Experiments with Analogy-X for Software Cost Estimation

Jacky Keung[1] , Barbara Kitchenham[2]
[1]*National ICT Australia Ltd, Australia*
[2]*Keele University, UK.*
[1]*Jacky.Keung@nicta.com.au*, [2]*b.a.kitchenham@cs.keele.ac.uk*

Abstract

We developed a novel method called Analogy-X to provide statistical inference procedures for analogy-based software effort estimation. Analogy-X is a method to statistically evaluate the relationship between useful project features and target features such as effort to be estimated, which ensures the dataset used is relevant to the prediction problem, and project features are selected based on their statistical contribution to the target variables. We hypothesize that this method can be (1) easily applied to a much larger dataset, and (2) also it can be used for incorporating joint effort and duration estimation into analogy, which was not previously possible with conventional analogy estimation. To test these two hypotheses, we conducted two experiments using different datasets. Our results show that Analogy-X is able to deal with ultra large datasets effectively and provides useful statistics to assess the quality of the dataset. In addition, our results show that feature selection for duration estimation differs from feature selection for joint-effort duration estimation. We conclude Analogy-X allows users to assess the best procedure for estimating duration given their specific requirements and dataset.

Keywords: Software effort prediction, duration prediction, case-based reasoning, analogy, Mantel's correlation, Analogy-X, ISBSG

1. Introduction

Software effort estimation by analogy is an approach based on retrieving one or more past completed software projects that resemble the new target project that is to be estimated. The Analogy-based approach is quite well known to software researchers [1], and is supported by tools such as Shepperd's ANGEL tool [2].

One of the challenges in the analogy-based approach is the identification of relevant project features to be used in its application [2]. Related literature reveals that a number of search algorithms are currently in use for selecting useful project features, but none of these algorithms provide useful statistics to determine the quality and the relevance of the features selected by using one of the search techniques. In most cases, features are selected using brute-force and other forms of search heuristics, which requires a large amount of computing power and time to complete its entire search function, yet provides no evidence to support its identified features for estimation purposes. This is especially problematic when dealing with large datasets.

Furthermore, software development duration estimation is something of an afterthought. For example, in COCOMO duration is estimated using a model that uses effort as an input. In practice, the model would be driven by the effort estimate obtained from the effort estimation model [3]. Using regression analysis, researchers have proposed estimating duration using either standard product factors or effort estimates [4]. Putnam's SLIM model is a notable exception, since it is based on a model that relates size to effort and duration [5].

We have developed a solution called Analogy-X in [6, 7] to overcome many of these issues when selecting relevant project features based on Mantel's correlation statistics and randomization tests.

In this paper, we apply Analogy-X on a much larger publicly available dataset to evaluate its performance and its capability in dealing with outlying data points. Also we argue that it is possible to determine a set of project features that are appropriate for joint effort and duration prediction. We demonstrate this on a company dataset where duration and effort are dependent variables. Our method can also identify whether effort and duration should be estimated using different feature sets. However, if this is the case, we must caution estimators (or practitioners) to be aware that the joint feasibility of the estimates cannot be confirmed.

Section 2 provides the background and an outline of our research approach and its applications to large dataset and joint effort-duration estimation. Section 3 defines Mantel's correlation and randomization test in more detail and explains our approach to feature selection. In Section 4, we present an experiment using Analogy-X on a large dataset. We define the dataset we use to illustrate our approach to joint effort-duration estimation and present our analysis in Section 5. We discuss and conclude our results in Section 6.

2. Related Work

2.1 The Basic Assumption Underlying Analogy

Our recent research has proposed a method for assessing whether data-intensive case-based reasoning (sometimes referred to as analogy) is an appropriate method for predicting effort on a specific dataset [7] [6]. Although usually unstated, the basic hypothesis underlying the use of data-intensive case-based reasoning for software project effort estimation is:

"Projects that are similar with respect to project and product factors such as size and complexity will be similar with respect to project effort."

