Analogy-X: Providing Statistical Inference to Analogy-Based Software Cost Estimation

Jacky Wai Keung, *Member*, *IEEE*, Barbara A. Kitchenham, *Member*, *IEEE Computer Society*, and David Ross Jeffery, *Member*, *IEEE Computer Society*

Abstract—Data-intensive analogy has been proposed as a means of software cost estimation as an alternative to other data-intensive methods such as linear regression. Unfortunately, there are drawbacks to the method. There is no mechanism to assess its appropriateness for a specific data set. In addition, heuristic algorithms are necessary to select the best set of variables and identify abnormal project cases. We introduce a solution to these problems based upon the use of the Mantel's correlation randomization test called Analogy-X. We use the strength of correlation between the distance matrix of project features and the distance matrix of known effort values of the data set. The method is demonstrated using the Desharnais data set and two random data sets, showing 1) the use of Mantel's correlation to identify whether analogy is appropriate, 2) a stepwise procedure for feature selection, as well as 3) the use of a leverage statistic for sensitivity analysis that detects abnormal data points. Analogy-X thus provides a sound statistical basis for analogy, removes the need for heuristic search, and greatly improves its algorithmic performance.

Index Terms—Cost estimation, management, software engineering, analogy.

1 Introduction

Software project managers require reliable methods for estimating software project costs. For more than 25 years, there has been considerable research effort directed toward software cost estimation. There is still no definitive solution as to the best method of software cost estimation. Among data-intensive algorithmic cost estimation methods, analogy-based software cost estimation (also referred to as Case-Based Reasoning (CBR)) is currently a popular alternative to constructing estimation models using linear regression. However, there are problems inherent in the method:

- There is no method to measure the appropriateness of the analogy approach for a specific data set.
- The existing methods for selecting the best set of variables and project cases to build an analogy is based either on brute-force search or heuristic algorithms (e.g., greedy search or hill climbing).
- There is no simple method to identify abnormal projects.

These issues are well recognized but as yet unresolved. For example, ANGEL [1], [2], which is the most well-known analogy-based estimation tool, currently uses a variety of different search algorithms to identify appropriate feature subsets and abnormal cases and has no method for identifying data sets for which analogy is inappropriate. As the research and development of CBR proliferates, resolving these issues becomes imperative. The goal of this paper is to introduce a method we call Analogy-X that provides a solution to these problems.

In Section 2, we set the context for our study with discussion of the current status of analogy-based cost estimation. This is followed in Section 3 with a detailed discussion of the underlying theory of our approach. Section 4 describes statistical constructs of our approach. Section 5 describes how our approach supports stepwise variable selection and discusses the impact of categorical variables on a distance matrix. This section also describes a leverage metric based on the Jackknife method that can be used to identify abnormal data points (i.e., abnormal cases) in the data set and to permit sensitivity analysis. Section 6 demonstrates the application of Analogy-X on the Desharnais data set using its sensitivity analysis and stepwise variable selection procedure and further investigation of our sensitivity analysis on two randomized data sets is discussed in Section 7. In Section 8, we conclude by providing a summary and recommendations.

2 Background

2.1 Case-Based Reasoning/Analogy

Analogy-based software estimation is a typical example of a CBR strategy [3]. Although analogy or CBR was originally regarded as a means of human problem-solving and decision making, analogy-like methods based on automated analysis of project data are currently a popular alternative approach to software cost estimation compared with closed-form algorithmic models such as Boehm's COCOMO system [4] or other data-intensive model building methods based on linear regression. Data-intensive analogy-based estimation was

- *J.W. Keung is with ESE/NICTA, Bay 15, Australian Technology Park, Garden St, Eveleigh NSW 1430, Sydney, Australia. E-mail: Jacky.Keung@nicta.com.au.*
- *B.A. Kitchenham is with Keele University, Keele, Staffordshire ST5 5BG, UK. E-mail: Barbara.Kitchenham@nicta.com.au.*
- *D.R. Jeffery is with the School of Computer Science and Engineering, University of New South Wales, Sydney 2052, Australia. E-mail: Ross.Jeffery@nicta.com.au.*

Manuscript received 18 May 2006; revised 17 June 2007; accepted 7 May 2008; published online 15 May 2008.
Recommended for acceptance by P. Jalote.
For information on obtaining reprints of this article, please send e-mail to: tse@computer.org, and reference IEEECS Log Number TSE-0109-0506.
Digital Object Identifier no. 10.1109/TSE.2008.34.

popularized in the late 1990s by Shepperd and Schofield [5], who successfully demonstrated its potential for software effort prediction. The general principle of automated analogy [3], [5] is to reuse experience in the form of project cases stored in a repository. First, case retrieval is performed to extract similar cases to the target case based on their feature similarity (measured by a distance matrix). Then, case adaptation is applied to the selected similar cases in order to obtain a prediction of effort for the target case. Empirical experiments with tools such as ESTOR [6] and ANGEL [3], as well as other studies [2], [3], [5], [7], [8], have demonstrated that software effort estimation by analogy is a viable alternative to other conventional estimation methods in terms of predictive accuracy and flexibility. Proponents of analogy point out that it can be used with partial knowledge of the target project at an early stage of a project and that the concept of analogy-based estimation (i.e., looking for a project that is similar to the project to be estimated) may be more intuitive to managers than regression-based estimation [9].

2.2 The Choice of Estimation Method

In the context of software cost estimation research (as opposed to practice), the most commonly applied software effort prediction methods are regression [9] and data-intensive analogy [5]. Previous empirical software cost estimation studies have attempted to determine which method is best; however, these studies have produced conflicting results. For example, Shepperd et al. [3], [5] claimed analogy provided better prediction accuracy. This was supported by Angelis and Stamelos [10], who found analogy-based systems were far superior to other methods, and by the more recent work of Mendes et al. [11], [12] on a large heterogeneous data set. In contrast, Myrtveit and Stensrud [13] replicated previous studies described in [5], but, instead, they found analogy was not better than regression and also suggested that the results are sensitive to experimental design. Similarly, Briand et al. [14] found analogy-based systems were less robust than other methods, particularly when dealing with heterogeneous data sets. Jeffery et al. [15] also concluded stepwise regression outperformed analogy-based systems with the ISBSG data set.

Recently, Mair and Shepperd [16] undertook a systematic review to investigate these contradictory results. They reviewed 20 primary studies comparing regression and analogy conducted during the past decade and concluded that there was no clear indication that regression was better than analogy or vice versa. They concluded that the mixed results are due to the characteristics of the data set and the individual data points [17]. The implication is that the resulting prediction is sensitive to the data quality of individual data sets. Shepperd and Kadoda have studied this issue using simulation [18].

These data-intensive software effort prediction models have not delivered results that are regarded as satisfactory by industry; an alternative is to utilize the knowledge of an experienced expert. Expert opinion is a human intensive approach that is the most commonly adopted estimation method in the software industry [19], [20]. Estimates are usually produced by a domain expert based on their own personal recollection of similar past events in the organization. It is flexible and intuitive in the sense that it can be applied in a variety of circumstances where other estimating techniques do not work, for example, when there is a lack of historical data. Although it is widely practiced, there are no standard methods for expert opinion-based estimation [21]. The process by which estimates are derived are not usually made explicit. Therefore, the process is nontransferable and the estimates themselves are not repeatable. Jorgensen [20] conducted an extensive systematic review of empirical evidence about expert opinion estimation that synthesizes 15 articles and concluded that expert opinion estimates were not systematically worse than estimates based on cost models. Furthermore, there were theoretical reasons to explain under what conditions expert opinion estimates would be likely to outperform cost models and vice versa [22].

As suggested by Mair and Shepperd [16] and Jorgensen [20], we believe researchers should be focusing on when to use technique A rather B, as opposed to attempting to identify one method as the "best" method. What is important to a project manager is which method is most appropriate in his/her specific circumstances. We would like to stress that the purpose of this paper is not an attempt to promote or compare the performance of regression and analogy but to focus on the issue of improving the rigor of the analogy method itself.

