ELSEVIER

The Journal of Systems and Software 64 (2002) 57–77

The Journal of Systems and Software

www.elsevier.com/locate/jss

# An empirical study of maintenance and development estimation accuracy

Barbara Kitchenham [a,*], Shari Lawrence Pfleeger [b], Beth McColl [c], Suzanne Eagan [c]

[a] *Department of Computer Science, Keele University, Keele, Staffs, ST5 5BG, UK*
[b] *Systems/Software Inc., 4159, Davenport St. NW, Washington DC 20016-4415, USA*
[c] *Computer Sciences Corporation, 100 Winnendon Road (Vergason Building) Norwich, CT 06360, USA*

Received 20 February 2001; received in revised form 28 May 2001; accepted 27 September 2001

**Abstract**

We analyzed data from 145 maintenance and development projects managed by a single outsourcing company, including effort and duration estimates, effort and duration actuals, and function points counts. The estimates were made as part of the company's standard project estimating process that involved producing two or more estimates for each project and selecting one estimate to be the basis of client-agreed budgets. We found that effort estimates chosen as a basis for project budgets were, in general, reasonably good, with 63% of the estimates being within 25% of the actual value, and an average absolute error of 0.26. These estimates were significantly better than regression estimates based on adjusted function points, although the function point models were based on a homogeneous subset of the full data set, and we allowed for the fact that the model parameters changed over time. Furthermore, there was little evidence that the accuracy of the selected estimates was due to their becoming the target values for the project managers.

*Keywords:* Estimation accuracy; Maintenance estimates; Development estimates; Function points; Effort; Duration

## 1. Introduction

A major goal of project managers and software developers is to produce accurate estimates of the effort and time required to complete a software development or maintenance project. Usually, estimates are made when the project is conceived. When the project is complete, the actual values for effort and duration are compared with estimates to determine estimation accuracy. Many papers have suggested methods to improve estimation, and each proposed estimation process is compared with previously defined estimation techniques to see which one is better on a given project or set of projects. But as far as we know, there have been no studies reporting the accuracy of estimates made as part of the commercial process of determining project cost.

It is important for developers and maintainers to know the accuracy of their own estimation processes, many of which are variations on the theoretical models proposed by researchers or vendors, but some of which are home-grown techniques or tools. After all, the more accurate the estimates, the more commercially successful are the companies building or maintaining the software.

In this paper, we report the results of analyzing a data set that includes not only actual effort, duration and function point counts (Albrecht and Gaffney, 1983), but also estimates made as part of Computer Science Corporation's (CSC's) project costing activities. Both CSC and its clients were under the impression that a function point-based estimation process would significantly improve estimate accuracy. Thus, CSC's motivation for analyzing the data set was to demonstrate the improvement that might be expected if the estimation process were based on a suitable function point model.

This focus on improvement raises two issues:

- why CSC presupposed that its current estimating process was poor, and
- why it believed a function point estimation model would be better.

[*] Corresponding author. Tel.: +44-1782-583413; fax: +44-1782-713082.
*E-mail address:* barbara@cs.keele.ac.uk (B. Kitchenham).

0164-1212/02/$ - see front matter © 2002 Elsevier Science Inc. All rights reserved.
PII: S0164-1212(02)00021-3

The software engineering community's view of human estimation accuracy originates from the model of the relationship between development phase and effort estimation presented by Boehm (1981, Fig. 21-1). This model is intended to encapsulate the uncertainty inherent in predicting the costs of new software applications. Boehm's model suggested that early estimates could be more than 100% inaccurate. There are no comparable models, whether theoretical or empirical, for maintenance projects.

DeMarco (1982) discusses one project manager who thought he was an awful estimator based on a recent fiasco, although the manager offers no details about his actual estimating performance. This lack of information suggests that it is possible for managers' perceptions of estimation accuracy to be dominated by a specific bad memory, rather than a balanced and rational assessment of overall performance. Furthermore, DeMarco noted that most average managers rated themselves as below-average estimators. He suggested that this misperception occurred because people will usually rate themselves as poor at a task they do not perform very often. Although data on estimates had been collected, there had been no systematic assessment of estimate accuracy, so accuracy was assumed to be poor. To address this problem, DeMarco recommended the formation of a specialized estimation group to ensure appropriate levels of practice.

One of the reasons that estimators in industry assume their estimates are poor may be because there have been few empirical studies of actual estimation processes and, in particular, we are not aware of any published data that record the contemporary estimates used when projects were undertaken. For example, Hughes (1997) investigated how people in industry construct estimates, but he did not present any information about how accurate they were.

Another reason is that empirical studies are usually based on demonstrating the value of some algorithmic estimating method or data-intensive tool (for examples, see Boehm, 1981; Kemerer, 1987; Shepperd and Schofield, 1997). Thus, it is easy for estimators in industry to believe that if they do not use algorithmic models or data-intensive estimation processes their estimates will be inaccurate. In addition, there is considerable emphasis on function point size measures in the literature (for examples, see Low and Jeffery, 1990; Matson et al., 1994; Kemerer, 1987). Thus, it easy for industrial practitioners to assume that algorithmic models based on function point measures are bound to out perform any process that utilizes human experience.

However, there is some evidence of the value of the human input to the procedures or tools advocated by researchers and vendors. For example, Stensrud and Myrveit (1998) found in a post hoc estimation experiment that human estimators working with estimation tools produced better estimates than those produced directly by the tools. Moreover, they pointed out that some expert estimators produced good estimates without the aid of tools. These results are consistent with results reported by Vicinanza et al. (1990, 1991). Using the Banker and Kemerer (1989) data set (ten of the projects were used for training and five projects for validation), they found that expert estimators relying only on expert judgment substantially out-performed COCOMO and function point-based estimates. Vicinanza et al. discovered that the human estimators were using case-based reasoning. Furthermore, using a verbal protocol of the estimation process used by the best estimator, the researchers were able to develop a tool that mimicked the experts' reasoning process. Although the tool was not as good as the human estimator in a predictive situation (that is, when estimating the validation set of five projects), it performed better than COCOMO.

In this paper, we present a data set that enables us to investigate the actual accuracy of industrial estimates and to compare those estimates with estimates produced from various function point estimation models. However, we must make it clear that any models derived from the current data set are context-specific. They are based on data from a single company with a specific estimation process (that depends on the particular skills of the estimation staff), and a specific client set. Thus, the population to which any statistical inference can be made is the population of maintenance and development projects undertaken by the specific company for a specific set of its clients. Our analysis may provide support for general software engineering hypotheses, or provide counter-examples that cast doubt on the generality of other empirical study results, but our models, whether predictive or descriptive, will not be directly applicable to other companies.

## 2. Case study context

The data come from CSC and relate to its outsourcing activities maintaining and developing software products on behalf of client organizations. Thus, the projects span different products from different sources. The maintenance projects include the usual categories of maintenance: corrective changes, adaptive changes, preventive changes, and perfective changes (Lientz and Swanson, 1980), other projects are user support activities and development projects. Each type of project is defined as follows:

- *Corrective* projects are performed to identify and correct existing processing, performance, or implementation problems. For example, processing failures might include abnormal program termination or incorrect program output. Similarly, performance failures might be indicated by slow response time or inadequate transaction processing rates. Implementation failures can include standards violations, inconsistencies, or incompleteness in program design or documentation.

- *Adaptive* projects are performed to modify an application because of mandatory or regulatory requirements changes, such as tax code changes, contractual obligations, or changes required to support operating system upgrades.
- *Preventive* projects increase the system's future maintainability by preventing likely errors. These projects can include restructuring code (such as making it more modular) or updating documentation.
- *Perfective* projects incorporate changes to accommodate new or changed user requirements to an existing system. For instance, a perfective change may enhance the user interface or add a new feature to make the application more useful.
- *User support* projects involve short-term consulting assignments, where the user requests reports such as data or usage summaries.
- *Development* projects involve creating a new application or replacing an existing one.

The project manager is responsible for assigning a category to a project. If a project appears to involve more than one category, it is usually broken down into several smaller projects, each of which is in a single category.

CSC's estimation process involves six steps.

1. First, the project manager discusses the scope of the project with an independent estimator to determine which estimation methods might be best-suited to the project. The independent estimators are drawn from CSC's Software Process Group.
2. Next, the project manager reviews the estimate with the independent estimator to verify accuracy and completeness.
3. The project manager and program manager also review the estimate to resolve discrepancies.
4. If necessary, the estimate is then reviewed by the corporation's senior management, since the estimate will form the basis for making financial commitments to the project.
5. The independent estimator documents the basis for the estimate and supplies it to a corporate metrics analyst.
6. A corporate metrics analyst evaluates this estimate in the context of others, and analyzes monthly and annual estimating performance for senior management.

Table 1 lists the eight different estimating methods that can be used; multiple estimates are usually required. From multiple estimates, one is selected by the project manager and the independent estimator to be the basis of the client contract and project budget.

The eight approved estimating methods were chosen based on past success with each technique, determined at a workshop attended by experienced corporate staff from various parts of CSC. Each method was then tailored to particular local environments and calibrated with corporate historical data. Usually, two estimating methods are used when the project is less than 200 h, and three estimating methods are used otherwise. In practice, the most commonly used methods are expert judgment, average, CA-Estimacs and function points. The other methods are seldom used.

Although only the Expert Judgment and Delphi methods are expert opinion-based estimating methods, human expertise affects the estimation process in two other ways:

1. The human estimators must decide which estimation methods to use to create the initial set of two or more estimates.
2. The human estimators must decide which of the initial estimates to use as the basis for the project budget. At this point, the estimators have the option of choosing a particular estimate or constructing an average estimate.

Thus, we regard the current estimation process as a human-centered estimation process, in contrast to a model-centered estimation process such as one based solely on a function point estimating model.

Throughout this paper we refer to the estimate used for the project budget as the *selected* estimate and the other estimates as the *rejected* estimates.

A significant problem can occur when comparing contemporary estimates with model-generated estimates; the contemporary projects estimates can inadvertently or intentionally be used as targets by the project managers. One can argue that, unless project managers are given totally unreasonable budgets, they will usually manage to finish the project within budget constraints. Thus, the contemporary estimates may be more accurate than any post hoc model-based estimates because they became the project targets, acting as self-fulfilling prophecies. Abdel-Hamid and Madnick (1989) contend that by imposing different estimates on a software project "... we would, in a real sense, be creating different projects". This view might be taken to imply that there is no possible way of comparing contemporary estimates with any alternative estimates. However, an advantage of our data set is that we have not only the selected estimates but also the rejected estimates. Thus, we can compare the accuracy of the selected estimates with the accuracy of the rejected ones. This allows us to assess the extent of the "self-fulfilling prophecy" effect on the accuracy of the selected estimates, and, if

Table 1
Approved estimating methods

| Method | Definition |
|---|---|
| Average | This method averages two or more of the estimates prepared for the project using the other methods. The initial choice of estimating methods is made by the Project Manager and the Independent Estimator. Then, some or all of the estimates are averaged, again based on the expertise of the Project Manager and the Independent Estimator. |
| CA-Estimacs | This method is based on a commercial software tool, CA-Estimacs 7.0, that queries the user for project characteristics and applies information from a historical database to develop an estimate. The tool has not been calibrated with the past history of the corporate projects; estimates are made based on the database supplied with the tool. The estimate is expressed both in hours and in function points. The input questions vary according to whether the project is client/server, object-oriented, real-time, information engineering, maintenance or generic. The independent estimator answers the questions posed by the tool after consulting with the project manager. The independent estimator helps the project manager to answer the questions consistently. |
| Comparison | This method compares the target project to other completed projects that were similar in scope and type. A reference project is chosen, and its actual hours are used as a basis for the target project estimate. |
| Delphi | Estimates are developed independently by several experienced application developers. The team leader and independent estimator analyze the individual estimates and calculate the median value; then, they call a meeting to discuss them and reach consensus on a single, final estimate. |
| Expert judgment | An experienced project leader compares the requirements for the current project to the requirements for other projects with which he or she has experience. This comparison is performed by decomposing the requirements into tasks, and estimating the hours for each task based on experience. From this analysis, the project leader estimates the likely effort for the new project. |
| Function points | The proposed system is decomposed into functional units, according to the IFPUG standard 4.0: inputs, outputs, inquiries, logical internal files and external interface files. The function point count for a maintenance update to an existing system is based on assessment of the added functionality and is therefore an estimate of the actual function point value. The application's complexity is calculated and used to modify the initial function points count, yielding an adjusted function points count. The conversion from function points to effort is based on past history at CSC; multipliers are generated to reflect hours per function point.<br>The function point estimates used for preliminary effort estimation are not currently held in the corporate database. The values held in the corporate database are the actual values from completed projects (see Table 19 and Kitchenham et al., 2001). |
| Proportion | This method uses estimates or actuals from one or more phases of an existing project. Then, the current estimate is generated by extrapolating to the total development hours using a standard distribution percentage (such as 3–6% for vision and strategy, 12–18% for business systems design, and 3–7% for integration). |
| Widget counting | This method identifies widgets (repeated characteristics of system development) for the project, counting the number of each and assigning a complexity factor. Past history is used to suggest the number of hours required to produce each widget. The widget estimates are summed. Then, effort for supporting tasks is added to the widget estimate to determine total project hours. Predefined widgets include design, test plans, code, code reviews, unit tests and test reviews. |

necessary, we can compare the accuracy of model-based estimates with the accuracy of the rejected estimates.