Based on this principle, tools such as ANGEL [2] [8], compute a similarity measure using project and product features between a new project and projects in an historical database [2]. An effort estimate for the new project is then based on the actual effort of the *k* most similar projects in the database. The value of *k* is determined by trial and error for particular dataset. There are several alternative strategies for constructing the estimate for the new project, for example a simple average of the *k* most similar projects, or a weighted average.

One major problem with this method is that tools such as ANGEL will provide an estimate even if the data set is completely inappropriate for case-based estimation [8]. However, our recent research [7] has identified a method for testing whether the hypothesis underlying analogy is valid for a particular dataset that is analogous to assessing whether a regression line produces a statistically significant fit for a particular dataset.

2.2 An eXtension for Analogy

Our approach, called Analogy-X [7] [6], uses Mantel's correlation and randomization test to test the basic hypothesis. To test that a dataset is appropriate for case-based reasoning, we construct a similarity matrix for effort as well as a similarity matrix for project factors. We then construct Mantel's correlation from the related elements of each similarity matrix.

If projects that are similar with respect to project features are also similar with respect to project effort, a correlation derived from the similarity matrix element for effort and the corresponding similarity matrix element for project features will be relatively large. If there is no association between pairs of similarity matrix elements, Mantel's correlation (R_M) will not be significantly different from zero and the data set is inappropriate for case-based estimation.

We use Mantel's Randomization test to test whether the value of Mantel's R_M is significantly different from zero. We have also developed a process for stepwise feature selection and sensitivity analysis based on a Jack-knife method [6] [9] used to provide confidence limits on values of R_M that are significantly different from zero.

We have empirically evaluated Analogy-X on a number of small datasets (such as Kemerer, Albrecht, Desharnais and Telecom-1 [2, 10]). Results show that it enables dataset quality evaluation and provides a more robust feature selection process for analogy. In summary, Analogy-X's novelty is:

- It delivers a statistical basis for analogy, which until now has been missing.
- It is able to detect a statistically significant predictive relationship and reject non-significant predictive relationships.
- It provides a simple mechanism for project feature selection.
- It is able to identify abnormal data points within a dataset.
- It supports sensitivity analysis that can detect spurious correlations in a dataset.

2.3 The differences in the estimation processes

The following (Figure 1) overviews the difference between a classical analogy-based system, stepwise regression and Analogy-X. It clearly shows the main difference between ANGEL and Analogy-X lies in the pre-processing dataset evaluation stage, where exhaustive search algorithms can be replaced by a more systematic and robust Analogy-X system. Analogy-X is able to reject the hypothesis when there is no suitable analogy model for the dataset, in a much similar fashion to that of stepwise regression. The application process of Analogy-X remains the same to that of classical analogy (i.e. ANGEL), where potential source analogues must be first identified and case adaptation performed to derive a final estimate.