2.3 Issues in Analogy

Although the analogy approach has been used successfully in the industry for more than 10 years, it has two major weaknesses.

First, analogy does not provide any objective quantitative measure of statistical significance to confirm that is a suitable approach for a particular data set. Kirsopp et al. [23] recognize that analogy-based methods suffer from the problem of poor explanatory value, offering no justification that is helpful for the project manager to determine the usefulness of the prediction. This problem is also recognized in [17], where it proved difficult to decide whether an analogy-based approach was a useful technique or not. Analogy-based estimation currently has no underlying statistical basis. In contrast, linear regression's goodness of fit can be determined using statistical significance tests with a simple p-value.

Second, there is no agreed algorithmic method for best feature and case selection for analogy-based systems as best feature selection is based either on search heuristics or a full search of all possible combinations of features. Shepperd and Schofield [5] suggest using human-based expert opinion to establish those features of a case that are believed to be significant in determining similarity. More recently, Kirsopp et al. have suggested using machine learning (ML) techniques to search for appropriate feature subsets [23]. They examined a variety of approaches to partially search the data set comprised of each possible combination of features, including Random, Hill Climbing, and Forward Sequential Selection search techniques, in an attempt to solve this problem. Their findings suggest that some form of heuristic-based initialization might prove

useful for this problem, but a best subset cannot be guaranteed and may introduce spurious effects. These heuristic-based searching algorithms are not very efficient and are somewhat time-consuming, even with a modern computer.

In essence, both of the issues are largely due to the fact that data-intensive analogy is unable to directly assess its model fitness. To the best of our knowledge and at the time of writing, there is no other better alternative to heuristic-based search for identifying best feature subsets and abnormal cases.

Recently, Li et al. [24] have identified other weaknesses in analogy with respect to distance matrix construction and constructing a prediction. Constructing a prediction is outside the scope of Analogy-X; however, their work is relevant to ours because Analogy-X manipulates distance matrices. With respect to distance matrix construction, Li et al. suggest novel approaches to constructing distance matrices for both nonstandard project properties, such as value ranges, sets, etc., including nominal scale values, and for handling missing values. In principle, our results are independent of the means by which the distance matrices are constructed, so our approach will support Li et al.'s method as well as the standard analogy. However, as will be shown later, we treat nominal scale values differently.

2.4 Accuracy Statistics/Model Performance Measures

Another issue is that analogy relies on fitness functions such as the Mean Magnitude of Relative Error (MMRE) statistic to assist the selection of features and the identification of abnormal cases. For a set of n projects, MMRE is calculated as follows:

$$\mathrm{MMRE} = \frac{\sum \mathrm{MRE}_i}{n}, \tag{1}$$

where MRE_i is the magnitude (absolute) relative error for project i, that is,

$$\mathrm{MRE}_i = \frac{|Actual\_Effort_i - Estimated\_Effort_i|}{Actual\_Effort_i}. \tag{2}$$

Although other goodness-of-fit functions can be used, MMRE still appears to be the most popular. However, Foss et al. [25] and Kadoda et al. [26] have shown that MMRE is a very unreliable metric to assess estimation accuracy since it is inherently biased in favor of underestimates and they strongly recommend not using MMRE to evaluate and compare prediction models.

3 Overview of the Mantel's Method

We propose using Mantel's correlation and randomization test [27], [28], [29], [30] to overcome the shortcomings of analogy. Mantel's correlation can be used to compare the association between the elements in two distance matrices and the randomization test is used to assess the statistical significance of the association.

The fundamental assumption underlying analogy-based estimation is that projects that are similar with respect to a set of project features will also be similar with respect to effort. In this case, effort is the response variable we usually need to predict for software development. Mantel's randomization test allows us to formally test this assumption. In this section, we provide an overview of the underlying theory to support our proposal.

3.1 The Theoretical Principle

Mantel's correlation for comparing two distance matrices was first introduced by Mantel in 1967 as a solution to the problem of detecting space and time clustering of diseases for cancer research [27]. It has since been widely adopted in ecology, biology, and psychology research to address this kind of problem [29]. A classical example in ecology is attempting to explain the distribution of species based on constraints of their environmental variables. The operative question in these ecology experiments is: "Do samples that are close with respect to Xs (environmental variables) also tend to be close with respect to Ys (species variables)?" The question is analogous to the questions we want to ask in an analogy-based software cost estimation approach, i.e., "Do projects that are close with respect to Xs (project features) also tend to be close to Ys (development effort)?" In the previous example, a distance matrix was constructed to measure environmental property distances from the related environmental variables and, similarly, species geographical distances were measured by a distance matrix constructed from species variables; Mantel's method considers these two distance matrices as parameters. The objective is to test a hypothesis that there is a relationship between the predictor distance matrix and the response distance matrix.

Although Mantel discussed more general situations and findings in his original study [27], Manly [28], [30] provides more comprehensive examples of Mantel's method. He describes the basic principle of Mantel's method, which is to measure the association between the corresponding elements in two distance matrices by a suitable statistic, usually using the Pearson correlation. Theoretically, other correlation coefficients may also be used, such as Spearman's correlation and Kendall's correlation. The significance of the correlation is then determined by a permutation procedure in which the original value of the test statistic is compared with the distribution of the statistics found by randomly reallocating the order of the elements in one of the distance matrices. To understand the underlying principles and the permutation procedure of Mantel's method, consider the following example: Let matrix A consist of the distances of predictor variables and matrix B consist of the distances of response variables as follows:

$$A = \begin{bmatrix} 0 & a_{12} & \cdots & a_{1n} \\ a_{21} & 0 & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & b_{12} & \cdots & b_{1n} \\ b_{21} & 0 & \cdots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \cdots & 0 \end{bmatrix}. \tag{3}$$

The distance matrix is a matrix of n cases (e.g., projects). Each case has a distance measure constructed from p features (variables). Thus, for example, the distance

element between case 1 ($\times 1$) and case 2 ($\times 2$) is calculated using the following simple euclidean distance:

$$a_{21} = \sqrt{\sum_{i=1}^{p} (x1_i - x2_i)^2}. \quad (4)$$

Equation (4) considers the values of all p variables for each pair of cases. Before the diagonal elements can be constructed, the variables have to be standardized by transformation so that they are all equally weighted and comparable. The same procedure is used for analogy-based estimation. In ANGEL, a euclidean distance is used after transforming all values to be in the range of 0 and 1 (i.e., by dividing each value by max-min, where max is the maximum value in the data set for the relevant variable and min is the minimum value). Li et al. [24] adopt the same principle for the wider range of project factors.

Because of symmetry, only the lower diagonal elements in the above matrices (3) need to be considered when constructing the Mantel's test statistic. Thus, the Mantel's correlation coefficient is

$$R_m = \frac{\sum a_{ij} b_{ij} - \sum a_{ij} \sum \frac{b_{ij}}{m}}{\sqrt{\left[\left\{\sum a_{ij}^2 - \frac{\left(\sum a_{ij}\right)^2}{m}\right\} \times \left\{\sum b_{ij}^2 - \frac{\left(\sum b_{ij}\right)^2}{m}\right\}\right]}}, \quad (5)$$

where m is the number of diagonal elements in the distance matrix and it is given by

$$m = \frac{n(n-1)}{2}. \quad (6)$$

For the randomization test, the distance matrix elements are randomly permuted for one of the matrices, for example, matrix A (3). A random order matrix A_{Random} (7) can be constructed based on the random ordering of elements. For example, one randomization of the elements of A gives the matrix A_{Random}:

$$A_{Random} = \begin{bmatrix} 0 & a_{68} & a_{18} & \cdots & a_{38} \\ a_{68} & 0 & a_{16} & \cdots & a_{36} \\ a_{18} & a_{16} & 0 & \cdots & a_{31} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{38} & a_{36} & a_{31} & \cdots & 0 \end{bmatrix}. \quad (7)$$

The entry in column 1, row 2 is the distance between data items 8 and 6; the entry in column 2, row 3 is the distance between data item 6 and 1, and so on. The value of the Mantel's correlation is then computed using matrix B (3) and A_{Random} (7). Repeating the same procedure many times produces the randomization statistic distribution. Using the randomization distribution, we can test whether the value of the Mantel's correlation derived from the original pair of distance matrices is significantly different from zero. If the Mantel's correlation is significantly different from zero, we can be sure that projects that are close together with respect to project features are close together with respect to effort and that analogy-based estimation is an appropriate method for the data set under investigation. We assume here that the projects are a sample from a well-defined population and that any future projects requiring estimation would come from the same population. These are the same assumptions underlying regression.