Another potential problem occurs during the process of agreeing on the selected estimate. This agreement is made by the project manager and the independent estimator working together, so there is a danger that the project manager may urge selection of the estimate that is easiest for him or her achieve. The CSC data set comprises time and materials projects. There is no incentive for the project manager to beat the first estimate. He or she is expected to meet the final approved budget and schedule estimates. However, we can check for estimate selection bias by investigating whether or not the selected estimate is always the largest estimate (which is clearly the estimate most likely to be achieved).

We realize that some researchers may regard any comparison of selected estimates and model-based estimates as invalid, since human estimators clearly have more information available to them than a simple model. However, we were in an industrial situation investigating whether the human-driven estimation process could be replaced by a simpler model. Thus, a comparison between the contemporary selected estimate accuracy and the post-hoc model estimate seemed the only basis for a rational decision.

CSC's estimating process encourages the production of revised estimates throughout the project. However, this paper concentrates on the first estimate agreed with the client, since it is the basis for determining the client's expectations about how much the project is likely to cost. Thus, estimation accuracy affects the company's bottom line in at least two ways. First, client expectations affect customer satisfaction; wild disparities between estimates and actuals can lead to disgruntled customers. Second, the estimates are often the basis of decisions about resource allocation: which developers will be assigned to which projects. If estimates differ significantly from actuals, then developers will not be available for the tasks that follow soon after the expected conclusion of current tasks.

## 3. The data set

Our data set comprised 145 projects. Effort and duration estimates and actuals, project start and end date, and total adjusted function points are shown in Table 19. The full data set including the function point elements and the rejected estimates is published in Kitchenham et al. (2001).

A "project" represents a specific maintenance change to an existing application or a new product development. (Changes and enhancements requiring less than 200 h of staff time are tracked separately. We do not include them in the data set analyzed here.) The data are collected at various times during a project's lifecycle, and the independent estimator enters them in a metrics repository. All estimating information is stored, even for the rejected estimates. Associated with each estimate (selected or rejected) are the project name, the type of estimating method, the estimated date at which the product will be ready for testing, the estimated date at which the tested project will be delivered, and the estimated hours required to complete the project. Project duration is also estimated, but not as early in the project as project effort. At the end of the project, the independent estimator adds to the repository the actual hours, actual delivery date, the final function point count, and other project-specific attributes.

Periodically, the metrics repository is audited to ensure that the data entered are correct. Schedule and budget information is verified by comparing the data with the project's bi-weekly status reports and monthly project closeout reports. Other data elements are audited as required, usually semi-annually or annually depending on the type of data.

The full corporate data set, implemented in Microsoft Access with additional tools in Word and Excel, comprises more than 145 projects. However, we restricted our analysis to those projects for which actual effort, estimated effort and function point counts were all available. The projects were undertaken between 1994 and 1998.

Kitchenham et al. (1999) developed a model of the factors that influence software maintenance. Their model was intended to identify those factors that need to be reported in order to understand results of empirical studies of software maintenance. Compared with their model, the information available from the corporate data set has a major flaw: it provides no information about the application being maintained. This lack of application knowledge means that we are unable to investigate the impact of factors such as the application's size, age, or quality on any function point-based estimation model.

## 4. The analysis procedure and methods

The main reason for analyzing this data set is to compare the estimates obtained from a function point-based effort estimation model with the contemporary estimates made by the independent estimators. To develop a valid function point model, we need first to identify the possible candidate models and then decide which candidate model provides best estimates.

### 4.1. Function point-based effort estimation models

There are three aspects involved in determining candidate estimation models:

1. deciding on the appropriate independent variables;
2. deciding on the possible functional forms of the model;
3. deciding on the likely structure of the model error term(s).

Once the variables, functional forms and error terms are determined, the data can be analyzed to specify the model. It is important to remember that the appropriate method of statistical analysis depends on all three aspects, not just the first two.

Consideration of the available independent variables suggests three candidate function point models:

*Model 1:* A model using the adjusted function point count as the independent variable.
*Model 2:* A model using the raw function point count as the independent variable.
*Model 3:* A model based on the individual function point counts (i.e. the input, output, logical master file, interface and inquiry counts) and the function point complexity adjustment as a set of six independent variables.

The possible functional form of the model is more difficult to assess. Boehm (1981) suggested that the functional form of a model with a single size measure is:

$$\text{Effort} = \beta_0 \,\text{Size}^{\beta 1} \qquad (1)$$

He also suggested that the parameter $\beta_1$ was greater than 1 implying a diseconomy of scale. In order to estimate $\beta_0$ and $\beta_1$, many researchers have applied a natural logarithmic transformation to the effort and size variables and used least-squares analysis to find a linear model of the form:

$$\text{Ln(Effort)} = \text{Ln}(\beta_0) + \beta_1 \,\text{Ln(Size)} \qquad (2)$$

The logarithmic transformation followed by ordinary least-squares analysis is appropriate if the model error term is multiplicative. That is, Eq. (1) is more properly represented as:

$$\text{Effort} = \beta_0\,\text{Size}^{\beta 1} \times (1+\varepsilon) \tag{3}$$

where $\varepsilon$ is distributed (approximately) Normally with mean $\mu = 0$ and variance $\sigma^2$. This means that the logarithmic transformation leads to a model with an additive error:

$$\text{Ln(Effort)} = \text{Ln}(\beta_0) + \beta_1\,\text{Ln(Size)} + \text{Ln}(1+\varepsilon) \tag{4}$$

Such a model can correctly be analyzed using least-squares regression analysis. Furthermore, it is correct to apply the transformation even if $\beta_1 = 1$, since the transformation corrects for the non-stable variance (i.e. heteroscedasticity) implied by Eq. (3).

However, if the error term in Eq. (3) is additive not multiplicative, the logarithmic transformation is incorrect. The data set should be analyzed untransformed using a non-linear regression method (Ratkowsky, 1983). The CSC data set show a fan-shaped pattern when effort is plotted against size (see Fig. 1), so it would seem that Eq. (3) is the most suitable model.

Other researchers have suggested other forms of relationship. For example, Banker and Kemerer (1989) suggested that individual data sets exhibit mixed economies of scale where smaller projects exhibit economies of scale and larger projects exhibit diseconomies of scale. However, recent research, which used genetic algorithms as a flexible method of model fitting, did not find any significant deviations from a linear model (Dolado, 2001). Thus, unless there is very clear evidence to suggest otherwise, we believe it is appropriate to use least-square regression after a logarithmic transformation if there is evidence of heteroscedasticity. Although this procedure is appropriate when we have a single-valued size measurement, such as the adjusted function point count or the raw function point count, it is not clear that it can be applied to models where the weighted function point elements are treated as separate independent variables.

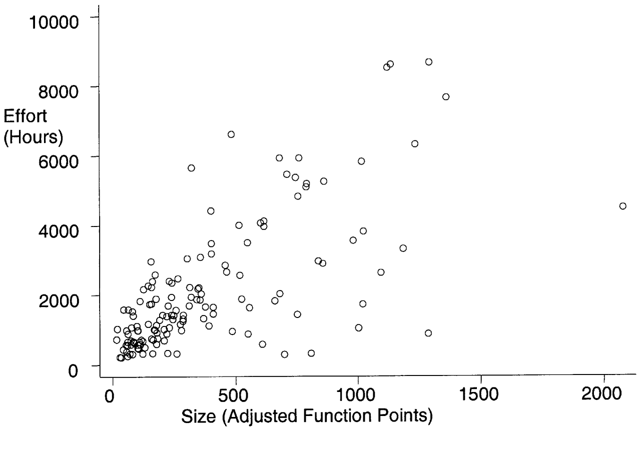

Fig. 1. Scatterplot of effort (hours against size (adjusted function points) excluding outlier project.

Using simulated data sets, Pickard et al. (1999a,b) found that non-parametric median regression was quite robust in the presence of skewed input variables, unstable variances and errors in the independent variables. This result suggests that it is appropriate to use median regression on the raw data to obtain a model based on the function point elements.

### 4.2. Data set refinement

In addition to considering the functional form of the candidate models, it is also necessary to consider the nature of the data set that will be analyzed to specify each model. The current data set is mixed with respect to client, project age and project type, and any of these factors may have a systematic effect on the candidate models. There are two ways of handling factors that have a systematic effect on a model:

1. Include the factor as an independent variable in the model.
2. Partition the data set into homogeneous subsets based on the factor, and build models for each partition.

If the factor has a simple additive effect on the model (that is, it simply increases or decreases the mean effect), the first option can be used. If the impact of the factor is not so simple, the second option is preferable. In either case, the first thing to do is decide whether the factors have an impact on the model.

Initially, it was clear that project 102 was an exceptional point; the total effort required for the project (113 930 h) was 50 times greater than the average value for the remaining 144 projects. In investigating the nature of this outlier, we found the project to be anomalous in many ways. It was CSC's first project using a client-server architecture, using a larger-than-normal team with no prior experience in this work. In particular, the project manager had never managed a client-server project before. The project was also larger than normal in terms of lines of code or function points. CSC estimated the effort for this project using the standard estimating process, however they did not have similar past data to properly calibrate the estimation process. Thus, this project was not used for any of our model building.

The procedure we adopted to assess whether the data set needed to be partitioned was first to build each candidate model on the remaining data set (144 projects), and then to build each candidate model for different data partitions and check whether the models differed.

### 4.3. Selecting the best candidate model

Our main purpose was to compare the estimation accuracy of the most appropriate function point-based estimates with the accuracy of the contemporary estimates made by the estimation group (i.e. the selected estimate). There are several different criteria by which competing models can be judged. In our case we are concerned with two criteria:

1. The sensitivity and stability of the model with respect to the particular data set from which it is derived.
2. The accuracy of the estimates produced by the model.

With separate criteria, it is possible for a model to perform well on one but not on the other. In terms of trade-off, our preference is for a stable model that is not oversensitive to the data set from which it was derived. An overly sensitive model is less likely to give accurate predictions, even though it might be a better fit to a particular data set.

