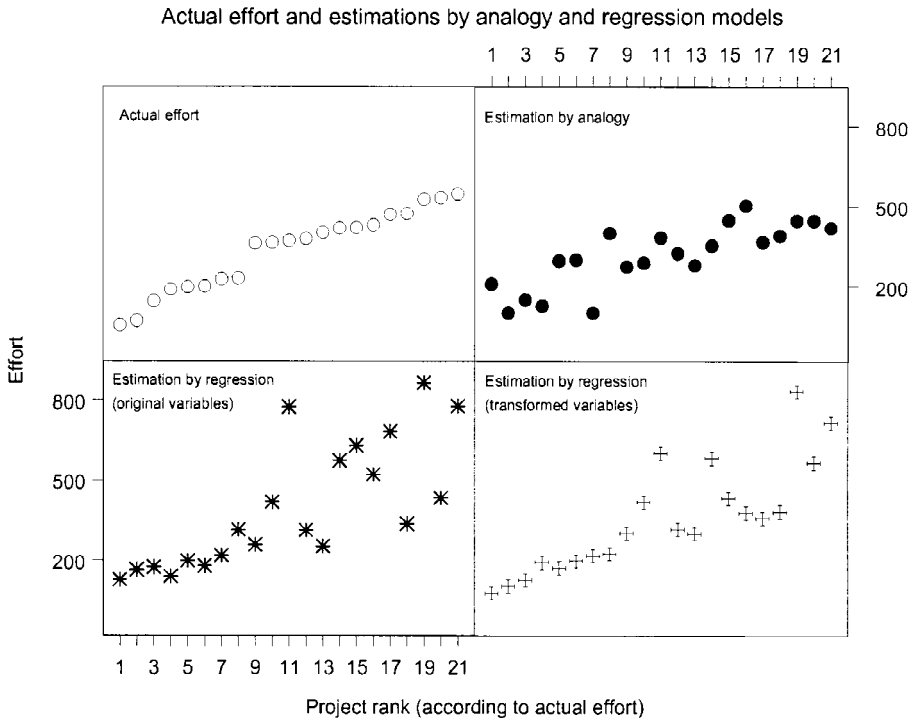

Figure 12. The actual efforts of the Abran–Robillard data set with estimations by analogy and regression models.

The reader should refer to Shepperd and Schofield (1997) for further discussion on the advantages and drawbacks of estimation by analogy and comparison with traditional estimation methods, such as the use of regression models. In general no method may be rejected: regression models seem more suitable for relatively large datasets where strong data relationships are detected, while analogy may be better for relatively small datasets that do not permit the generation of sound statistical models. On the other hand, the success of both methods is subject to the quality of the data and the assumptions used, and to correct interpretation of the estimates produced. Regression models and analogy based estimation should be seen as complementary methods: a software organization should consider both when deciding its cost estimation strategy. Detailed multisource cost and schedule estimation is also recommended by Boehm (1991) as a risk management technique addressing the problem of unrealistic schedules and budgets.

9. Automating Statistical Simulation

In this section, we discuss shortly the automation of the calculations presented above. The analogy method is already automated through ANGEL (ANaloGy Estimation tooL,