Some researchers refer to a similarity matrix (where 1 means two different projects are equivalent with respect to all project factors) and others to a distance matrix (where 1 means two projects are completely different with respect to all project factors); however, elements of a distance matrix are just 1-elements of a similarity matrix, so there is substantially no difference between the two terms.

3.2 Number of Permutations Needed for a Randomization Test

When calculating Mantel's correlation, it is generally impractical to consider all possible permutations of distance matrix elements when the number of cases is large. Nevertheless, it is very important that the significance level generated from permutations be close to the level that would be obtained if all possible permutations were calculated.

Manly [30] has adopted Marriott's principle [31], which describes the number of permutations needed for randomization in Monte Carlo tests, and demonstrated that it is generally practical to determine the full randomization distribution with available computing power for a distance matrix with up to nine elements or 9!. Alternatively, 1,000 randomizations are a realistic minimum for estimating a significance level of about 0.05 and 5,000 randomizations are a realistic minimum for estimating a significance level of about 0.01. Further examples can be found in [30], [31].

3.3 Statistic Package and Library

Mantel's randomization test has been implemented in a number of software packages. We use the open-source GNU R statistic package [32] with the ADE-4 [33] and VEGAN [34] libraries, which provide all of the functions necessary to compute the Mantel's test and distance matrices. Simulation results in this study were also produced with these statistical tools.

4 Analogy-X: Statistics

We suggest that the principles of Mantel's method can be applied in the case of analogical reasoning for software cost estimation. We propose extending the basic Analogy method to include statistical inference based on Mantel's method; we call this method Analogy-X, i.e., an eXtension for Analogy. Before we present the basic Analogy-X procedures in the next section, we first introduce the statistical theory behind the procedures in this section.

As explained in Section 3, we can use Mantel's approach to determine the correlation R between the two distance matrices and use the randomization test to identify whether R is significantly different from zero. If R is not significantly different from zero, analogy is not suitable for the specific data set. However, we need additional techniques to support the following:

- Identification of abnormal cases, i.e., cases that distort the relationship between distance matrices

and cause us to overestimate or underestimate R. This we refer to as a sensitivity analysis. It requires a method to detect cases that significantly distort the value of R.

- Identification of the subset of project factors (i.e., dependent variables (DVs)) that significantly influence the value of R. This involves a stepwise introduction of project factors that significantly increase R. It requires a robust estimate of R (i.e., one unaffected by abnormal values) and a means of determining the confidence limits on R (when R is nonzero) in order to assess whether adding a new project factor to the distance matrix significantly increases the value of R.
- A procedure for managing nominal scale project factors. We demonstrate the need for a special procedure in Section 4.2.

We use a Jackknife approach to support these requirements, as described in Section 4.1.

4.1 The Jackknife Estimator of Mantel's R

Analogy-X supports stepwise variable selection and sensitivity analysis. These applications are required to have a robust measure of Mantel's correlation R and this is provided using the Jackknife method. The Jackknife method was first proposed by Tukey [35] for use in statistics as a general approach for hypothesis testing and calculating confidence intervals and it is commonly known as the "Leave-One-Out" approach [36].

We use the Jackknife method to provide an unbiased estimator of the Mantel's R. The Jackknife estimator $\bar{R}$ of Mantel's R can be calculated as

$$\bar{R} = \frac{\sum_{i=1}^{n} R_i}{n}, \tag{8}$$

where n is the total number of cases and R_i is the Mantel's correlation of all cases excluding the ith case (project) in turn. The Jackknife estimator $\bar{R}$ will be distributed normally (approximately) with an unknown variance S^2 that can be estimated as

$$S^2 = \frac{\sum_{i=1}^{n} (R_i - \bar{R})^2}{n-1}. \tag{9}$$

Then, the Jackknife confidence interval of $\bar{R}$ at the 0.05 significance level (approximately) can also be estimated as

$$CI_{Jackknife} = \bar{R} \pm 2 \times S. \tag{10}$$

The Jackknife estimator of Mantel's R and its confidence interval provide the basis of the Analogy-X method of variable selection and sensitivity analysis. Although the computation required is largely determined by the sample size n, it is not computationally intensive by present day standards.

4.2 Analysis of Categorical (Nominal) Variables

During our initial studies of the viability of our approach, we tried out our ideas on the Desharnais 77 data set. This data set is publicly available through the software package of ANGEL and has been the focus of important Analogy studies, for example, in [2], [5], [17], [26]. The original version of the data set had 81 projects, but four of the projects had missing values and were excluded from our analyses, which is common with most other studies.

In common with many software data sets, the Desharnais 77 data set includes features measured with a mix of continuous, ordinal, and categorical (nominal) variables. We found that Mantel's method worked well with continuous and ordinal scale variables. However, we observed some unexpected effects when we incorporated the categorical variable:

- Including the categorical variable and a continuous variable in the distance matrix could substantially reduce the value of the Mantel's correlation, even when the Mantel's correlation using the continuous variable alone was extremely strong.
- Using the categorical variable to partition the data set, then applying Mantel's correlation to projects within each partition led to Mantel's correlations in some subsets that were substantially greater than the value obtained for the entire data set. This implied that a categorical variable was important for identifying similar cases, but the effect could not be observed if the categorical variable was included in the distance matrix.

Thus, unlike regression, where the introduction of inappropriate dummy variables merely leads to the regression coefficient for the dummy variable not being significantly different from zero, the introduction of an inappropriate categorical variable may completely undermine the association between effort and project factors. This means that the impact of categorical variables cannot be assessed in the same way as continuous and ordinal scale variables.

The reason this occurs can be understood using a simple example based on the hypothetical data set shown in Table 1, which has one DV and six independent variables, three of which are continuous (IV1, IV2, and IV3) and three of which are categorical (G1, G2, and G3). In this data set, the variable DV is strongly correlated with IV3 and G1.

Table 2 shows the value of the distance matrix elements for each pair of cases based on each variable separately. It confirms that the Mantel's correlation is largest for IV3 (0.948) and the correlation for G1 was also large (0.923). The association between DV and IV3 is shown graphically in Fig. 1.

Distance matrices elements constructed from IV3 and G1 and from IV3 and G3 are shown in Table 3. The effect of including both highly correlated variables (IV3 and G1) is to reduce the overall Mantel's correlation. The reason this occurs is shown in Fig. 2, where it is clear that the categorical variable has distorted the relationship between the distance matrix elements because the impact of the categorical variable (which is 0.5 or 0) is so much greater than the impact of the continuous variable (which is between 0 and 0.15).

Fig. 3 demonstrates how severe the distortion can be for an inappropriate categorical variable.