#### 4.3.1. Model sensitivity and stability

The sensitivity of each model (i.e. the extent to which it is affected by high-influence data points) can be checked using residual plots. These plots compare the residual (i.e. the estimate minus the actual) with the estimate, the independent variable or another variable such as project age. For models with Normal errors and error-free independent variables that have been derived using least-squares regression, a variety of more sophisticated plots and tests are available based on "studentized" residuals, leverage statistics and influence analysis (Cook and Weisberg, 1982). However, since one of our models violates most of these assumptions, we have restricted our sensitivity analysis to simple residual plots. We consider a residual plot to be acceptable if there is no systematic pattern in the plot that might indicate that the model is biased or is a poor fit for particular ranges of the independent variable.

Model stability can be assessed by evaluating the changes to model parameters found when the model is derived from different partitions of the database. In general, we would prefer a model that fitted several different data set partitions, so the partitions can be joined to form a larger data set.

#### 4.3.2. Model accuracy

When models are built using different statistical techniques, it is no longer possible to compare standard goodness of fit statistics such as the multiple correlation coefficient or the mean square error; these statistics are meaningful only for ordinary least-squares analysis. Furthermore, when these goodness of fit statistics are derived from fitting a model to transformed data they cannot be compared to those derived from fitting a model to the raw data.

Two measures of estimate accuracy that are popular in the cost estimation community are the mean magnitude relative error (MMRE) and the Pred(25) statistics, both of which were suggested by Conte et al. (1986). The mean magnitude of relative error is calculated from the following formula:

$$\mathrm{MMRE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|\mathrm{actual} - \mathrm{estimate}|}{\mathrm{actual}} \tag{5}$$

where $n$ is the number of projects.

Pred(25) is the proportion of project estimates within 25% of the actual. For example, if Pred(25) is 0.60, then 60% of the estimates are within 25% of the actuals. These statistics give an overall summary of estimate accuracy. However, Pickard et al. (1999a) point out that it is often more useful to look at boxplots of the residuals, where the residual is the difference between the estimated effort and the actual effort. Residual boxplots allow different models to be compared visually. They also show whether or not the estimates are biased (i.e. whether the median differs from zero) and whether the model has a tendency to under- or over-estimate.

To test whether the estimates obtained from one estimation model were significantly better than the estimates obtained from another, we used a paired $t$-test of the difference between the *absolute* residual for each model, as proposed by Stensrud and Myrveit (1998) for the absolute relative error.

The accuracy statistics, boxplots and paired $t$-tests were also used to compare the estimates obtained from the best regression-based model with the selected estimates.

To assess a model's accuracy, our use of accuracy statistics and boxplots is related to which projects are estimated and on what basis. We have a number of alternative methods for determining the accuracy of predicting the actuals:

1. Use the model to predict all the data points that were used to construct it.
2. Remove one data point from the data set at a time, derive the model from the remaining data and predict the excluded data point from the derived model.
3. Separate the data set into two random partitions. Use one partition to generate the model, and then use that model to estimate the effort of projects in the other partition.

In the first case the accuracy statistics relate to the goodness of fit of the model to the data. In the second two cases the accuracy statistics relate to the predictive accuracy of the model.

The third option may appear to be best for assessing predictive accuracy. However, it is viable only if the data set does not need to be partitioned into a number of smaller data sets. Furthermore, although the first option appears significantly worse than the other options, Cook and Weisberg (1982) show that there is a functional relationship between the residuals obtained from the full data set and the residuals obtained when one data point at a time is excluded. They point out that the main difference when comparing models using accuracy statistics based on the residuals derived from the full data set with accuracy statistics residuals derived from the "leave one out" data set, is that the former will favor models that fit the mid range of values best, while the latter will favor models that fit the extreme values best.

Thus, to assess the accuracy of the alternative function-point regression models, we decided to use the simplest procedure, obtaining residuals from the all the data points used to generate the model. Although this approach is less optimal than using a separate validation data set, it is equally unfair to all the competing models.

### 4.4. Comparing selected estimates and regression-based estimates

Before comparing the selected estimates with regression-based estimates, we compared the accuracy of the selected estimates with the accuracy of rejected estimates. If the rejected estimates were significantly less accurate than the selected estimates, this would indicate that the selected estimates acted as self-fulfilling prophesies. If this were the case, we intended to use the best rejected estimate rather than the selected estimate as a basis for comparison with the model-based estimates.

To compare the best regression model with the selected estimates, we used a simple comparison of the residuals obtained from fitting the model with the residual obtained by subtracting the actual value from the selected estimate. However, in addition we used a time-series approach to investigate the accuracy of *predictions*. For each project, we have the start and end date. Using them, it is possible to simulate the growth of the data set to produce a time series of estimates based on the projects that were completed when a new project was about to start. We did not undertake a completely accurate simulation of the estimation process because that would mean generating a different data set for each project (i.e. for each project in the data set, generating a related data set based on those projects that finished before that project started). For simplicity, we have ordered the projects by completion date; for each project, we construct a model using projects that completed prior to the completion of that project, and then use the model to estimate the effort for that project.

There are factors that complicated this analysis procedure:

(1) We needed to decide on the minimum number of projects that can be used to construct a model. In general, we prefer to have 30 projects before constructing a model. However, that number depends on the extent to which the data set needs to be partitioned. Thirty is the preferred minimum data set size both because such data set sizes are not atypical of those reported in the literature and because a sample size of 30 is about the size at which sample statistics begin to converge to their asymptotic properties. When we have many partitions or many projects that need to excluded from the analysis because they represent very small partitions, we must accept a smaller data set size; nevertheless, we did not want to have a data set of fewer than 20 projects.

(2) We needed to decide whether the data set was allowed to grow or whether old projects were to be removed from the data set, keeping constant the number of projects used to construct an estimate. This decision depends on the extent to which the model changes over time. If the model is not affected by project age, the data set can be allowed to grow; if the model is affected by project age, older projects must be excluded when new projects are added.

### 4.5. Duration

CSC's corporate database included preliminary estimates of the project delivery date and actual delivery date, so it is possible to assess the accuracy of schedule estimates. In practice, researchers usually convert schedule estimates and actuals into duration estimates because duration measured in days is easier to analyze than dates. However, the database did not include any information about the expected start date used as the basis for the estimate of project delivery date. Thus, it is not possible to assess duration estimate accuracy unless we assume the project start date was unchanged after the estimate of project end date was obtained. We have made this assumption but it implies that our duration estimate accuracy may appear to be worse than it really is.

Duration estimates are constructed after the effort estimates are reviewed and the selected effort estimate chosen. Thus, there are few alternative duration estimates recorded in CSC's corporate database and we are unable to assess the extent to which duration estimates become self-fulfilling prophecies.

Duration estimation models are usually based on the relationship between effort and duration. Most researchers agree that, at the project level, duration is not linearly related to effort (for example, see Boehm, 1981; Kitchenham, 1992). Thus, to investigate the relationship between effort and duration, we used a logarithmic transformation followed by a least-squares analysis. In addition, we investigated the relationship between adjusted function points and duration, and effort *estimates* and duration.

## 5. Results

Statistical analysis was performed using the Windows version of STATA version 5.0 (STATA Corporation, 1997).

As noted earlier in this paper, preliminary scrutiny of the data set revealed one project to be an extreme outlier. The average effort per project of the other 144 projects was 2344 h, with a minimum of 219 h and a maximum of 15 673 h. The outlier project took 113 930 h, some 50 times larger than average. This project was removed from all model-building and evaluation analyses.

Fig. 1 shows a scatterplot of effort against size (adjusted function points) for the 144 projects. Tables 2–7 show various summary statistics. Table 2 presents summary statistics for the variables to be used in building and testing models: duration, effort, selected effort estimate, adjusted function points, raw function points, the function point technology adjustment value, and the values of each of the function point element counts.

Table 3 shows the number of projects performed for each client. It is clear that most of the projects have been undertaken for one client. Table 4 shows summary statistics for effort, duration and adjusted function points for the two clients with the most projects. Inspection of Table 4 suggests that projects undertaken for Client 1 were smaller but took longer than projects for Client 2. These results suggest that partitioning the data set on the basis of the client may be necessary to obtain a homogeneous data set. This approach will be discussed in a later section of this paper.

Table 5 shows the number of projects of each project type. Most of the projects in this data set were either development or perfective maintenance projects. Table 6 shows summary statistics for development and perfective maintenance projects. The development projects appear to be slightly larger than the perfective projects. This difference indicates that project type is another possible basis for partitioning the data set.

Table 2
Summary statistics

| Variable | Mean | Median | Standard deviation (SD) | Minimum | Maximum |
|---|---|---|---|---|---|
| Duration (days) | 201.31 | 170 | 119.39 | 37 | 604 |
| Effort (h) | 2343.56 | 1544.5 | 2508.55 | 219 | 15673 |
| Estimated effort (h) | 2325.36 | 1726 | 2153.13 | 200 | 14226 |
| Adjusted function points | 405.38 | 259.59 | 386.06 | 15.36 | 2075.8 |
| Raw function points | 394.64 | 267.5 | 363.56 | 15 | 1940 |
| Technology adjustment factor | 36.22 | 36 | 11.43 | 0 | 61 |
| External inputs | 132.2014 | 75.5 | 146.658 | 0 | 850 |
| External outputs | 101.9028 | 59.5 | 119.4836 | 0 | 627 |
| Internal logical files | 59.3125 | 28 | 86.81396 | 0 | 555 |
| External interfaces | 13.81944 | 0 | 56.70185 | 0 | 614 |
| External inquiries | 87.40278 | 49 | 109.9355 | 0 | 618 |

Table 3
Distribution of projects across clients

| Client code | Number of projects |
|---|---|
| 1 | 16 |
| 2 | 115 |
| 3 | 4 |
| 4 | 4 |
| 5 | 4 |
| 6 | 1 |

Table 4
Summary statistics for most frequently served clients

| Variable | Client number | Number of projects | Mean | Median | Standard deviation (SD) |
|---|---|---|---|---|---|
| Duration (days) | 1 | 16 | 238.9 | 198 | 155.4 |
| Duration | 2 | 115 | 193.9 | 163 | 115.6 |
| Effort (h) | 1 | 16 | 2223.31 | 1287 | 3151.74 |
| Effort (h) | 2 | 115 | 2474.2 | 1660 | 2522.74 |
| Adjusted function points | 1 | 16 | 253.1 | 135.6 | 312.35 |
| Adjusted function points | 2 | 115 | 406 | 267 | 395.85 |

Table 5
Distribution of projects across project types

| Project type | Number of projects |
|---|---|
| Adaptive | 4 |
| Corrective | 2 |
| Development | 51 |
| Perfective | 75 |
| Preventive | 1 |
| User support | 1 |
| Unknown | 10 |

Table 7 shows the extent to which different estimating methods were used to produce the first estimates on different projects. It is clear from Table 7 that most of

Table 6
Summary statistics for most frequent project type

| Variable | Project type | Number of projects | Mean | Median | Standard deviation (SD) |
|---|---|---|---|---|---|
| Duration (days) | Development | 51 | 198.9 | 186 | 107.72 |
| Duration | Perfective | 75 | 190.4 | 151 | 114.71 |
| Effort (h) | Development | 51 | 2802.9 | 1650 | 3370.3 |
| Effort (h) | Perfective | 75 | 2117.7 | 1431 | 1904.84 |
| Adjusted function points | Development | 51 | 466.7 | 370 | 391.68 |
| | Adjusted function points Perfective | 75 | 359.7 | 230 | 373.39 |

Table 7
Number of projects estimated using specific estimation methods

| Selected estimate estimation method | Number of projects | Rejected estimate estimation methods | | | | | |
|---|---|---|---|---|---|---|---|
| | | CA-Estimacs | Function point | Expert opinion | Delphi | Comparative | Proportion |
| Average | 34 | 18 | 24 | 32 | 4 | 3 | 2 |
| CA-Estimacs | 1 | 0 | 0 | 1 | 0 | 1 | 0 |
| Comparative | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| Function point | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Delphi | 3 | 1 | 0 | 3 | 0 | 0 | 0 |
| Expert opinion | 104 | 32 | 31 | 0 | 4 | 2 | 0 |
| Proportion | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Widget counting | 1 | 0 | 1 | 1 | 0 | 0 | 0 |
| Total | 144 | 51 | 57 | 37 | 8 | 6 | 2 |

the selected first estimates were expert opinion based (used on 73% of projects including the projects estimated using the Delphi technique). The next most frequent estimation type was an average (used on 24% of projects). Among the rejected estimates where other techniques were used, CA-Estimacs was applied on 35% of projects, and function point methods were used on 40%. Expert opinion was also used as an alternative estimating method on 26% of projects, and contributed to all but two projects that used the average as the selected estimate.