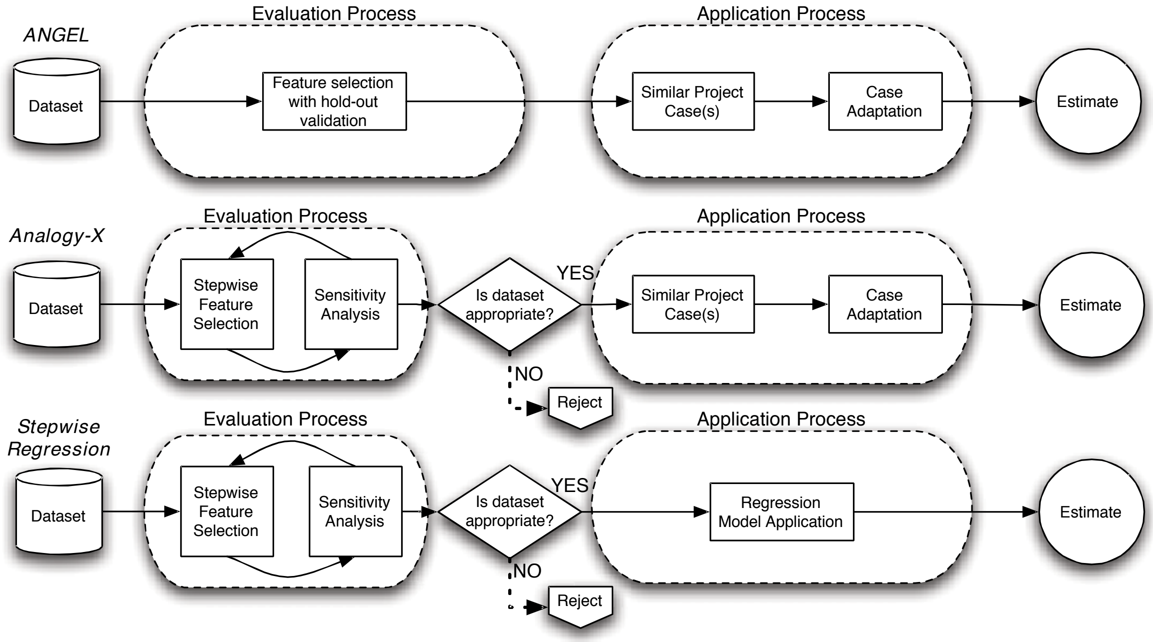

Figure 1 Estimation Processes of Analogy (ANGEL), Analogy-X and Stepwise Regression.

3. Analogy-X Methodology

This section provides the principle of the Mantel statistics and an outline of our Analogy-X approach.

3.1 Mantel's Correlation and Randomization test

Mantel's correlation for comparing two distance or dissimilarity matrices was first introduced as a solution to the problem of detecting space and time clustering of diseases for cancer research [11]. It has since been widely adopted in ecology, biology and psychology researches to address this kind of problem [12].

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 *X*s (environmental variables) also tend to be close with respect to *Y*s (species variables)?" The question is analogous to the questions we want to ask in CBR-based software cost estimation approach i.e. "Do projects that are close with respect to *X*s (project and product features) also tend to be close to *Y*s (development effort)?"

Although Mantel discussed more general situations and findings in his original study, Manly provides more comprehensive examples of Mantel's method [12, 13]. The basic principle of Mantel's method is to measure the association between the corresponding elements in two distance matrices by a suitable statistic, usually the Pearson correlation statistic. 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 re-ordering the elements in one of the distance matrices. The normal statistical tests for the Pearson correlation coefficient are inappropriate in this case because the elements in a distance matrix are not independent.

Assuming distance matrices *A* for predictor variables and *B* for response variables are constructed as follow:

$$A = \begin{bmatrix} 0 & a_{12} & \cdots & a_{1n} \\ a_{21} & 0 & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & 0 \end{bmatrix} \quad B = \begin{bmatrix} 0 & b_{12} & \cdots & b_{1n} \\ b_{21} & 0 & \cdots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \cdots & 0 \end{bmatrix} \qquad (1)$$

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 (*x1*) and case 2 (*x2*) is calculated using simple Euclidean distance:

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

Equation (2) considers the values of all *p* variables for each pair of cases. Note that 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 usual transformation is to divide each value by the difference between the maximum and minimum value.

Because of symmetry, only the lower diagonal elements in the above matrices (Equation 1) need to be considered when constructing and testing the Mantel's correlation. The Mantel correlation coefficient is:

$$R_M = \frac{\sum a_{ij} b_{ij} - \sum a_{ij} \sum \frac{b_{ij}}{m}}{\sqrt{\left[\left\{\frac{\sum a_{ij}^2 - \left(\sum a_{ij}\right)^2}{m}\right\} \times \left\{\frac{\sum b_{ij}^2 - \left(\sum b_{ij}\right)^2}{m}\right\}\right]}} \qquad (3)$$

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

$$m = \frac{n(n-1)}{2} \qquad (4)$$

For the randomizations test the distance matrix elements are randomly permuted for one of the matrix, matrix A (Equation 1) says. For example one randomization of the elements of A, gives the matrix A_{Random} (Equation 5):

$$A_{Random} = \begin{bmatrix} 0 & a_{68} & a_{18} & \cdots & a_{38} \\ a_{68} & 0 & a_{16} & \cdots & a_{36} \\ a_{18} & a_{16} & 0 & \cdots & a_{31} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{38} & a_{36} & a_{31} & \cdots & 0 \end{bmatrix} \qquad (5)$$

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 correlation is then computed using matrix B (Equation 1) and A_{Random} (Equation 5).