If all of the project feature variables are categorical, it may sometimes be appropriate to use distance matrices to select

TABLE 1
Hypothetical Data Set

Case	DV	G1	G2	G3	IV1	IV2	IV3
P1	1.64	1	0	0	0.40	1.00	0.60
P2	1.71	1	0	0	0.50	0.75	0.70
P3	0.82	0	0	1	0.70	0.50	0.30
P4	0.88	0	1	0	0.80	0.00	0.20
P5	0.38	0	0	1	0.00	0.90	0.00
P6	0.94	0	1	0	1.00	0.30	0.10
Correlation with DV		**0.923**	**-0.228**	**-0.694**	**0.132**	**0.299**	**0.948**

TABLE 2
Distance Matrix Elements for Each Pair of Cases for Each Individual Variable

Cases	DV	G1	G2	G3	IV1	IV2	IV3
P1,P2	0.07	0	0	0	0.10	0.25	0.10
P1,P3	0.82	1	0	1	0.30	0.50	0.30
P1,P4	0.76	1	1	0	0.40	1.00	0.40
P1,P5	1.26	1	0	1	0.40	0.10	0.60
P1,P6	0.70	1	1	0	0.60	0.70	0.50
P2,P3	0.89	1	0	1	0.20	0.25	0.40
P2,P4	0.83	1	1	0	0.30	0.75	0.50
P2,P5	1.33	1	0	1	0.50	0.15	0.70
P2,P6	0.77	1	1	0	0.50	0.45	0.60
P3,P4	0.06	0	1	1	0.10	0.50	0.10
P3,P5	0.44	0	0	0	0.70	0.40	0.30
P3,P6	0.12	0	1	1	0.30	0.20	0.20
P4,P5	0.50	0	1	1	0.80	0.90	0.20
P4,P6	0.06	0	0	0	0.20	0.30	0.10
P5P6	0.56	0	1	1	1.00	0.60	0.10
Mantel's correlation with DV		**0.834**	**-0.199**	**0.219**	**0.246**	**-0.052**	**0.873**
p-value		**0.192**	**0.979**	**0.109**	**0.278**	**0.628**	**0.044**

appropriate subsets of variables (as is done in other disciplines), but it is important to treat the nominal variables appropriately. Classical regression analysis handles nominal scale variables by transforming them into a set of $n-1$ dichotomous (dummy) variables, where n is the number of distinct categories in the nominal scale. This approach is invalid for distance matrices.

Again, a simple example can explain the problem. Consider the data set shown in Table 4, which shows six project cases with one categorical variable CatVar, which has four categories and would be transformed into three dummy variables if we were undertaking a regression analysis.

The distance between case P1 and the other cases is illustrated in Table 5 for the categorical variables and the three dummy variables. For the categorical variable, the numbers 1 to 4 are treated as labels "1," "2," "3," and "4."

If the category labels are the same for two project cases, the distance between the cases is 0 and it is 1 otherwise. Thus, the distance between P1 and P2 is 0 because their category labels are the same. The difference between P1 and the other cases is 1 because their category labels are different. However, for the dummy variables, only Dum1 has the same distance values as CatVar for P1. The other dummy variables behave differently because pairs of cases are treated as similar if they both have the value of 1 or if they both have the value of 0. Thus, the distance measured by dummy variables has a major problem. If you decide to select Dum2 (rejecting Dum1 and Dum3), it appears that P1 is similar to P2, P5, P6, P7, and P8. However, the categorical variable makes it clear that P1 is only similar to P2. Thus, for analogy, the distance matrix must be based on the nominal scale variable and not on the dummy variables.

4.3 Within-Group Matrix Correlation for Categorical Variables

In response to the undesirable aspects of categorical variables in distance matrix correlation discussed above, we suggest using only the relevant distance matrix elements to calculate Mantel's R. This is achieved by using only the distance matrix elements corresponding to comparisons between projects in the same group. Consider a simple data set that has three groups and six projects (two in each group: P1 and P2 are in group 1, P3 and P4 are in

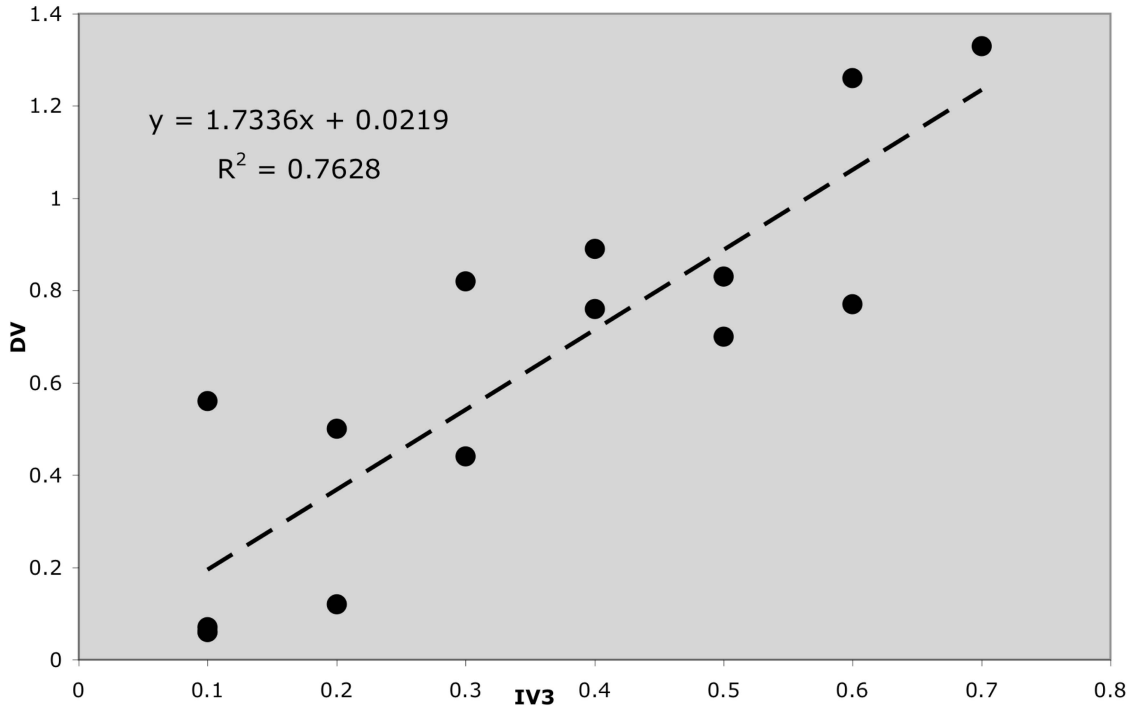


Fig. 1. Association between distance matrix elements for DV and IV3.

group 2, and P5 and P6 are in group 3), then the two distance matrices $distX$ and $distY$ will have a structure as shown in Table 6.

To calculate Mantel's R irrespective of groups, the correlation coefficient is based on all 15 matrix elements. To calculate R allowing for the effect of groups, omit all elements that correspond to projects that do not share the same group. In Table 6, the within-group correlation is calculated based on the three elements $a12$, $a34$, and $a56$. This approach enables the same Jackknife procedure to calculate the confidence limits and provides a test statistic equivalent to the one used for nonnominal variables in Analogy-X.

However, one theoretical problem is that the within-group R is based on fewer matrix element values than the simple Mantel's correlation, so the confidence intervals for the within-groups R will be larger than the confidence limits for the cross-groups R. In addition, this requires modification to the Mantel's correlation algorithm, where a distance matrix matching and isolation function must be executed before the Mantel's correlation and randomization test takes place.

TABLE 3
Distance Matrix Elements for Distance Matrices Based on Two Variables

Cases	DV	IV3+G1	IV3+G3
P1,P2	0.07	0.05	0.05
P1,P3	0.82	0.52	0.52
P1,P4	0.76	0.54	0.20
P1,P5	1.26	0.58	0.58
P1,P6	0.70	0.56	0.25
P2,P3	0.89	0.54	0.54
P2,P4	0.83	0.56	0.25
P2,P5	1.33	0.61	0.61
P2,P6	0.77	0.58	0.30
P3,P4	0.06	0.05	0.50
P3,P5	0.44	0.15	0.15
P3,P6	0.12	0.10	0.51
P4,P5	0.5	0.10	0.51
P4,P6	0.06	0.05	0.05
P5P6	0.56	0.05	0.50
Mantel's correlation with DV		0.866	0.477
p-value		0.044	0.018

5 Analogy-X: Procedures

We have presented the theoretical principles of Analogy-X in the previous sections, providing all of the necessary basic constructs for the Analogy-X procedures. In this section, we first present the basic procedures used by Analogy-X to analyze extreme data points and to remove these outlying cases to ensure the stability of the data set and avoid spurious correlation. We then present an Analogy-X stepwise variable selection technique to select an appropriate set of features for the resultant data set obtained in the sensitivity analysis.