Table 7 also reveals that not all projects had two estimates. In fact, 53 projects had only one estimate made, with 52 of them using expert opinion. A single estimate occurred because the data were being collected before the estimating standards were finalized. Once the estimating standards were defined and distributed, at least two estimates were required; all projects started in 1998 or later had at least two estimates.

We noted earlier that bias might be introduced into the choice of the selected estimate because the project manager would have a vested interest in choosing the largest estimate. However, Table 7 provides evidence of no consistent bias. The use of an average estimate confirms that the selected estimate is not always the largest estimate, since by definition the average of a set of estimates must be less than or equal to the largest estimate. In addition, further analysis of the 52 projects that used expert opinion as the selected estimate but also had an alternative estimate revealed that 33% (17) projects used an estimate smaller than the largest available.

### 5.1. Function-point based regression models

To compare the accuracy of the current estimation process with the estimates produced by an appropriate function point-based model, we compared three possible models:

$$\mathrm{M1}: \mathrm{Ln}(\mathrm{Effort}) = \beta_0 + \beta_1 \,\mathrm{Ln}(\mathrm{AdjustedFP})$$

$$\mathrm{M2}: \mathrm{Ln}(\mathrm{Effort}) = \beta_0 + \beta_1 \,\mathrm{Ln}(\mathrm{RawFP})$$

$$\begin{aligned}\mathrm{M3}: \mathrm{Effort} = {} & \beta_0 + \beta_1 \,\mathrm{Ins} + \beta_2 \,\mathrm{Outs} + \beta_3 \,\mathrm{Files} \\ & + \beta_4 \,\mathrm{Interfaces} + \beta_5 \,\mathrm{Inqs} + \beta_6 \,\mathrm{TAF}\end{aligned}$$

Model 1 uses the adjusted function points as an input variable. Model 2 uses the raw function point count as an input variable. Model 3 uses each of the individual function point elements i.e. Inputs (Ins), Outputs (Outs), Logical Master Files (Files), Interface Files (Interfaces) and Inquiries (Inqs), and the technology adjustment factor (TAF) as input variables. With Model 3, we decided to fit all the variables and then remove non-significant variables one at a time (least significant first) until only significant variables were left. No adjustment was made if the constant term was not significant.

Model 1 and model 2 were constructed using least-squares regression. Model 3 was constructed using non-parametric median regression. Each model was constructed on different partitions of the data set, including:

- the full data set;
- projects performed for Client 1;

- projects performed for Client 2;
- development projects;
- perfective maintenance projects;
- development projects for Client 2;
- perfective projects for Client 2;
- development and perfective projects for Client 2.

The results of these analyses are described fully in Kitchenham et al. (2001). In summary we found that:

1. Model 3 produced significantly different models for each database partition. In each case a different selection of variables were identified as significant input variables. This indicated that Model 3 was extremely sensitive to the particular data set partition and, therefore, could be eliminated as a candidate regression-based estimation model.
2. Model 1 and Model 2 generated very similar models for each data set partition but the adjusted multiple regression coefficient for Model 1 was slightly larger than the adjusted multiple regression coefficient for Model 2, for all data set partitions.
3. The models for Client 1 and Client 2 projects were rather different suggesting the two data sets partitions could not be joined.
4. For Client 2, both Model 1 and Model 2 generated regression models for the 38 development project and for the 67 perfective maintenance projects that were fairly similar to each other, and to the models generated for the combined data set.

We concluded that any comparison between the accuracy of the contemporary estimates and the function point model-based estimates should be based on Model 1 (the adjusted function point model) using the Client 2 development and perfective maintenance projects. To confirm this view, we investigated the residuals for Models 1 and 2 for the suggested partition, to address two issues:

1. To confirm that there was no evidence of systematic problems with either model.
2. To confirm that Model 1 was better than Model 2.

To address the first issue, we plotted the residuals (without transforming back to the raw data scale) first against the fitted values and then against project start age (measured as the elapsed days relative to 1 January, 1995).

Both models had similar residual plots. Plots of the residuals against the fitted values showed evidence of heteroscedasticity with larger residuals showing a larger variation than smaller residuals but no evidence of any high influence points. When the residuals were plotted against a measure of the start date of the project (i.e. days from 1 January, 1995 to the project start date), the larger residuals were associated with the projects that started during 1995 and 1996 as shown in Fig. 2 for Model 1 (Model 2 showed a similar pattern).

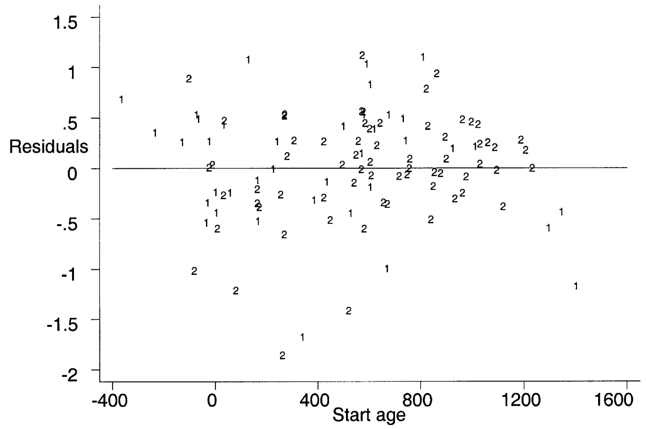

Fig. 2. Residuals vs. age of the project when it started (in days relative to 1/1/95) for Model 1. (1 indicates Development projects, 2 indicates Perfective projects).

We used a paired *t*-test of the difference between the *absolute residuals* for each model to test whether Model 1 was significantly better than Model 2 (Stensrud and Myrveit, 1998). The residuals used for this test were obtained after converting the predictions from each model to the raw data scale. The results of this test together with the MMRE and Pred(25) statistics for each model are shown in Table 8. The *t*-test result confirms that Model 1 provides a better fit to the data than Model 2. The *t*-test result is consistent with the MMRE and Pred(25) values derived from the models; the MMRE is slightly smaller for Model 1 and the Pred(25) is slightly greater.

The relationship between residual size and time shown in Fig. 2 suggests that Model 1 is not stable over time. To check the extent of the instability, Model 1 was applied to four further subsets of the Client 2 development and perfective maintenance projects. The projects were sorted by start age and divided into four groups according to start age:

Table 8
Comparison of Model 1 and Model 2 accuracy (based on predictions after transformation back to the raw data scale)

| Statistic | Model 1 | Model 2 |
|---|---|---|
| MMRE | 0.51 | 0.55 |
| Pred(25) | 0.39 | 0.34 |
| Average absolute residual | 1009.85 | 1059.81 |
| SE of difference | 17.632 | |
| *t* value of difference | −2.193 ($p < 0.01$) using a two tailed test for the alternative hypothesis that Model 1 absolute residuals ≠ Model 2 absolute residuals | |

Table 9
Model 1 results

| Data set partition | Projects | $\beta_1$ | Standard error $\beta_1$ | $\beta_0$ | Standard error $\beta_0$ | Adjusted $R^2$ |
|---|---|---|---|---|---|---|
| Start age <184 days | 26 | 1.026** | 0.1107 | 1.664* | 0.605 | 0.77 |
| Start age 236–587 days | 26 | 0.497** | 0.1212 | 4.381** | 0.6440 | 0.59 |
| Start age 588–842 days | 26 | 0.708** | 0.1164 | 3.686** | 0.6965 | 0.50 |
| Start age >851 days | 27 | 0.834** | 0.1034 | 2.729** | 0.5962 | 0.42 |

** significant at $p < 0.01$, * significant at $p < 0.05$.

- the 26 oldest projects with a start age less than 184 days,
- the 26 next oldest projects with a start age between 236 and 587 days,
- the 26 next oldest projects with a start age between 588 and 842 days,
- the 27 newest projects with a start date greater than 851.

These results are shown in Table 9, which shows that the best-fitting regression line has changed substantially over time, both in terms of the multiplicative parameter $\beta_1$ and the constant $\beta_0$. Since $\beta_1$ changes as well as $\beta_0$, it is not possible to treat start age group as a dummy variable. Furthermore, if start age is used as an independent variable in a simple multiple regression model, its regression coefficient is not significantly different from zero. The only consistency in Table 9 is that the adjusted $R^2$ value appears to have decreased steadily over time.

These results suggests that any comparison between the estimates arising from the current estimation process and the estimates obtained from Model 1 should be based on predicting project $n$ from projects $n-30$ to $n-1$. That is, estimation models based on simulating the growth of the data set over time should allow older projects to drop out of the estimation data set.

### 5.2. Analysis of contemporary estimates

#### 5.2.1. Accuracy of selected estimates

CSC estimators produced one or more estimates at or before the start of each project. In the case where they produce more than one estimate, they either negotiate with the project manager to select one of the estimates to be the project target, or agree with the project managers to use the average value of two or more of the alternative estimates as the project target. The accuracy of the contemporary selected estimates is summarized in Table 10. The values for MMRE and Pred(25) suggest that the performance of the current estimation process across this *non-homogeneous* data set is quite good. Conte et al. (1986) suggest a good estimation process should have an MMRE of 0.25 or less and a Pred(25) of 0.75 or more.

It is interesting to note that the MMRE and Pred(25) are not affected by the outlier, even though the average absolute residual is strongly influenced by the outlier. In addition, the pseudo-$R^2$ (calculated from the correlation between the estimate and the actual) is also affected by the outlier. Table 10 indicates that the *relative* accuracy of the estimate of the outlier project was good but the actual accuracy was very poor. These results cast some doubt on the wisdom of researchers' reporting only relative accuracy measures.

Table 10
Selected estimate accuracy

| Statistic | Full data set | Data set excluding outlier |
|---|---|---|
| Number of projects | 145 | 144 |
| Average absolute residual | 707.35 | 475.74 |
| SD absolute residuals | 2883.955 | 736.425 |
| (Pseudo) Adjusted $R^2$ | 0.99 | 0.88 |
| MMRE | 0.26 | 0.26 |
| Pred(25) | 0.63 | 0.63 |

In the previous section, we found that the function point models were affected by client. However, this was not the case for the selected estimates (see Table 11). Using the Kruskal–Wallis non-parametric analysis of variance (because of the heteroscedasticity), we found no significant difference in accuracy between projects from the two clients or between perfective and development maintenance projects (see Table 12).