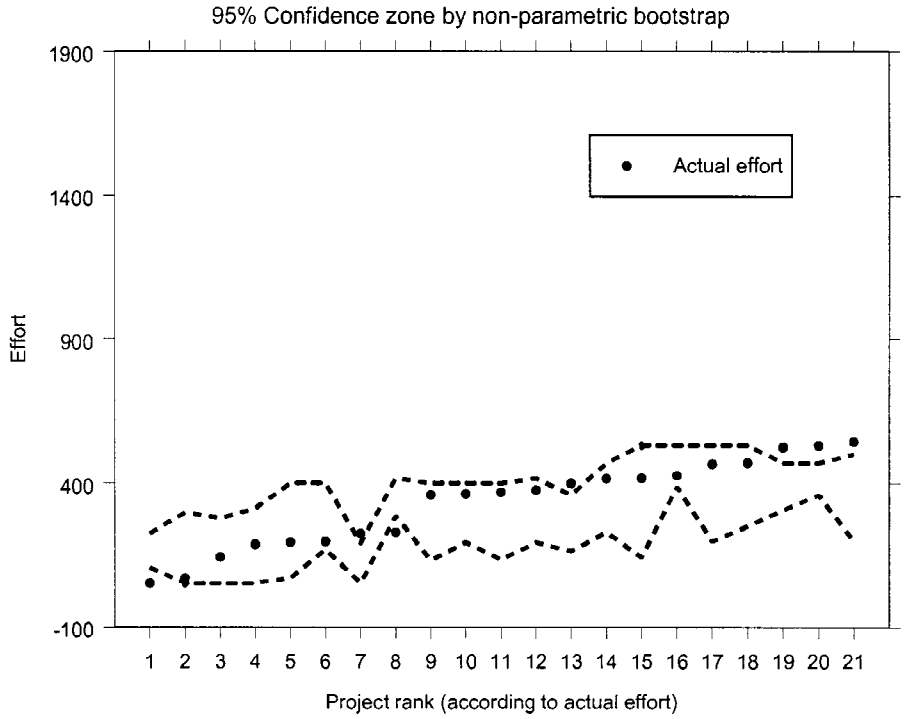

Figure 13. A 95% confidence zone for estimation by analogy using non-parametric bootstrap (Abran–Robillard data set).

Shepperd et al., 1996), a software tool that supports the stages of data collection and effort prediction, and provides some tuning facilities. Bootstrap has been applied on the data sets of this paper through the help of the widely used statistical package S-Plus. The authors have prepared routines written in the proprietary statistical language of S-plus. However, a typical software practitioner may find cumbersome the application of the statistical techniques such as the ones described here. It is evident that a software tool supporting statistical simulation for cost estimation will help the dissemination of the method in the software industry.

The entire procedure involving calibration, estimation by analogy and non-parametric bootstrap, can be easily programmed in every known programming language since the algorithm includes only trivial calculations, random number generation and data sorting. A straightforward approach may be the implementation of the statistical routines in a stand-alone tool, e.g. by incorporating them into ANGEL.

On the other hand, the parametric bootstrap needs tests for normality and decisions on the transformation of the original attributes. It also needs the implementation of an algorithm providing random samples from the multivariate normal distribution. A reasonable approach for complete automation is the integration of a tool performing typical analogy estimation tasks, such as project data collection and estimate generation with a statistical

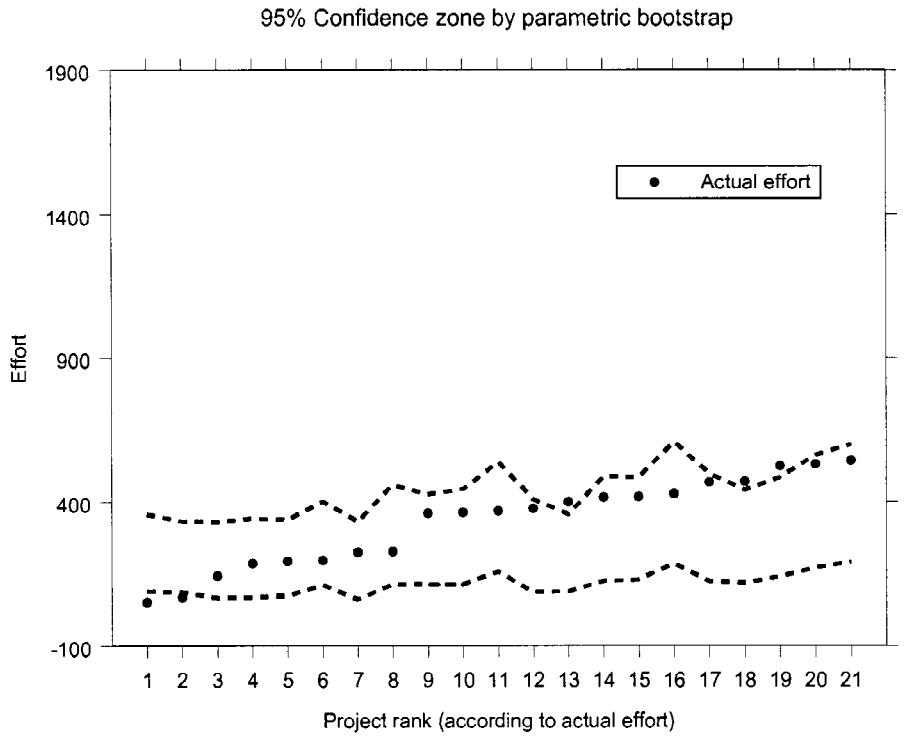

Figure 14. A 95% confidence zone for estimation by analogy using parametric bootstrap (Abran–Robillard data set).

package providing all necessary routines for parametric bootstrap calculations (such as S-Plus). Estimation tools are usually provided with easy-to-use interfaces that must be enhanced with commands implementing the necessary tasks for bootstrap, such as parameter selection (e.g. number of analogies, sample size) and calculation of confidence intervals. Statistical packages usually can be programmed only through proprietary languages. Integration may be accomplished through a command interface that will translate the bootstrap specific commands into script files, containing the necessary calls to the statistical routines, and through auxiliary commands managing files and variables. The interface will invoke the statistical tool with the appropriate script file and will also manage and read back the statistical results into the main estimation module.