5.1 Sensitivity Analysis

Shepperd and Kadoda [17] performed simulation studies of the accuracy of the current techniques used, which are: Stepwise Regression, Rule Induction, and CBR. A particularly interesting part of their study involved presenting the techniques with random data sets. They found that stepwise regression was capable of finding relationships where none existed (i.e., when presented with random data

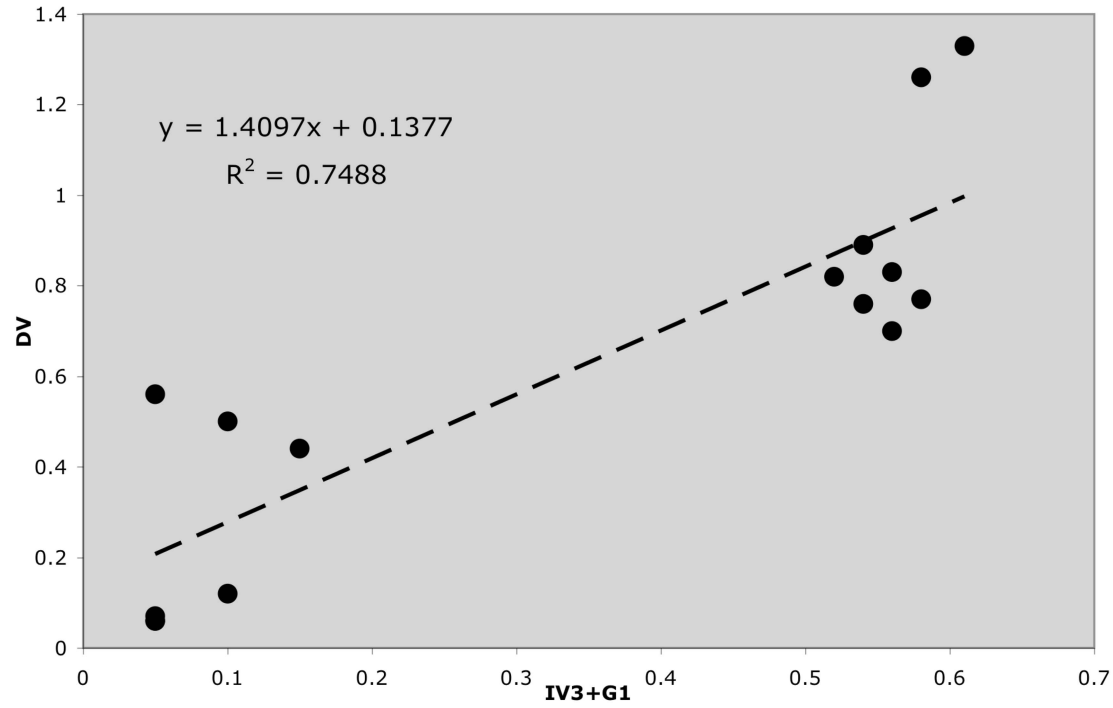


Fig. 2. Association between distance matrix elements for DV and IV3+G1.

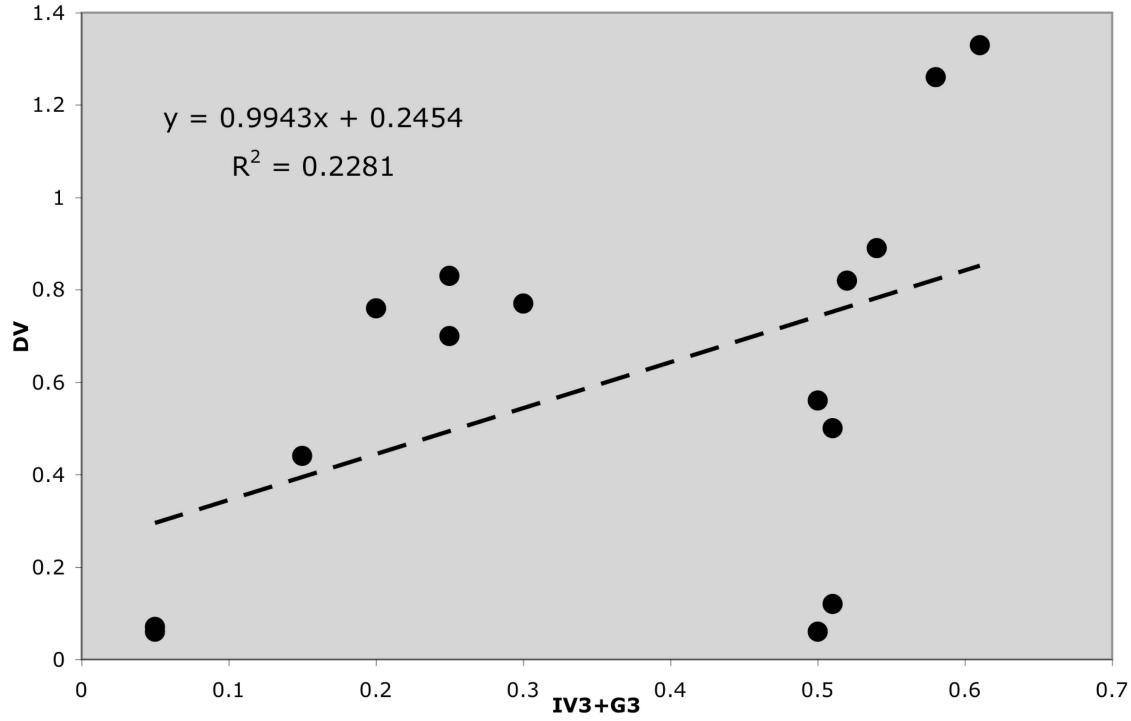


Fig. 3. Association between distance matrix elements for DV and IV3+G3.

sets) particularly for small data sets. Rule Induction was even more vulnerable to detecting spurious relationships. They also point out that CBR has no means of dealing with random data sets and will "always endeavor to predict no matter what the circumstances." The procedure described in this section avoids the problem of always predicting even if the data set is inappropriate while also providing a mechanism to prevent Analogy-X being vulnerable to detecting relationships in random data sets.

Statisticians are well aware of the danger of abnormal data points creating spurious correlations in data sets and have introduced a number of techniques (broadly described as sensitivity analysis) to investigate the extent to which statistical results could be an artifact of particular data points. One approach is to identify "high leverage" data points. High leverage data points have a large impact on the results of the analysis. Sensitivity analysis involves assessing whether relationships are stable to the removal of

TABLE 4
A Data Set with One Nominal Scale Variable with Four Categories

	Categorical Variable	Dummy Variable		
Project	CatVar	Dum1	Dum2	Dum3
P1	1	1	0	0
P2	1	1	0	0
P3	2	0	1	0
P4	2	0	1	0
P5	3	0	0	1
P6	3	0	0	1
P7	4	0	0	0
P8	4	0	0	0

high leverage data points from the data set. In this section, we explain how Mantel's correlation provides us with the ability to identify high leverage data points in the context of distance matrices and so apply sensitivity analysis to analogy-based estimation.

We have developed the Mantel Leverage Metric to support sensitive analysis for analogy, based on the same principles as the Jackknife method [36] and the properties of the standard normal distribution, i.e., a normal distribution with a mean of 0 and a variance of 1 $\mathrm{N}(0,1)$. The Mantel Leverage metric (LM) is based on calculating the Mantel's correlation excluding each case (project) in turn. This indicates the extent to which the Mantel's correlation for the complete data set is influenced by each individual case. To calculate LM for each case i, let R_i be the Mantel's correlation for the data set excluding case i and $\bar{R}$ (8) be the Jackknife Estimator of overall Mantel's R, then

$$LM_i = R_i - \bar{R}. \tag{11}$$

LM_i is the difference (residual) between the overall Mantel's R and R_i indicating the impact of the specific case i on the overall R. Under the null hypothesis that case i is not abnormal, R_i will be an unbiased estimator of $\bar{R}$ and will be approximately $\mathrm{N}(0, S^2)$. The following z test provides a mechanism to formally verify whether the value of R_i is an abnormal one. For each case i, LM_i can be converted to its standard normal form:

$$z_i = \frac{LM_i}{S}. \tag{12}$$

If $|z_i|$ is greater than 2, the case is significantly different from 0 at the 0.05 significance level (approximately), if $|z_i|$ is greater than 4, the significance level is 0.001 (approximately).