Table 11
Estimate accuracy for Client 1 and Client 2 projects

| Statistic | Client 1 | Client 2 |
|---|---|---|
| Number of projects | 16 | 115 |
| Average absolute residual | 514.5 | 476.9 |
| SD absolute residual | 1315.11 | 649.39 |
| MMRE | 0.18 | 0.23 |
| Pred(25) | 0.62 | 0.64 |

Table 12
Estimate accuracy for development and perfective maintenance projects

| Statistic | Development | Perfective |
|---|---|---|
| Number of projects | 51 | 75 |
| Average absolute residual | 538.3 | 448.5 |
| SD absolute residual | 946.39 | 604.26 |
| MMRE | 0.27 | 0.25 |
| Pred(25) | 0.63 | 0.61 |

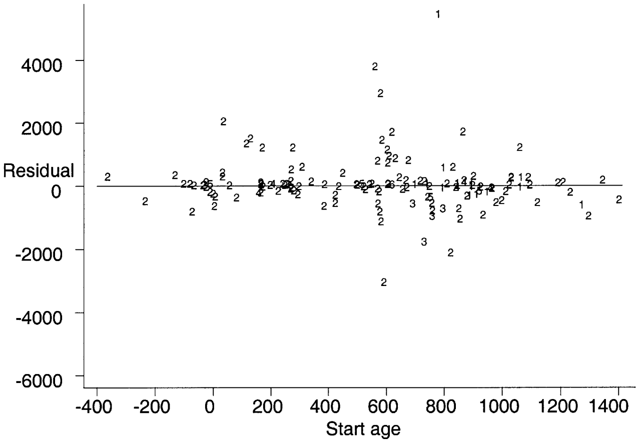

Fig. 3. Residuals vs. project start age for the selected estimates (the symbols 1–6 indicate the client code).

Fig. 3 shows a residual plot for the selected estimates plotted against the start age of the project. This scatterplot suggests that estimate accuracy of the prediction process has not changed much over time, although accuracy was poor for projects that started in the middle of the time period.

Most of the selected estimates were produced by expert opinion or averaging (138 out of 144). The standard accuracy statistics suggested that the expert opinion estimates were marginally better than the average estimates. The MMRE for average estimates was 0.273, compared with 0.254 for expert opinion estimates. Similarly, the Pred(25) for average estimates was 0.529 but was 0.663 for expert opinion estimates. We also calculated the residuals (i.e. the estimate minus the actuals) for each estimate type. The residual boxplots are shown in Fig. 4. The mean value of the absolute residuals of the 34 average estimates was 628.1 and the mean value of the absolute residuals of the 102 expert opinion estimates was 427. A one-way analysis of variance of the absolute residuals confirmed that the expert opinion estimates were more accurate than the average estimates ($p < 0.05$). This is slightly surprising since many estimators would expect the average of several independent estimates to be more accurate than a single estimate.

### 5.2.2. Accuracy of rejected estimates

Before comparing the contemporary estimates obtained by the CSC estimation process with estimates obtained from Model 1, it is necessary to assess how much the accuracy of the contemporary estimates could be attributed to their becoming project targets.

Looking at the data set of 138 expert opinion or average estimates, we found that 85 projects had multiple estimates, that is, they also had estimates that were not selected as the project target. Table 7 shows that expert opinion, CA-Estimacs and Function points were the most popular type of rejected estimate. Only one project of the 85 had a rejected estimate of another type. Fig. 5 shows boxplots of the residuals of each major type of rejected estimate. It is clear that the estimates produced by the function point method are biased (i.e. the median value is substantially less than zero). In addition, the rejected expert opinion estimates seem to be marginally more accurate than the rejected CA-Estimacs estimates. We performed a paired $t$-test of the absolute residuals of the each estimate type compared with its matched selected estimate. The results are shown in Table 13. Since, we are performing three related tests of the hypothesis that selected estimates have the same accuracy as rejected estimates, we have used the Bonferroni procedure to adjust the significance level of each test to be $p < 0.017$ so that the overall significance of the three tests is 0.05 (Rosenberger, 1996).

Table 13 indicates that there is no statistically significant difference between the absolute residuals of the selected estimates and the absolute residuals of the rejected estimates when the rejected estimates were produced by expert opinion or CA-Estimacs. However, the rejected estimates produced by the function point

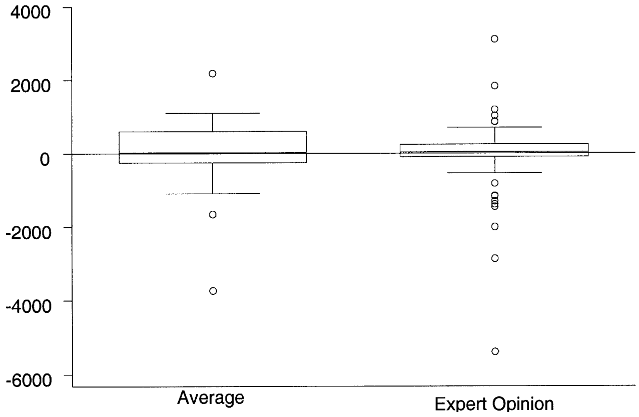

Fig. 4. Comparison of the residuals of expert opinion and average estimates.

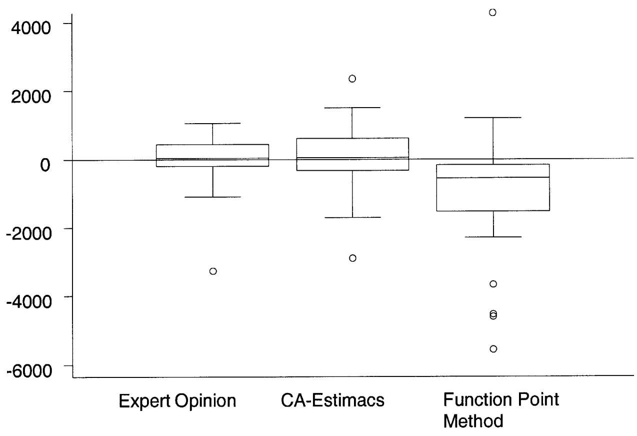

Fig. 5. Boxplots of residuals for rejected estimates.

Table 13
Results of paired $t$-tests comparing selected and rejected estimates ($^*p < 0.017$)

| $t$-test comparison | Projects | Selected estimate absolute residual mean | Rejected estimate absolute residual mean | Standard error of difference | $t$ Statistic |
|---|---|---|---|---|---|
| Selected estimate vs. rejected expert opinion estimate | 32 | 615.69 | 525.59 | 70.781 | 1.273(n.s) |
| Selected estimate vs. rejected CA-Estimacs estimate | 50 | 532.8 | 638.5 | 103.240 | −1.024 (n.s) |
| Selected estimate vs. rejected function point method estimate | 55 | 558.95 | 1158.22 | 137.138 | −4.370* |

Table 14
MMRE and Pred(25) for rejected estimates

| Rejected estimate type | Projects | MMRE | Pred(25) |
|---|---|---|---|
| Expert opinion estimate | 32 | 0.249 | 0.594 |
| CA-Estimacs estimate | 50 | 0.371 | 0.509 |
| Function point method estimate | 55 | 0.760 | 0.218 |

method were significantly worse than the selected estimates. The MMRE and the Pred(25) values for the rejected estimates are shown in Table 14. These compare with an MMRE of 0.259 and a Pred(25) of 0.63 for the 138 selected estimates. We interpret these results to mean that the main reason for significant differences between the selected estimate and the rejected estimates is the method used to construct the rejected estimate. We conclude that the effect of using the selected estimate as a project target is relatively small *in this database.*

### 5.3. Comparison of selected estimates and regression-based estimates

To compare the accuracy of the estimates produced by the current estimation process with the accuracy of estimates produced by Model 1, we restricted our attention to the 105 projects from Client 2 that were identified as development or perfective maintenance projects.

Our preliminary assessment of estimate accuracy is shown in Table 15. The MMRE and Pred(25) values for Model 1 were obtained from the predictions from the four time-based versions of Model 1 shown in Table 9, where the predictions have been transformed back to the raw data scale.

Table 15 shows that the selected estimates are much better than the estimates obtained from Model 1. It is however interesting to note that in comparison with Table 14, the Model 1 estimates are more accurate than the function point estimates made at the start of the project. This may be because the Model 1 estimates are based on post hoc function point counts and the initial estimates are based on function point estimates.

Table 15
MMRE and Pred(25) for the selected estimates and the Model 1 estimates

| Estimate type | Projects | MMRE | Pred(25) |
|---|---|---|---|
| Selected estimates | 105 | 0.236 | 0.63 |
| Model 1 estimates | 105 | 0.441 | 0.48 |

The average absolute residual for the selected estimates was 481 and for the Model 1 estimates was 893.2. A paired $t$-tests confirmed that the absolute residuals of the selected estimate were smaller than the Model 1 absolute residuals ($p < 0.05$).

To assess the accuracy of the function point-based models in a *predictive* situation, we ordered the projects according to start age and predicted the 31st to 105th projects such that the estimate for project $n$ was based on the model calculated from projects $n - 31$ to $n - 1$. The boxplot of the residuals is shown in Fig. 6. The boxplot of the selected estimate residuals is also based only on projects 31–105. The residuals for Model 1 were calculated after converting the logarithmic predictions back to the raw data scale. Thus, the boxplots are directly comparable. It is clear from Fig. 6 that the selected estimates are substantially more accurate than the Model 1 estimates. This is confirmed by Table 16 which shows the MMRE and Pred(25) values for the estimates. Note. As would be expected, the predictive accuracy statistics for Model 1 show in Table 16 is slightly worse than the goodness of fit accuracy statistics shown in Table 15.

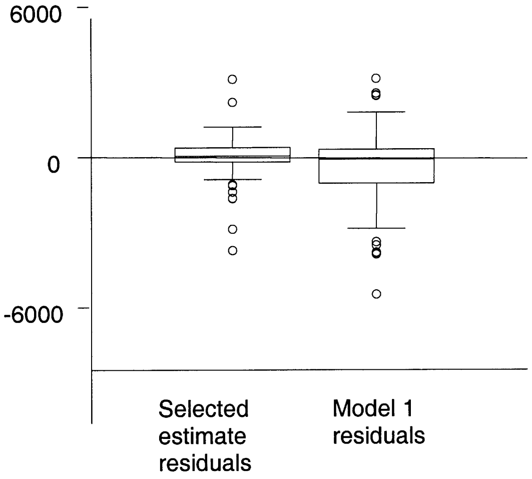

Fig. 6. Residual boxplots for the selected estimates and Model 1 predictions.

Table 16
MMRE and Pred(25) for the selected estimates and the Model 1 predictions

| Estimate type | Projects | MMRE | Pred(25) |
|---|---|---|---|
| Selected estimates | 75 | 0.244 | 0.61 |
| Model 1 estimates | 75 | 0.485 | 0.38 |

The average absolute residual for the selected estimates was 553.1 and for the Model 1 estimates was 994.9. A paired *t*-tests confirmed that the absolute residuals of the selected estimate were smaller than the Model 1 absolute residuals ($p < 0.05$).

### 5.4. Duration prediction

The accuracy of duration estimates made by CSC estimators is shown in Table 17 (note there were three missing duration estimates). These results suggest that duration estimates are slightly more accurate than effort estimates (see Table 10). This may indicate that project managers placed more emphasis on meeting delivery dates than on meeting budget constraints.

As discussed in Section 4.5, we expected to find a non-linear relationship between effort and duration. Fig. 7 illustrates the relationship between effort and duration for Client 2 development and perfective maintenance projects. It shows some evidence that larger projects take longer to produce than smaller. However, for large values of effort, the graph appears to be composed of a series of different lines with slightly different slopes. The different-lines effect could be due to team size differences—all things being equal, a team of three will take less time to complete a project than a team of two, and that effect should remain true for projects requiring different amounts of effort. Fig. 7 also suggests the distinct lines have different slopes, and the lines that correspond to larger teams (i.e. lines associated with projects that are completed relatively quickly) have a smaller gradient than lines that correspond to smaller teams. The effect of the different gradients would in turn lead to an appearance of heteroscedasticity in the scatterplot. These results are consistent with the concept of different staffing level rates used in the SLIM model (Londeix, 1987). However, since the current data set does not include an independent measure of team size, this interpretation of the scatterplot cannot be formally tested.