10. Conclusions and Future Research

In this paper we have proposed the use of a statistical simulation procedure in order to improve the applicability and the reliability of the estimation by analogy method for software

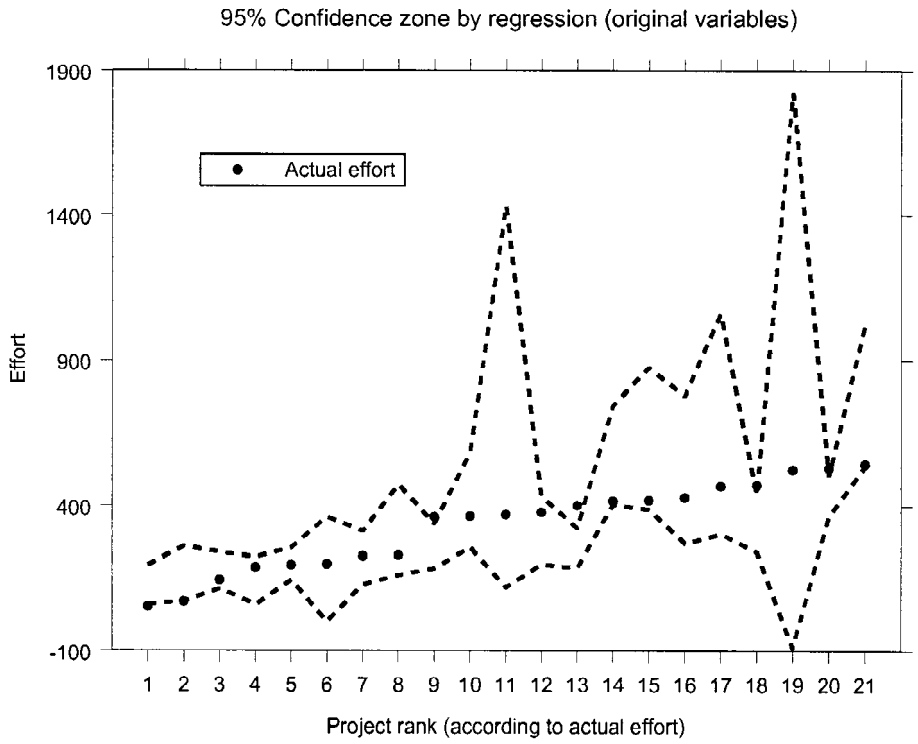

Figure 15. A 95% confidence zone for estimation by regression with the original variables (Abran–Robillard data set).

projects. In particular, we have explored the problem of determining the optimum method parameter configuration before application. Such method parameters may be the distance between projects, the number of analogies and the statistic used to predict efforts. We have also shown how interval estimates may be generated both with and without assumptions about the theoretical distribution of the data set. We believe that in this way we have contributed to the tuning phase of the method and have helped in rendering the method more attractive for practical, risk oriented, software cost estimation.

The ideas presented can be automated and the authors have initiated the development of a software tool that will support the user in the application of the bootstrap method for cost estimation. Nevertheless, as in all automatic software estimation approaches, the application of the method is not without problems. The organization applying analogy should review the accuracy of the method at various time periods or even avoid using it when projects with radically new characteristics must be estimated. The user should carefully assess the quality of his/her historical project cost database in order to be able to have a sound idea about the reliability of the analogy based estimates.