Repeating the same procedure many times produces the randomization statistic distribution. Using the randomization distribution we can test whether the value of the Mantel correlation derived from the original pair of distance matrices is significantly different from zero. If the Mantel 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 dataset under investigation.

The traditional correlation approach do not work for correlating distance matrices, as change in the order of cases will also change the distance matrix elements, which violates the assumption of independence in correlation. This randomization procedure is critical as it removes the independence issue and provides a suitable test of significance for distance matrices correlation.

Based on Marriott [14], Manly [12] notes that 1,000 randomizations is a realistic minimum for estimating a significance level of about 0.05 and 5,000 randomizations is a realistic minimum for estimating a significance level of about 0.01.

3.2 Stepwise Feature Selection in Analogy-X

An important element of Analogy-X is to use Mantel's R_M to support stepwise feature selection. This was never considered as an issue by Mantel or Manly, but it is important in software cost estimation because there may be many features, some of which may not be relevant for identifying projects that are similar with respect to effort (or duration) and may distort measures of similarity.

The stepwise procedure depends on being able to assess whether an increase in the value of R_M is statistically significant or not. We cannot use the randomization test for this because the randomization test is only applicable for testing the null hypothesis that $R_M = 0$. When $R_M > 0$, we used a Jack-knife method to construct a Jack-knife estimate of R_M and the 95% confidence intervals for the Jack-knife estimate. Then when we add a new feature and obtain an increased value of R_M, the new feature is assumed to have made a significant contribution to assessing similarity if the new value of R_M is larger than the upper 95% confidence limit constructed for the previous jack-knife estimate of R_M.

The Jack-knife estimate is obtained by omitting one project at a time from the dataset and calculating R_i which is the Mantel's correlation for the dataset excluding project i. The jack-knife estimate of R_M for a dataset of size n is

$$\hat{R}_M = \left(\sum_{i=1}^{n} R_i\right) \Big/ n \qquad (6)$$

$\hat{R}_M$ will be approximately normally distributed with variance:

$$S^2 = \sum_{i=1}^{n} (R_i - \hat{R}_M)^2 \Big/ (n-1) \qquad (7)$$

The UCL (Upper 95% Confidence Limit for $\hat{R}_M$) is

$$\hat{R}_M + 1.96 \times S \qquad (8)$$

This Jack-knife method also supports sensitivity analysis, since it is possible to identify individual projects which have a significant impact on the value of $\hat{R}_M$ [6, 7].

4. Experiment 1 - Large Dataset

Experiment 1 was designed to evaluate the performance and the efficacy of Analogy-X on a much larger dataset. The ISBSG release 9 was selected for this purpose. This dataset is believed the most well known publicly available industrial dataset collected by the International Software Benchmarking Standards Group (ISBSG). Many analogy studies have been based on the ISBSG dataset such as in [15].

The dataset used in this study is the latest release (version 9) that contains over 3,000 software projects. Project entries in the dataset have been carefully validated by ISBSG organization, and a quality rating for the credibility of each project is given. Nevertheless, there are many projects with missing values in the dataset that requires imputation and other missing value techniques. For this reason, only a portion of the dataset is used. 10 useful project features are considered, and only projects with quality rating "A" are considered, projects with missing values are discarded. This resulted in 502 completed project cases in the dataset. The following table shows the project features of the 502 projects from the ISBSG (v9) dataset. Dependent variables are Effort and Duration of the completed projects in the dataset.

Table 1: Project Features of the ISBSG dataset used in the experiment.

Feature	Description	Mantel-R	p-value
Adj.FPs	Adj. Function Points	0.6141	0.001
Input	No. of Input	0.5156	0.001
Output	No. of Output	0.4745	0.001
Enquiry	No. of Enquiry	0.4782	0.001
File	No. of File	0.5566	0.001
Interface	No. of Interface	0.2326	0.007
Added	No. of Added Features	0.6258	0.001
Changed	No. of Changes	0.0920	0.041
Deleted	No. of Deleted Features	0.0563	0.051
Resource	Resource Level	0.2147	0.001

Given the size of the dataset, traditional analogy-based systems (such as ANGEL [2]) require a lot more time to compute, and is not able to identify and remove unrelated cases. This dataset is used to evaluate Analogy-X for its scalability in circumstances where large number of project features and cases exist in the dataset.