Table 17
Duration estimate accuracy

| Data set partition | Number of projects | MMRE | Pred(25) |
|---|---|---|---|
| All projects | 142 | 0.23 | 0.65 |
| All projects excluding outlier (project 102) | 141 | 0.23 | 0.65 |
| Development and perfective maintenance for Client 2 (excluding project 102) | 102 | 0.20 | 0.64 |

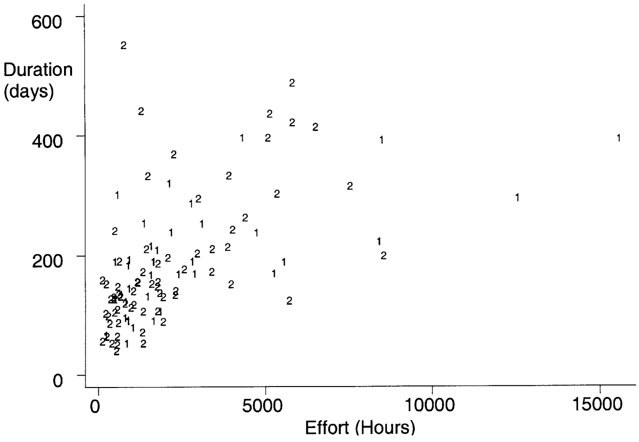

Fig. 7. Duration against for Client 2 development (shown by the symbol 1) and perfective maintenance (shown by the symbol 2) projects.

Treating the problem as a simple curve-fitting problem, we used the logarithmic transformation of the data and the time series partitions used previously. The basic format of the model was:

$$\text{Ln}(\text{Duration}) = \beta_0 + \beta_1 \,\text{Ln}(\text{Effort})$$

Using the four time-based data set partitions, we found that the all four models were quite similar (Kitchenham et al., 2001), so we based our analysis on all 105 Client 2 development and perfective maintenance projects giving the model:

$$\text{Ln}(\text{Duration}) = 2.212 + 0.386\,\text{Ln}(\text{Effort})$$

Both the gradient and the intercept were significantly different from 0 ($p < 0.05$). After transforming the model estimates back to the raw data scale, the MMRE for the model was 0.37 and the Pred(25) was 0.48. In spite of the fact that the simple effort-duration model ignores any team size effect, the goodness-of-fit statistics for the model are moderately good. However, they are substantially worse than the accuracy of the selected duration estimates. The average absolute residual for the selected duration estimates was 42.1 and for the logarithmic duration-effort model was 60.2. A paired *t*-test of the absolute residuals confirmed that the selected estimates were significantly more accurate than the estimates obtained from the duration-effort model ($p <$ 0.001 with 101 degrees of freedom).

In practice actual effort is not known, so it cannot be used as an input to a predictive model. There are two possible surrogates for actual effort: predicted effort and size estimated in terms of function points. We therefore built two more duration models (using the logarithmic format): one used the selected effort estimate, and the other used adjusted function points (Kitchenham et al., 2001). After transformation back to the raw data scale,

Table 18
Results of paired *t*-tests comparing the accuracy of the duration models (* significant at $p < 0.025$ applying the Bonferroni adjustment)

| *t*-test | | Projects | Average absolute residuals | | Standard error of difference | *t* Statistic (One-tailed) |
|---|---|---|---|---|---|---|
| Group 1 | Group 2 | | Group 1 | Group 2 | | |
| Actual effort logarithmic model | Estimated effort logarithmic model | 105 | 60.053 | 64.018 | 1.8857 | −2.103* |
| Actual effort logarithmic model | Adjusted function point logarithmic model | 105 | 60.053 | 65.228 | 3.1219 | −1.658 |

the MMRE for the model based on effort estimates was 0.40 and the Pred(25) was 0.44, the MMRE for the adjusted function point model was also 0.40 and the Pred(25) was 0.44. The MMRE and Pred(25) values suggest that there is nothing to distinguish one model from the other.

We performed paired *t*-tests on the absolute residuals to investigate whether there were any significant differences between the logarithmic effort-duration model and each of the other two models. We used the one-tailed test, so that our alternative hypothesis was that the logarithmic model was more accurate than the other models. The results of these tests are shown in Table 18.

Formally, we are able to reject the hypothesis that the accuracy of the model based on actual effort is no better than the model based on estimated effort. However, inspection of the average absolute residuals in Table 18 shows that the magnitude of the difference between each model is very similar and is also relatively small.

In the previous section, we did not find any major differences with our results when we based our accuracy assessments on predictions rather than goodness-of-fit, except that the MMRE and Pred(25) values were lower for predictions. Thus, we did not perform an analysis of the predictive accuracy of the duration models.

## 6. Discussion

Looking at the accuracy of the selected estimates across all the projects (see Table 10), we see no evidence of the large relative errors reported for new projects by Boehm (1981) and DeMarco (1982). This data set includes maintenance projects and new projects, so it is tempting to infer that estimate accuracy has improved since Boehm and DeMarco's work in the early 1980s. However, we must treat such inferences with caution since, like all project effort data sets, this data set is biased because it does not include information about projects that were abandoned without being completed.

The results described in Section 5.3 show that contemporary estimates produced by the CSC estimation process are considerably better than the estimates produced by a logarithmic adjusted function point model. In terms of the MMRE, the contemporary estimates are nearly twice as accurate as predictions based on a logarithmic adjusted function point model (see Table 16). In fact, the regression model results are worse that this, because the function point model was suited only to 105 of the 145 projects in the data set; for 28% of projects, no function point prediction could be made.

The outlier project (102) presents a particular problem. The contemporary estimate was actually within 30% of the actual, which would be considered good by most estimating standards. However, it was an underestimate of 34 060 h. The total difference between estimate and actual for the other 144 projects was 2612 h (i.e., an average of 18.2 h underestimated per project). An average overrun of 18 h per project is not likely to cause major problems for a company or for its clients. However, an overrun of 34 060 h is likely to cause severe problems. In a fixed cost situation, the company would lose a substantial amount of money. On a time-and-materials contract, the company would both find itself at odds with its client, and find resources intended for other projects being absorbed by unplanned overruns on a single project. In the past, the IT industry has hoped that improved cost estimation would avoid such problems. However, this data set makes it clear that even a good estimation process will not always adequately address an atypical project. For such projects, it is more important to assess the risks of doing the project than to strive to improve estimation accuracy. We recommend that a company embrace a portfolio view, looking at a collection of projects to assess whether a particular project for which they are invited to bid will expose them to undue risk (Kitchenham and Linkman, 1997). Risk must be evaluated not only in terms of direct loss for the specific project, but also in terms of the carryover effects of resource shortages on other projects.

Our results are consistent both with the results of Stensrud and Myrveit (1998), who found that incorporating human expertise improved the accuracy of tool-based estimates, and with the results of Vicinanza et al. (1990, 1991), who found that human experts out-performed COCOMO and a function point model on the Kemerer data set (Kemerer, 1987). These findings emphasize that a human-mediated estimation process can be considerably more accurate than a simple algorithmic model. It is possible that the contemporary estimates became project targets, so their accuracy may have been artificially inflated by becoming "self-fulfilling prophecies". However, the accuracy of the rejected estimates produced by CA-Estimacs and expert opinion were comparable with the accuracy of the selected estimates, so the effect of self-fulfilling prophecies appears to be relatively small in this data set.

Many people in the IT industry are suspicious of human-based estimation, and in particular of estimation that relies on expert opinion. Instead, they desire the claimed objectivity of algorithmic models and tools. But our results show that a human-centered estimating process incorporating expert opinion can substantially outperform simple function point models. If a tool-based solution were required, the CA-Estimacs tool appears promising. On the occasions when it was used it produced estimates of similar accuracy to expert opinion estimates. However, it is not clear why CA-Estimacs, which is a function point based tool, performed so much better than the function point models, particularly since it was not calibrated to the CSC data set. This can be contrasted with the results reported by Kemerer (1987) who found CA-Estimacs performed very poorly when it was not calibrated. In addition, CA-Estimacs was not used for all the projects and it would be necessary to determine whether it could be used on all CSC projects before it could be considered as a viable alternative to the existing process.

Our findings may have resulted from the particular characteristics of the estimation team; the CSC estimators have at their disposal far more information than is incorporated into straightforward function point models. If that is the case, models and tools must incorporate such information to be competitive with human estimators. However, we do not know what the "extra information" is, nor how human estimators use it. Klein (1998) and Pfleeger (2000) have shown that new approaches are needed, as well as new models that incorporate the instinctual ways that humans organize and manipulate the information at their disposal. At a minimum, identifying the essential information used by practicing estimators is an important area for further research.

Kitchenham (1992) observed little difference between the models using adjusted function points and those using raw function points. She suggested that the extra effort and subjectivity involved in collecting the function point technology adjustment factors was not necessary. In this data set, we found that a logarithmic effort model using adjusted function points was significantly more accurate (in terms of the size of absolute residuals) than a logarithmic model based on raw (unadjusted) function points. However, the improvement, although significant, was quite small; the mean absolute residual from the raw function point model was about 5% larger than the mean absolute residual from the adjusted function point model. This level of increased accuracy is small compared with the improvement needed to make a function point model comparable with the selected estimates. If function point models need additional information to improve their accuracy, and bearing in mind that all data collection costs, we believe it may be preferable to drop the costs of collecting data that add only minimal accuracy improvement and look for other factors with a better pay-off. In practice, such an assessment could not be made until appropriate extra variables were identified and their costs of data collection compared with the costs of collecting the technology adjustment factor data.

Many researchers have suggested that duration is related non-linearly to effort (e.g., Boehm, 1981; Kitchenham, 1992). Investigation of a logarithmic model relating duration and effort was consistent with a non-linear relationship. More interesting is the actual scatterplot of duration against effort shown in Fig. 7. It is dangerous to over-interpret a scatterplot because it is easy to observe patterns that are actually spurious. However, Fig. 7 appears to show a series of different lines, which could be due to team size effects. It would be useful to collect a measure of team size more regularly to investigate these effects.

Although it is clear that actual effort is associated with actual duration, actual effort does not provide any help in predicting duration, because it is not known until the end of the project. For this data set, we found that estimated effort and adjusted function points were both reasonable surrogates for actual effort in making duration predictions. However, like effort estimating models, duration estimating models are significantly less accurate than the duration estimates created by the CSC estimators.

## 7. Conclusions

The conclusions we can draw from this study are somewhat limited, because the projects we studied were undertaken by a single company. Thus, we do not expect any of the models presented in this paper to generalize automatically to other maintenance or development situations. We, therefore, restrict our conclusions to demonstrations that a phenomenon may or may not occur in a specific situation. In other words, empirical studies embedded in a single company provide proofs of existence and counter examples but little more general results.