To complete the sensitivity analysis, all cases identified as abnormal should be removed from the data set and the Mantel's correlation is recalculated. We propose the use of $|z_i| > 4$ as an indicator of an abnormal data point because we only want to remove extreme data points. We do not want to be oversensitive to individual cases when we perform an iterated case removal process.

If the relationship between the distance matrices is resilient to the removal of the abnormal cases measured by the p-value of Mantel's correlation, we can be confident that analogy is appropriate for the reduced data set and the strength of the correlation or the Jackknife estimator $\bar{R}$ can be used to indicate the explanatory power of the Analogy model for the data set.

5.2 Stepwise Analogical Evaluation

An essential issue in CBR is to identify which features are important for the purpose of case retrieval. Mantel's method allows us to adopt a stepwise procedure for variable selection analogous to stepwise regression (for clarity, we assume that R is significant for at least one project factor):

1. For each feature X_i, construct a project feature distance matrix $distX$ for that variable alone. Compute Mantel's correlation between the effort distance matrix $distY$ and project feature distance matrix $distX$ and determine the significance of the

TABLE 5
Distances between P1 and the Other Cases for Each Variable

	Distance			
Distance Matrix Elements for P1	Dum1	Dum2	Dum3	CatVar
(P1,P2)	0	0	0	0
(P1,P3)	1	1	0	1
(P1,P4)	1	1	0	1
(P1,P5)	1	0	1	1
(P1,P6)	1	0	1	1
(P1,P7)	1	0	0	1
(P1,P8)	1	0	0	1

TABLE 6
Structure of Distance Matrix

	P1	P2	P3	P4	P5	P6
P1	-					
P2	**a12**	-				
P3	a13	a23	-			
P4	a14	a24	**a34**	-		
P5	a15	a25	a35	a45	-	
P6	a16	a26	a36	a46	**a56**	-

correlation R. Identify the project variable for which R is greatest (call that X_1).

2. Calculate the Jackknife version of Mantel's R for X_1 (i.e., $\bar{R}_1$) and its upper confidence limit (UCL) (10).
3. For the remaining variables, calculate a distance matrix for each project feature in combination with X_1. If the best Mantel's correlation obtained is less than or equal to the UCL value of $\bar{R}_1$, stop and use X_1 alone for analogy-based estimation.
4. If there are other project features which, in combinations with X_1, give a larger Mantel's correlation that is also greater than the UCL of $\bar{R}_1$, choose the variable that in combination with X_1 gives the largest correlation and call that new set of variables X_2 and accept both project features in analogy-based estimation.
5. Repeat Steps 3 to 4, adding one project feature at a time, until a maximum Mantel's correlation value is reached, determined by the UCL of Jackknife estimator of Mantel's correlation derived from the previous variable combination, then stop.

The procedure described above is a simple forward stepwise algorithm. However, implementing a full stepwise algorithm that includes a test for the "least useful predictor in the model" [37] is perfectly feasible.

To integrate sensitivity analysis with stepwise variables selection, we suggest the standard process used in regression: First perform stepwise variable selection and then perform sensitivity analysis. If significantly abnormal cases are detected, remove them and repeat the stepwise variable selection procedure.

5.3 Incorporating Categorical (Nominal) Variables

We have introduced a stepwise variable selection procedure and a sensitivity analysis technique to remove abnormal data points in this section. To address data sets with categorical and numerical data, we propose a two-stage procedure as follows:

Stage 1. Identify and select all the numerical variables for the feature distance matrix and follow all procedures discussed in Section 4.2. This will ensure that all statistically relevant noncategorical features are detected. A Jackknife estimator $\bar{R}$ and its confidence intervals are also produced.

Stage 2. Evaluate the best subgroups for project cases based on the categorical (nominal) variables. After a subset of numerical variables is found in Stage 1, partition the data set into subgroups based on the nominal variables. This should be done for each categorical variable in turn and then the categorical variables that significantly improve the Mantel's correlation can be identified in a stepwise fashion. This process should stop either when further subgrouping does not significantly improve Mantel's correlation or when the subgroup sizes become too small (e.g., < 5 cases). Note that we have not investigated a cut-off value, but we would not expect the algorithm to work with a very small number of projects in each subgroup.

In order for a stepwise procedure to work, we must have a way of obtaining any overall correlation from the Mantel's correlation obtained from the individual Mantel's correlation for each subgroup. The aggregated correlation from all groups can be calculated using our within-group Mantel's correlation discussed in Section 4.3. Similarly, the aggregated upper 95 percent confidence limit can be derived using the same Jackknife principle discussed in Section 4.1.

To confirm that the impact of the selected categorical variable is statistically significant, a similar approach can be adopted as discussed in Section 5.2. If the within-group $R >$ UCL of $\bar{R}$ of all of the variables identified in Stage 1, then we are confident that the grouping effect is significant and should be included in the final prediction system.

This two-stage procedure is the simplest and most straightforward approach we have found to solve the categorical variable problem. However, it differs substantially from the approach adopted by Li et al. [24]. In our opinion, the distortion caused by mixing nominal and numerical values implies that our two-step approach has some merit compared with Li et al.'s approach, but further research is needed to assess which approach is more reliable.

6 Example 1: Applying Analogy-X on the Desharnais Data Set

In this section, we demonstrate the use of Analogy-X on the Desharnais 77 data set. This data set has nine independent variables, which are shown in Table 7. We use Actual_Effort in person hours as the DV. Note that, of the nine independent variables, Dev.Env is a categorical variable and, therefore, we have to perform the two-stage analysis described in the previous section.

6.1 Stage 1: Variable Selection for Nonnominal Variables

Applying the stepwise analysis to the original Desharnais data set (all 77 cases), we found that the Adj.FPs distance matrix, the RawFP distance matrix, the Transactions distance matrix, and the Entities distance matrix were significantly correlated with the Actual Effort distance matrix (see Table 8).

TABLE 7
Desharnais 77 Project Features

Independent Variable	Description	Type
Adj.FPs	Adjusted Function Points	Continuous
RawFPs	Raw Function Points	Continuous
Transactions	Number of. Transactions	Discrete (Ratio scale)
Entities	Number of Entities	Discrete (Ratio scale)
Adj.Factor	Technology Adjustment Factor	Continuous
Year.Fin	Year Finished	Discrete (Interval Scale)
ExpProjMan	Experience of Proj. Mgmt	Discrete (Ordinal)
ExpEquip	Experience of Equipment	Discrete (Ordinal)
Dev.Env	Development Environment	Discrete Categorical

Furthermore, no combination of Adj.FPs with another variable improved the Mantel's correlation. The new Jackknife estimator of Adj.FPs's R is $\bar{R} = 0.602$, $\mathrm{UCL} = 0.656$.

To ensure that the variable selected is not an artifact of any abnormal cases, we use sensitivity analysis to identify abnormal cases based on all 77 cases and given selected variable Adj.FPs.

Table 9 shows case 77 is an extremely influential data point in the data set. The exclusion of case 77 causes the Mantel's correlation to be reduced from 0.603 to 0.385 (R_{-77}), although the Mantel's correlation is still significantly greater than zero. This means that the results are strongly influenced by the specific data value, but that there is an underlying relationship that can be used for analogy-based estimation. Next, we remove project case 77 to ensure that the data set is stable (free of extreme cases).