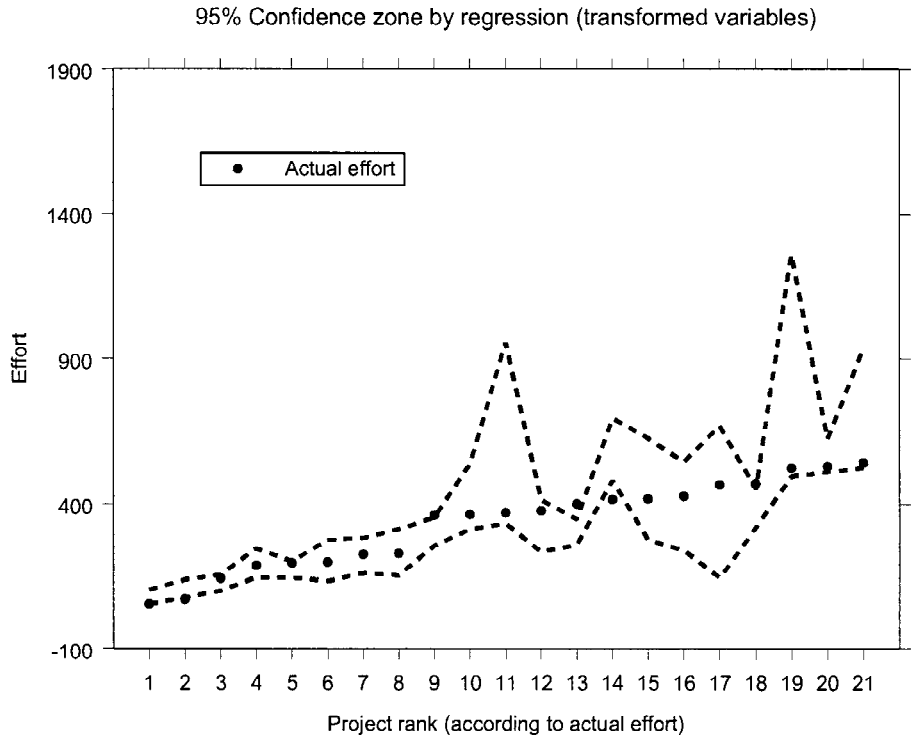

Figure 16. A 95% confidence zone for estimation by regression with transformed variables (Abran–Robillard data set).

In the future we plan to apply the ideas presented in this paper on other cost data sets, as well as the COCOMO data base (see Boehm, 1981) in order to fully explore the issue of similarity between projects with nominal attributes. We would also like to experiment with distance metrics that require a preliminary analysis of the attributes in the cost data set, such as the scaled Euclidean distance.

Another point of interest is the determination of a threshold distance between projects to decide whether analogy based estimations should be applied or not. Such information will add credibility to the entire approach, helping its dissemination among software developers. It is also widely accepted that, whenever possible, it is preferable to apply more than one technique (e.g. combine expert judgement and an algorithmic cost model, or more than one algorithmic cost models) and compare and combine their results to reach the final estimate. Therefore, we would like to investigate the issue of generation of estimates combining analogy and one or more other estimation methods.

Finally, we would like to exploit the power of the bootstrap methods in other areas of software engineering.

Acknowledgments

The authors would like to thank the two reviewers for their comments, which helped in improving the paper.

References

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

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

Boehm, B. W. 1981. *Software Engineering Economics*. Prentice-Hall.

Boehm, B. W. 1991. Software risk management: principles and practices. *IEEE Software* 8(1): 32–41.

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

DeMarco T. 1982. *Controlling Software Projects*. Englewood Cliffs, N.J.: Prentice Hall.

Efron, B., and Tibshirani, R. 1993. *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Hoaglin, D. C., Mosteller, F., and Tukey, J. W. 1983. *Understanding Robust and Exploratory Data Analysis*. New York: John Wiley.

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

Kitchenham, B., and Linkman, S., 1997. Estimates, uncertainty and risk. *IEEE Software* 14(3): 69–74.

Krzanowski, W. J. 1993. *Principles of Multivariate Analysis: A User's Perspective*. Oxford University Press.

Law, A. M., and Kelton, W. D. 1991. *Simulation Modeling and Analysis*. McGraw-Hill.

Mukhopadhyay T., Vicinanza, S. S., and Prietula, M. J. 1992. Examining the feasibility of a case-based reasoning model for software effort estimation. *MIS Quarterly* 16: 155–171.

Johnson, R. A., and Wichern, D. W. 1988. *Applied Multivariate Statistical Analysis*. Prentice-Hall.

Shepperd, M. J., Shofield, C., and Kitchenham, B. A. 1996. Effort estimation using analogy. *Proc. 18th Int'l Conf. Software Eng*. Berlin, IEEE CS Press.

Shepperd, M. J. and Schofield, C. 1997. Estimating software project effort using analogies. *IEEE Trans. on Software Engineering* 23(12): 736–743.

Silverman, B. W. 1986. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

Venables, W. N., and Ripley, B. D. 1994. *Modern Applied Statistics with S-Plus*. New York: Springer-Verlag.

Lefteris Angelis received his BSc and Ph.D. degree in Mathematics from Aristotle University of Thessaloniki (A.U.Th.). He works currently as a Lecturer at the Department of Informatics of A.U.Th. His research interests involve computational methods in mathematics and statistics, planning of experiments and statistical applications to software engineering.

Ioannis G. Stamelos is a lecturer of computer science at the Aristotle University of Thessaloniki, Dept. of Informatics. He received a degree in Electrical Engineering from the Polytechnic School of Thessaloniki (1983) and the Ph.D. degree in computer science from the Aristotle University of Thessaloniki (1988). His research interests include software evaluation and management, software cost estimation and software measurement. He is a member of the IEEE Computer Society.