With respect to industry estimating practice, we conclude that a human-mediated estimation process can result in quite accurate estimates. Furthermore, such a process can substantially outperform simple algorithmic models. We would not recommend that any organization replace its current estimation process without

Table 19
Basic project data

| Project | Client code | Project type | Actual start date | Actual duration (Days) | Actual effort (Hours) | Adjusted function points | Estimated completion date | First estimate (Hours) | First estimate method |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | A | 10/12/96 | 107 | 485 | 101.65 | 15/04/97 | 495 | EO |
| 2 | 1 | D | 16/06/97 | 144 | 990 | 57.12 | 19/08/97 | 1365 | A |
| 3 | 1 | D | 01/03/97 | 604 | 13635 | 1010.88 | 30/06/98 | 8244 | EO |
| 4 | 1 | P | 23/06/97 | 226 | 1576 | 45.6 | 06/03/98 | 1595 | D |
| 5 | 1 | D | 20/01/97 | 326 | 3826 | 1022.58 | 01/01/98 | 3828 | A |
| 6 | 1 | P | 10/12/97 | 294 | 1079 | 77.04 | 07/08/98 | 879 | EO |
| 7 | 1 | A | 06/07/98 | 212 | 2224 | 159.6 | 17/02/99 | 2895 | EO |
| 8 | 1 | C | 08/12/97 | 175 | 1698 | 225.54 | 22/05/98 | 1800 | A |
| 9 | 1 | Pr | 05/05/97 | 584 | 1173 | 144.72 | 17/03/98 | 1160 | EO |
| 10 | 1 | D | 18/03/97 | 171 | 1401 | 84.42 | 30/07/97 | 885 | EO |
| 11 | 1 | D | 01/06/97 | 201 | 2170 | 126.42 | 04/12/97 | 2125 | EO |
| 12 | 1 | D | 14/08/97 | 195 | 1122 | 392.16 | 22/01/98 | 1381 | EO |
| 13 | 1 | U | 21/07/97 | 109 | 1024 | 18.9 | 04/11/97 | 1142 | D |
| 14 | 1 | P | 23/06/97 | 263 | 1832 | 112.14 | 01/04/98 | 1895 | EO |
| 15 | 1 | A | 10/07/97 | 165 | 1016 | 210.08 | 14/03/98 | 1339 | EO |
| 16 | 1 | D | 15/03/97 | 46 | 322 | 260.95 | 23/07/97 | 447 | EO |
| 17 | 2 | D | 18/12/95 | 186 | 580 | 609.7 | 12/02/96 | 507 | EO |
| 18 | 2 | D | 26/06/95 | 189 | 1003 | 169.85 |  | 952 | EO |
| 19 | 2 | P | 26/06/95 | 95 | 380 | 56 | 29/09/95 | 380 | EO |
| 20 | 2 | P | 25/09/95 | 53 | 220 | 30 | 17/11/95 | 220 | EO |
| 21 | 2 | P | 07/10/97 | 365 | 2356 | 241.86 | 18/12/98 | 2879 | EO |
| 22 | 2 | P | 19/12/94 | 438 | 1388 | 219.88 | 01/03/96 | 1483 | EO |
| 23 | 2 | P | 11/03/96 | 109 | 1066 | 229.71 | 28/06/96 | 1667 | EO |
| 24 | 2 | D | 05/08/96 | 283 | 2860 | 458.38 | 03/05/97 | 2125 | A |
| 25 | 2 | P | 12/01/98 | 137 | 1143 | 177.63 | 12/06/98 | 1175 | A |
| 26 | 2 | P | 12/05/97 | 102 | 1431 | 287.64 | 05/09/97 | 2213 | A |
| 27 | 2 | P | 09/06/97 | 103 | 1868 | 343.54 | 03/10/97 | 2247 | A |
| 28 | 2 | P | 01/07/97 | 192 | 2172 | 346.8 | 31/10/97 | 1926 | A |
| 29 | 2 | D | 14/08/96 | 219 | 8509 | 1121.48 | 31/12/96 | 5641 | EO |
| 30 | 2 | P | 17/02/95 | 484 | 5927 | 761.08 | 01/12/95 | 3928 | EO |
| 31 | 2 | P | 09/09/96 | 173 | 2663 | 464 | 15/01/97 | 1995 | A |
| 32 | 2 | P | 10/02/97 | 169 | 1425 | 203.01 | 01/03/97 | 2281 | EO |
| 33 | 2 | P | 05/01/98 | 207 | 3504 | 550.14 | 06/06/98 | 3305 | EO |
| 34 | 2 | P | 05/08/96 | 61 | 652 | 86.45 | 30/10/96 | 797 | EO |
| 35 | 2 | P | 25/07/96 | 311 | 7649 | 1362.11 | 08/03/97 | 3922 | A |
| 36 | 2 | P | 12/08/96 | 418 | 5927 | 681 | 10/05/97 | 6809 | A |
| 37 | 2 | P | 26/05/97 | 410 | 6607 | 485.1 | 22/04/98 | 4955 | A |
| 38 | 2 |  | 08/05/95 | 497 | 2591 | 172.96 | 16/09/96 | 1294 | EO |
| 39 | 2 | P | 16/08/96 | 259 | 4494 | 2075.8 | 19/05/97 | 5688 | EO |
| 40 | 2 | D | 24/01/97 | 234 | 4824 | 756.25 | 30/09/97 | 5245 | EO |
| 41 | 2 |  | 01/07/95 | 462 | 5094 | 789.66 | 07/08/96 | 3930 | EO |
| 42 | 2 | P | 21/04/97 | 291 | 3088 | 357 | 03/10/97 | 2562 | EO |
| 43 | 2 | P | 06/08/96 | 116 | 892 | 62.08 | 13/12/96 | 1526 | C |
| 44 | 2 | P | 26/06/95 | 128 | 750 | 157.56 | 01/11/95 | 1018 | EO |
| 45 | 2 | D | 01/04/97 | 185 | 5646 | 322.62 | 17/10/97 | 5646 | CAE |
| 46 | 2 | P | 03/10/94 | 207 | 1532 | 81.34 | 30/06/95 | 1532 | EO |
| 47 | 2 | P | 02/01/95 | 151 | 1280 | 191 | 02/06/95 | 1532 | EO |
| 48 | 2 | P | 24/10/94 | 99 | 313 | 121.52 | 28/02/95 | 314 | EO |
| 49 | 2 | P | 17/06/96 | 61 | 339 | 222.78 | 17/08/96 | 412 | A |
| 50 | 2 | P | 30/10/96 | 101 | 583 | 113.52 | 24/01/97 | 738 | EO |
| 51 | 2 |  | 29/07/95 | 462 | 726 | 15.36 | 15/11/96 | 763 | EO |
| 52 | 2 | P | 07/11/97 | 133 | 1939 | 320.12 | 09/03/98 | 1750 | A |
| 53 | 2 | P | 23/05/96 | 106 | 669 | 84.28 | 06/09/96 | 682 | A |
| 54 | 2 | P | 30/12/96 | 68 | 1413 | 248.88 |  | 1320 | EO |
| 55 | 2 | P | 17/11/95 | 239 | 4115 | 616.32 | 29/02/96 | 3573 | EO |
| 56 | 2 | P | 06/09/96 | 330 | 4009 | 515.07 | 13/06/97 | 2913 | A |
| 57 | 2 | P | 30/01/97 | 37 | 630 | 88.2 | 21/03/97 | 701 | EO |
| 58 | 2 | P | 08/07/96 | 187 | 718 | 115.14 | 27/01/97 | 725 | A |
| 59 | 2 | P | 09/08/96 | 329 | 1584 | 63.84 | 21/02/97 | 1826 | A |
| 60 | 2 | P | 04/10/96 | 120 | 5816 | 1015.98 | 01/04/97 | 5000 | EO |

Table 19 (*continued*)

| Project | Client code | Project type | Actual start date | Actual duration (Days) | Actual effort (Hours) | Adjusted function points | Estimated completion date | First estimate (Hours) | First estimate method |
|---|---|---|---|---|---|---|---|---|---|
| 61 | 2 | P | 07/02/97 | 85 | 2037 | 359.64 | 16/05/97 | 2640 | A |
| 62 | 2 | P | 16/05/97 | 49 | 1428 | 240.84 | 11/07/97 | 2534 | A |
| 63 | 2 | P | 03/08/97 | 152 | 1252 | 285.12 | 02/01/98 | 2231 | A |
| 64 | 2 | P | 11/03/96 | 47 | 655 | 61.2 | 27/04/96 | 1000 | EO |
| 65 | 2 |  | 03/11/95 | 148 | 1318 | 287.28 | 30/03/96 | 1645 | D |
| 66 | 2 | D | 22/03/96 | 141 | 995 | 172 | 13/07/96 | 1067 | EO |
| 67 | 2 | D | 09/09/96 | 235 | 2265 | 144.06 | 16/05/97 | 2270 | EO |
| 68 | 2 | D | 09/09/96 | 298 | 654 | 108.64 | 11/07/97 | 656 | EO |
| 69 | 2 |  | 02/02/96 | 99 | 718 | 165.36 | 11/05/96 | 121 | A |
| 70 | 2 | P | 05/04/96 | 127 | 2029 | 680.9 | 15/06/96 | 1685 | EO |
| 71 | 2 | D | 01/02/96 | 163 | 1650 | 409.4 | 13/07/96 | 2350 | EO |
| 72 | 2 | D | 25/07/97 | 316 | 2223 | 313.95 | 02/05/98 | 2308 | EO |
| 73 | 2 | D | 19/11/96 | 388 | 8600 | 1136.34 | 03/10/97 | 7850 | A |
| 74 | 2 | P | 01/09/97 | 152 | 1884 | 177 | 14/02/98 | 2004 | EO |
| 75 | 2 | D | 23/09/96 | 166 | 5359 | 746.24 | 08/11/96 | 3715 | W |
| 76 | 2 | P | 14/11/96 | 114 | 1159 | 274.92 | 22/03/97 | 1273 | A |
| 77 | 2 | P | 15/07/96 | 82 | 437 | 43.5 | 09/11/96 | 437 | A |
| 78 | 2 | D | 20/10/97 | 123 | 570 | 54.75 | 10/03/98 | 813 | EO |
| 79 | 2 | P | 20/01/95 | 49 | 502 | 130 | 18/03/95 | 900 | EO |
| 80 | 2 | P | 06/02/98 | 183 | 1877 | 525.96 | 21/08/98 | 2475 | A |
| 81 | 2 | P | 12/09/96 | 149 | 1693 | 311.85 | 15/11/96 | 799 | A |
| 82 | 2 |  | 15/10/95 | 370 | 3319 | 1185.08 | 23/03/96 | 2160 | EO |
| 83 | 2 | D | 26/08/95 | 128 | 1557 | 258.24 | 01/01/96 | 1770 | EO |
| 84 | 2 | P | 20/10/95 | 126 | 557 | 60 | 23/02/96 | 760 | EO |
| 85 | 2 | P | 10/10/95 | 200 | 3050 | 303.52 | 23/03/96 | 2588 | A |
| 86 | 2 | D | 26/05/96 | 76 | 1113 | 98.9 | 10/08/96 | 1148 | A |
| 87 | 2 | P | 19/08/96 | 299 | 5456 | 711.9 | 19/04/97 | 4064 | EO |
| 88 | 2 | D | 24/06/96 | 131 | 763 | 182.4 | 23/11/96 | 933 | EO |
| 89 | 2 |  | 30/05/97 | 140 | 2203 | 351.9 | 09/09/97 | 2096 | EO |
| 90 | 2 | P | 18/10/96 | 169 | 3483 | 401.98 | 18/05/97 | 3284 | EO |
| 91 | 2 | P | 14/04/97 | 130 | 2393 | 162.61 | 05/09/97 | 4576 | A |
| 92 | 2 | D | 22/05/95 | 389 | 15673 | 1210.99 | 09/02/96 | 14226 | EO |
| 93 | 2 | D | 26/08/96 | 166 | 2972 | 156.42 | 19/04/97 | 6080 | EO |
| 94 | 2 | P | 17/04/98 | 148 | 4068 | 603.58 | 06/09/98 | 4046 | EO |
| 95 | 2 | P | 04/05/98 | 131 | 698 | 73.92 | 26/09/98 | 649 | EO |
| 96 | 2 | P | 02/09/97 | 144 | 676 | 121.55 | 27/02/98 | 817 | EO |
| 97 | 2 |  | 17/09/95 | 369 | 6307 | 1234.2 | 13/07/96 | 6340 | EO |
| 98 | 2 | P | 03/07/95 | 155 | 219 | 35 | 27/10/95 | 300 | EO |
| 99 | 2 |  | 06/11/95 | 102 | 254 | 61.06 | 26/02/96 | 315 | EO |
| 100 | 2 | P | 03/04/95 | 149 | 324 | 162 | 30/09/95 | 750 | EO |
| 101 | 2 | P | 01/10/95 | 548 | 874 | 1285.7 | 30/09/96 | 898 | EO |
| 102 | 2 | D | 01/09/94 | 946 | 113930 | 18137.48 | 20/12/96 | 79870 | EO |
| 103 | 2 | D | 11/11/96 | 186 | 1722 | 1020.6 | 16/05/97 | 1600 | EO |
| 104 | 2 | D | 10/03/95 | 212 | 1660 | 377 | 08/10/95 | 1702 | EO |
| 105 | 2 | P | 09/10/95 | 84 | 693 | 210.45 | 01/01/96 | 592 | EO |
| 106 | 2 | D | 16/01/95 | 250 | 1455 | 410 | 29/09/95 | 2158 | EO |
| 107 | 2 | D | 28/06/95 | 86 | 988 | 279 | 22/09/95 | 994 | EO |
| 108 | 2 | D | 19/12/94 | 102 | 1940 | 240 | 11/04/95 | 1875 | EO |
| 109 | 2 | P | 10/10/95 | 137 | 2408 | 230 | 24/02/96 | 2527 | EO |
| 110 | 2 | D | 01/11/94 | 87 | 1737 | 150.29 | 27/02/95 | 2606 | EO |
| 111 | 2 | D | 07/11/94 | 291 | 12646 | 1940.68 | 25/08/95 | 12694 | EO |
| 112 | 2 | D | 14/01/94 | 392 | 4414 | 401 | 10/02/95 | 4176 | EO |
| 113 | 2 | D | 13/02/95 | 165 | 2480 | 267 | 28/07/95 | 2240 | EO |
| 114 | 2 | D | 11/09/95 | 88 | 980 | 102 | 18/12/95 | 980 | EO |
| 115 | 2 | D | 23/05/94 | 249 | 3189 | 403 | 27/01/95 | 3720 | EO |
| 116 | 2 | D | 12/12/94 | 186 | 2895 | 857 | 16/06/95 | 2914 | EO |
| 117 | 2 | D | 09/12/94 | 63 | 322 | 69 | 10/02/95 | 360 | EO |
| 118 | 2 | A | 09/10/95 | 192 | 3555 | 980.95 | 03/06/96 | 3700 | EO |
| 119 | 2 | P | 13/02/95 | 123 | 570 | 100.8 | 16/06/95 | 200 | EO |
| 120 | 2 | P | 02/05/97 | 123 | 464 | 105.28 | 15/09/97 | 578 | EO |
| 121 | 2 | D | 13/01/97 | 186 | 1742 | 158.4 |  | 1652 | EO |