Stepwise variable selection on the reduced 76 cases confirms that Adj.FPs is still the most significant variable and no combination of Adj.FPs with any other variable improved the Mantel's correlation. The new Jackknife estimator is then recalculated as $\bar{R} = 0.385$, $\mathrm{UCL} = 0.412$, $\text{p-value} = 0.001$. This concludes the Stage 1 stepwise analysis of all numerical variables.

TABLE 8
Mantel's Correlations for Each Project Factor Separately
(p Value Based on 1,000 Permutations)

dist(Variable)	Mantel R	p-value
Adj.FPs	0.603	0.001
RawFP	0.587	0.001
Transactions	0.498	0.001
Entities	0.254	0.004
ExpEquip	0.013	0.349
Adj Factors	0.099	0.078
Year.Fin	-0.061	0.802
ExpProjMan	-0.038	0.711

6.2 Stage 2: Variable Selection for Nominal Variables

The objective here is to investigate whether homogeneous subsets of project cases will improve Mantel's correlation. Using the reduced data set (76 cases) from Stage 1 and applying stepwise variable selection using the within-group Mantel's correlation technique, we found that the combination of Dev.Env with the variable Adj.FPs significantly improved the Mantel's correlation. The overall $R = 0.440$ is larger than $\mathrm{UCL} = 0.412$, based on Adj.FPs alone. This clearly indicates that Dev.Env should be included in any analogy-based prediction process. A further sensitivity analysis detected no abnormal cases. The final Jackknife estimator of Mantel's R is $\bar{R} = 0.440$ and its $\mathrm{UCL} = 0.466$.

6.3 Summary

The Analogy-X example presented in this section confirms that, for the Desharnais data set, case selection should be built upon the continuous variable Adj.FPs within the subgroups identified by the variable Dev.Env.

The variables selected using Analogy-X are the same as the variables selected using ANGEL's searching algorithms, so case selection will be identical for both methods. However, we have used a variable selection procedure based on Mantel's correlation as an alternative to Analogy's brute force and heuristic search algorithms.

7 Example 2: Sensitivity Analysis Applied to Random Data Sets

In the previous section, we demonstrated the application of Analogy-X to support stepwise variable selection on the Desharnais data set, as well as sensitivity analysis to ensure that influential observations are excluded from our model. To demonstrate what happens when Analogy-X is applied to data sets that are inappropriate for analogy, we show the effect of the Leverage statistic on two random data sets. The random data sets were generated to have similar properties to the Desharnais data set in terms of the number of cases and the distribution of variables, but were constrained to have no relationship between the dependent and independent

TABLE 9
Sensitivity Analysis for the Desharnais Data Set

$Case_i$	Mantel R	Mantel R_i	p-value$_i$	LM_i	$\|z_i\|$
77	0.603	0.385	0.001	0.218	8.162
41	0.603	0.653	0.001	0.049	1.891
51	0.603	0.634	0.001	0.031	1.190
76	0.603	0.634	0.001	0.030	1.183
44	0.603	0.623	0.001	0.019	0.765
46	0.603	0.620	0.001	0.017	0.670
…	…	…	…	…	…

variables. We generated 10,000 random data sets stratified according to Desharnais data set's mean and variance and selected two for analysis. For each random data set, we calculated the Mantel's correlation and selected the data set for which the correlation was largest (referred to as Desh-Random-0) and the data set for which the correlation was smallest (referred to as Desh-Random-1).

We applied our Leverage metric to each random data set for comparison. Each data point (R_i) is the Mantel's correlation excluding case i from the data set and its significance (p-value) is calculated using the permutation method. $Mantel\text{-}R$ is the Mantel's correlation for the entire data set.

Table 10 shows the largest five Leverage statistics for the Desh-Random-0 data set. In this case, the Mantel's R for the full data set is significantly different from zero, although the data set is random. Leverage analysis shows that one case (74) is extremely abnormal. Once it is removed from the data set, the Mantel's correlation is reduced from 0.357 to 0.088, which is not significantly different from zero. This implies that the relationship observed on the full data set is an artifact of case 74 and there is no underlying relationship.

Table 11 shows the largest five Leverage statistics for the Desh-Random-1 data set. In this case, there are abnormal data points, but there is no significant relationship in the data set even when the abnormal data points are excluded. Thus, Mantel's correlation correctly identified that there was no predictive relationship in the data set and the sensitivity analysis supports that conclusion.

The results from the two random data sets demonstrate that the Leverage statistic will identify abnormal cases. Subsequent sensitivity analysis based on reanalyzing the data set after removal of the high leverage cases will reduce the likelihood of building spurious prediction systems in a manner analogous to sensitivity analysis used to evaluate regression results.

8 Conclusion

The basic assumption underlying analogy-based estimation is that projects that are similar with respect to project features will be similar with respect to project effort. Analogy-X provides a means of formally testing this assumption using Mantel's correlation and randomization test for comparing two distance matrices. We have applied an approach similar to stepwise regression analysis in order to support feature selection using Analogy-X and have developed a Leverage metric to support sensitivity analysis.

In summary, Analogy-X's novelty is:

1. It delivers a statistical basis for analogy, which until now has been missing.
2. It is able to detect a statistically significant relationship and reject nonsignificant relationships.
3. It provides a simple mechanism for variable selection.

TABLE 10
High Leverage Cases for the Desh-Random-0 Data Set

77 cases Mantel-R: 0.357, p-value: 0.008)				
$Case_i$	R_i	p-value$_i$	LM_i	$\|z_i\|$
74	*0.088*	*0.105*	*0.269*	*8.147*
46	0.426	0.003	0.069	2.117
40	0.417	0.002	0.060	1.863
37	0.325	0.036	0.032	0.935
73	0.368	0.006	0.011	0.369
…	…	…	…	…

TABLE 11
High Leverage Cases for the Desh-Random-1 Data Set

(76 cases) Mantel-R: -0.019, p-value: 0.416				
$Case_i$	R_i	p-value$_i$	LM_i	$\|z_i\|$
6	*-0.040*	*0.607*	*0.021*	*4.539*
4	-0.001	0.361	0.018	3.853
16	-0.008	0.385	0.010	2.237
60	-0.009	0.370	0.010	2.173
41	-0.009	0.385	0.010	2.045
…	…	…	…	…

4. It is able to identify abnormal data points within a data set.
5. It supports sensitivity analysis that can detect spurious correlations in a data set.

All of these features are desirable and suggest that Analogy-X is a useful adjunct to the traditional analogy-based approach.

Analogy-X algorithms should be easily incorporated into current Analogy tools such as ANGEL [2] as an extension or a plug-in. The underlying procedures can be fully automated in the tool and would not be visible to the user other than to advise the user when analogy was not appropriate for the data set under investigation.

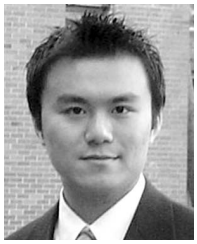
Like any other method, Analogy-X has limitations. Using categorical variables to partition the data set is not an attractive solution, particularly if there are a large number of such variables. However, we consider Analogy-X to be an important milestone for software effort estimation research because it provides a sound theoretical basis for analogy.

ACKNOWLEDGMENTS

National ICT Australia is funded through the Australian Government's Backing Australia's Ability Initiative, in part through the Australian Research Council. Funds were also provided by the University of New South Wales. The authors would like to thank Professor Martin Shepperd for his comments on a previous draft of this paper.

REFERENCES

[1] C. Schofield, "Software Support for Cost Estimation by Analogy," *Proc. Sixth European Software Cost Modeling Conf.,* 1995.
[2] C. Schofield and M.J. Shepperd, "Effort Estimation by Analogy: A Case Study," *Proc. Seventh European Software Control and Metrics Conf.,* 1996.
[3] M.J. Shepperd, C. Schofield, and B. Kitchenham, "Effort Estimation Using Analogy," *Proc. 18th Int'l Conf. Software Eng.,* 1996.
[4] B.W. Boehm, "Software Engineering Economics," *IEEE Trans. Software Eng.,* vol. 10, pp. 4-21, 1984.
[5] M.J. Shepperd and C. Schofield, "Estimating Software Project Effort Using Analogies," *IEEE Trans. Software Eng.,* vol. 23, pp. 736-743, 1997.
[6] T. Mukhopadhyay, S. Vincinanza, and M.J. Pietula, "Estimating the Feasibility of a Case-Based Reasoning Model for Software Effort Estimation," *MIS Quarterly,* vol. 16, pp. 155-171, 1992.
[7] K. Atkinson and M.J. Shepperd, "The Use of Function Points to Find Cost Analogies," *Proc. European Software Cost Modeling Meeting,* 1994.
[8] E. Stensrud and I. Myrtveit, "The Added Value of Estimation by Analogy: An Industrial Experiment," *Proc. Ninth European Software Control and Metrics Conf.,* 1998.
[9] F. Walkerden, "An Empirical Study of Analogy-Based Software Effort Estimation," *Empirical Software Eng.,* vol. 4, pp. 135-158, 1999.
[10] L. Angelis and I. Stamelos, "A Simulation Tool for Efficient Analogy Based Cost Estimation," *Empirical Software Eng.,* vol. 5, pp. 35-68, 2000.
[11] E. Mendes, N. Mosley, and S. Counsell, "Early Web Size Measures and Effort Prediction for Web Costimation," *Proc. Ninth Int'l Software Metrics Symp.,* pp. 18-39, 2003.
[12] E. Mendes and B. Kitchenham, "Further Comparison of Cross-Company and Within-Company Effort Estimation Models for Web Applications," *Proc. 10th Int'l Software Metrics Symp.,* pp. 348-357, 2004.
[13] I. Myrtvelt and E. Stensrud, "A Controlled Experiment to Assess the Benefits of Estimating with Analogy and Regression Models," *IEEE Trans. Software Eng.,* vol. 25, pp. 510-525, 1999.
[14] L.C. Briand, K. El Emam, D. Surmann, I. Wieczorek, and K.D. Maxwell, "An Assessment and Comparison of Common Software Cost Estimation Modeling Techniques," *Proc. 21st Int'l Conf. Software Eng.,* pp. 313-323, 1999.
[15] R. Jeffery, M. Ruhe, and I. Wieczorek, "Using Public Domain Metrics to Estimate Software Development Effort," *Proc. Seventh Int'l Software Metrics Symp.,* 2001.
[16] C. Mair and M. Shepperd, "The Consistency of Empirical Comparisons of Regression and Analogy-Based Software Project Cost Prediction," *Proc. Fourth Int'l Symp. Empirical Software Eng.,* pp. 491-500, 2005.
[17] M.J. Shepperd and G. Kadoda, "Using Simulation to Evaluate Prediction Techniques," *Proc. Seventh Int'l Software Metrics Symp.,* 2001.
[18] M.J. Shepperd and G. Kadoda, "Comparing Software Prediction Techniques Using Simulation," *IEEE Trans. Software Eng.,* vol. 27, no. 11, pp. 1014-1022, Nov. 2001.
[19] M. Jorgensen and D. Sjoberg, "Expert Estimation of Software Development Work," *Software Evolution and Feedback: Theory and Practice.* Wiley, 2006.
[20] M. Jorgensen, "A Review of Studies on Expert Estimation of Software Development Effort," *J. Systems and Software,* vol. 70, pp. 37-60, 2004.
[21] M. Jorgensen, "Practical Guidelines for Expert-Judgement-Based Software Effort Estimation," *IEEE Software,* vol. 22, pp. 57-63, 2005.
[22] M. Jorgensen, "Estimation of Software Development Work Effort: Evidence on Expert Judgement and Formal Models," *Int'l J. Forecasting,* 2007.
[23] C. Kirsopp, M. Shepperd, and J. Hart, "Search Heuristics, Case-Based Reasoning and Software Project Effort Prediction," *Proc. Genetic and Evolutionary Computation Conf.,* pp. 1367-1374, 2002.
[24] J. Li, G. Ruhe, A. Al-Emran, and M.M. Richter, A Flexible Method for Software Effort Estimation by Analogy, *Empirical Software Eng.,* http://www.kluweronline.com/issn/1382-3256, Apr. 2006.
[25] T. Foss, E. Stensrud, B. Kitchenham, and I. Myrtveit, "A Simulation Study of the Model Evaluation Criterion MMRE," *IEEE Trans. Software Eng.,* vol. 29, pp. 985-995, 2003.
[26] G.F. Kadoda, M. Cartwright, and M.J. Shepperd, "Issues on the Effective Use of CBR Technology for Software Project Prediction," *Proc. Fourth Int'l Conf. Case-Based Reasoning: Case-Based Reasoning Research and Development),* pp. 276-290, 2001.
[27] N. Mantel, "The Detection of Disease Clustering and a Generalized Regression Approach," *Cancer Research,* vol. 27, pp. 209-220, 1967.
[28] B.F.J. Manly, *Randomization, Bootstrap and Monte Carlo Methods in Biology,* second ed. Chapman & Hall/CRC, 1997.
[29] P. Legendre and L. Legendre, *Numerical Ecology,* second ed. Elsevier, 1998.
[30] B.F.J. Manly, *Multivariate Statistical Methods—A Primer,* second ed. Chapman & Hall/CRC, 1998.
[31] F.H.C. Marriott, "Barnard's Monte Carlo Tests: How Many Simulations?" *Applied Statistics,* vol. 28, 1979.
[32] R-Project, "The R Project for Statistical Computing," http://www.r-project.org, 2005.
[33] ADE4, "Ecological Data Analysis (ADE4) Package for R," http://pbil.univ-lyon1.fr/ADE-4/, 2004.
[34] VEGAN, "Vegan: R Functions for Community Ecology," http://cc.oulu.fi/~jarioksa/softhelp/vegan.html, 2004.
[35] J.W. Tukey, "Accurate Confidence Interval for the Ratio of Specific Occurrence/Exposure Rates in Risk and Survival Analysis," *Biometrical J.,* vol. 37, p. 611, 1958.
[36] B. Efron and G. Gong, "A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation," *The Am. Statistician,* vol. 37, pp. 36-48, 1983.
[37] N.R. Draper and H. Smith, *Applied Regression Analysis,* third ed. John Wiley and Sons, 1998.



Jacky Wai Keung received the BS degree (Hons) in computer science from the University of Sydney and the PhD degree in software engineering from the University of New South Wales for his research into statistical methods of software cost estimation. He is a researcher for the Empirical Software Engineering Research Group, National ICT Australia (NICTA), ATP Sydney, where he works in a range of technical roles including consulting in software measurement and cost estimation for a number of software development organizations in Australia and other countries. He also holds an academic fellow position in the School of Computer Science and Engineering, University of New South Wales. His research interests include software measurement and its application to project management, software resource and cost estimation, software quality and software engineering process, and product modeling. His most recent research has focused on the application of analogy-based systems to software cost estimation. He is a member of the Australian Computer Society and the IEEE.

Barbara A. Kitchenham is a professor of quantitative software engineering at Keele University, Keele, United Kingdom. From 2004 to 2007, she was a senior principal researcher at the National ICT Australia. She has worked in software engineering for nearly 30 years, both in industry and academia. Her main research interest is software measurement and its application to project management, quality control, risk management, and evaluation of software technologies. Her most recent research has focused on the application of evidence-based practice to software engineering. She is a chartered mathematician and a fellow of the Institute of Mathematics and Its Applications, a fellow of the Royal Statistical Society, and a member of the IEEE Computer Society.

David Ross Jeffery received the PhD degree from the University of New South Wales. He is the research group manager for Managing Complexity at the ATP Laboratory, National ICT Australia (NICTA). He also leads the Empirical Software Engineering Group in this laboratory. His research interests include software engineering process and product modeling, electronic process guides and software knowledge management, software quality, software metrics, software technical and management reviews, and software resource and cost estimation. He is an elected fellow of the Australian Computer Society, the founding chairman of the Australian Software Metrics Association, a member of the ACM, and a member of the IEEE Computer Society.