(*continued on next page*)

Table 19 (*continued*)

| Project | Client code | Project type | Actual start date | Actual duration (Days) | Actual effort (Hours) | Adjusted function points | Estimated completion date | First estimate (Hours) | First estimate method |
|---|---|---|---|---|---|---|---|---|---|
| 122 | 2 | D | 18/09/98 | 119 | 896 | 219.88 | 30/11/98 | 780 | A |
| 123 | 2 | P | 01/11/97 | 195 | 8656 | 1292.56 | 24/04/98 | 8690 | EO |
| 124 | 2 | P | 07/11/97 | 210 | 3966 | 616.08 | 29/05/98 | 3748 | EO |
| 125 | 2 | D | 04/09/94 | 180 | 989 | 103.4 | 03/03/95 | 710 | EO |
| 126 | 2 | P | 30/05/98 | 238 | 585 | 74.4 | 20/02/99 | 856 | EO |
| 127 | 2 | P | 16/09/97 | 144 | 1860 | 356.31 | 01/12/97 | 2436 | EO |
| 128 | 2 | P | 08/12/97 | 432 | 5249 | 862 | 23/11/98 | 4101 | EO |
| 129 | 2 | P | 27/06/97 | 392 | 5192 | 791.84 | 02/02/98 | 5231 | EO |
| 130 | 2 | D | 01/08/98 | 205 | 1832 | 661.27 | 20/11/98 | 2853 | A |
| 131 | 2 | D | 13/01/95 | 49 | 928 | 179 | 03/03/95 | 1246 | EO |
| 132 | 3 | P | 10/06/96 | 205 | 2570 | 518.4 | 15/01/97 | 2570 | EO |
| 133 | 3 | D | 01/01/95 | 145 | 1328 | 370 | 31/12/95 | 1328 | EO |
| 134 | 3 | D | 02/02/97 | 172 | 2964 | 839.05 | 22/08/97 | 3380 | EO |
| 135 | 3 | P | 21/07/97 | 137 | 1304 | 243.86 | 05/12/97 | 1522 | EO |
| 136 | 4 | D | 02/12/96 | 371 | 1631 | 557.28 | 04/09/97 | 2264 | EO |
| 137 | 4 | C | 10/01/97 | 217 | 955 | 485.94 | 29/08/97 | 2790 | EO |
| 138 | 4 | D | 07/02/97 | 308 | 286 | 698.54 | 26/12/97 | 1312 | EO |
| 139 | 4 | D | 17/03/97 | 217 | 1432 | 752.64 | 28/12/97 | 2210 | A |
| 140 | 5 | D | 23/09/96 | 40 | 321 | 809.25 | 12/11/96 | 337 | EO |
| 141 | 6 | P | 19/06/95 | 253 | 593 | 178.1 | 03/03/95 | 865 | EO |
| 142 | 6 | P | 27/08/97 | 405 | 302 | 81.48 | 06/10/98 | 441 | EO |
| 143 | 6 | P | 05/05/97 | 241 | 2634 | 1093.86 | 01/01/98 | 2731 | EO |
| 144 | 6 |  | 07/08/95 | 156 | 1040 | 1002.76 | 10/12/95 | 1039 | EO |
| 145 | 2 | D | 13/11/98 | 92 | 887 | 551.88 | 25/02/99 | 1393 | A |

Project types A = Adaptive, D = Development, P = Perfective, Pr = Preventive, C = Corrective, U = User support and estimate methods, EO = Expert opinion, A = Average, CAE = CA-Estimacs, D = Delphi, W = Widget counting.

knowing how good the current process is, and ensuring that any new process can perform better.

With respect to research into estimation processes, we agree with Vicinanza et al. (1990) that more emphasis should be put on understanding the information expert estimators need to produce good estimates and how they use that information. Furthermore, we believe such studies should be performed across a range of different organizations to investigate the extent to which estimating information is context dependent and the extent to which it is generic.

Another important issue is the extent to which the function point-based regression model changed over time. If algorithmic models are not stable over time, and we want to develop improved algorithmic models, the models must be capable of continual re-calibration, as new projects are completed and older projects become less and less like new projects. Furthermore, if project data (Table 19) get out-of-date quickly and our models require homogeneous data sets, we need procedures to generate models from relatively small data sets.

With respect to duration estimation, we found that effort estimates or function points were almost as good predictors as actual effort. However, we believe that it is important to capture information about team size before building prediction models. Duration is a controlled variable not a free variable, so any model that treats it as a simple free variable will never be very accurate.

## References

Abdel-Hamid, T.K., Madnick, S.E., 1989. Lessons learned from modeling the dynamics of software development. Commun. ACM 32 (12), 1426–1438.

Albrecht, A., Gaffney Jr., J., 1983. Software function, source lines of code and development effort prediction: a software science validation. IEEE Trans. Software Engng. 9 (6), 639–648.

Banker, R.D., Kemerer, C.F., 1989. Scale economies in new software development. IEEE Trans. Software Engng. 15 (10), 1199–1205.

Boehm, B.W., 1981. Software Engineering Economics. Prentice-Hall, Englewood Cliffs, NJ.

Conte, S.D., Dunsmore, H.E., Shen, V.Y., 1986. Software Engineering Metrics and Models. Benjamin/Cummings, California.

Cook, R.D., Weisberg, S., 1982. Residuals and Influence in Regression. Chapman and Hall, New York.

DeMarco, T., 1982. Controlling Software Projects. Prentice-Hall, Englewood Cliffs, NJ.

Dolado, J.J., 2001. On the problem of the software cost function. Inform. Software Technol. 43 (1), 61–72.

Hughes, R.T., 1997. An empirical investigation into the estimation of software. Ph.D. dissertation, University of Brighton, UK.

Kemerer, C.F., 1987. An empirical validation of software cost estimation models. Commun. ACM 30 (5), 416–429.

Kitchenham, B.A., 1992. Empirical studies of assumptions that underlie software cost estimation models. Inform. Software Technol. 34 (4), 304–310.

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

Kitchenham, B.A., Travassos, G.H., von Mayhauser, A., Niessink, F., Schneidewind, N.F., Singer, J., Takada, S., Vehvilainen, R., Yang, H., 1999. Towards an ontology of software maintenance. J. Software Maint. Res. Prac. 11, 365–389.

Kitchenham, B.A., Pfleeger, S.L., McColl, B., Eagan, S., 2001. An empirical study of maintenance and development projects. Technical Report TR/SE-0102, Keele University, Staffordshire, UK.

Klein, G., 1998. Sources of Power: How People Make Decisions. MIT Press, Cambridge, MA.

Lientz, B., Swanson, E.B., 1980. Software Maintenance Management. Addison-Wesley, Reading, MA.

Londeix, B., 1987. Cost Estimation for Software Development. Addison-Wesley, Reading, MA.

Low, G.C., Jeffery, D.R., 1990. Function points in the estimation and evaluation of the software process. IEEE Trans. Software Engng. 16 (1), 64–71.

Matson, J.E., Barrett, B.E., Mellichamp, J.M., 1994. Software development cost estimation using function points. IEEE Trans. Software Engng. 20 (4), 275–287.

Pfleeger, Shari Lawrence, 2000. Risky business: what we have yet to learn about risk management. J. Syst. Software 53 (3), 265–273.

Pickard, L.M., Kitchenham, B.A., Linkman, S.J., 1999a. Investigation of analysis techniques for software data sets. Technical report TR99-05, Department of Computer Science, Keele University, Staffordshire, UK.

Pickard, L.M., Kitchenham, B.A., Linkman, S.J., 1999b. An investigation of analysis techniques for software data sets. In: Proceedings Sixth International Symposium on Software Metrics (Metrics '99). IEEE Computer Society Press, Los Alamitos, CA.

Ratkowsky, D.A., 1983. Nonlinear Regression Modeling. Marcel Decker, New York.

Rosenberger, W.F., 1996. Dealing with multiplicities in pharmacoepidemiologic studies. Pharmacoepidem. Drug Safety 5, 95–100.

Shepperd, M.J., Schofield, C., 1997. Estimating software project effort using analogies. IEEE Trans. Software Engng. 23 (11), 736–743.

STATA Corporation, 1997. Intercooled STATA version 5.0 for Windows 95.

Stensrud, E., Myrveit, I., 1998. Human performance estimating with analogy and regression models: An empirical validation. In: Proceedings Fifth International Software Metrics Symposium (Metrics '98). IEEE Computer Society Press, Los Alamitos, CA, pp. 205–213.

Vicinanza, S., Prietula, M.J., Mukhopadhyay, T., 1990. Case-based reasoning in software estimation. In: Proceedings 11th International Conference on Information Systems, Copenhagen, Denmark, pp. 149–158.

Vicinanza, S., Mukhopadhyay, T., Prietula, M.J., 1991. Software effort estimation: An exploratory study of expert performance. Inform. Syst. Res. 2 (4), 243–262.