To compare the result with ANGEL, the reduced ISBSG dataset has been imported to the ANGEL system, and used its brute-force feature to detect useful features, Because of the large size of the dataset, ANGEL took more than 3 hours to execute, and when we stopped it manually, its feature selection process had not finished. The progress was between 1 to 5 percent completed. There were 4 different features selected by ANGEL, which seemed reasonable, the *AdjFPs* (Adjusted Function Points), *Output* (Number of Outputs), *Interface* (Number of Interfaces), and *Added* (Number of Added Features).

Applying the Analogy-X stepwise analysis to the ISBSG (v9) dataset, the distance matrix based on *Added* + *Adj.FP*s was significantly correlated with the Actual Effort distance matrix (see Table 2).

Table 2: ISBSG reduced dataset (502 cases, 14 features)

Iteration	Feature subset	Mantel R	p-value	Jackknife R	UCL
1	*Added*	0.6111	0.001	0.6230	0.6258
2	*Added* + *Adj.FPs*	0.6298	0.001	0.6258	0.6303

The Jackknife procedure used to derive the estimator of Mantel *R* is quite computation intensive, especially with a large number of project cases in the dataset. It took about 1 hour to fully compute and validate the results it produced. However, it is much more robust and efficient than the brute-force search technique.

Analogy-X's sensitivity analysis applied based on the distance matrix of *Added* and *AdjFPs*. We used the following leverage metric equation:

$$LM_i = R_i - \hat{R} \tag{9}$$

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 $\hat{R}_M$ and will be approximately $N(0,S^2)$. 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:

$$z_i = \frac{LM_i}{S} \tag{10}$$

The result shows three projects are extreme cases, see Table 3 below.

Table 3: High Leverage cases for ISBSG dataset

502 cases
Mantel-R: 0.6298, *p-value*: 0.001

Cases	Ri	*p*-value	LM	|*z*|
124	**0.669068**	**0.001**	**0.03900**	**18.646**
111	**0.619394**	**0.001**	**0.10594**	**5.0925**
437	**0.621024**	**0.001**	**0.00896**	**4.3135**
149	0.623079	0.001	0.00691	3.3314
285	0.636704	0.001	0.00672	3.1798
205	0.636510	0.001	0.00652	3.0871
…	…	…	…	…

Table 3 highlights the largest three extreme project cases with their leverage statistics (|z| > 4) for the extracted ISBSG dataset. In this case, the slightest variation in R_i significantly impacts the value of |*z*|, this is because the large sample size of the dataset (502 cases). Leverage analysis shows that these three cases are extremely abnormal. This also implies that the relationship observed on the full dataset is an artifact of these three abnormal cases although underlying relationship exists (p-value: 0.001).

Analogy-X's stepwise feature selection was reapplied with these three cases removed. The result confirms that project feature *Added* and *Adj.FPs* are still the most influential variables, with an improved Mantel's correlation coefficient. The final Mantel's correlation for features Added and *Adj.FPs* is 0.6551, p-value=0.001, the Jackknife estimator of Mantel R is $\hat{R}_M$ = 0.6552 with UCL = 0.6595. No further abnormal project cases exist in the dataset.

The feature subset selected using Analogy-X in here is different to the feature subset selected by ANGEL, however it is reasonable to believe that the combination of features *Added* and *Adj.FPs* will be useful for analogy-based prediction. Further examination on the Mantel's R for the ANGEL selection of features (*Adj.FPs, Output, Interface and Added*) without the 3 abnormal cases show that Mantel's R is 0.6046 and confirm that it is less than the UCL(0.6595) for features *Added* and *Adj,FPs* identified by Analogy-X. More importantly, Analogy-X provides statistical evidence to support its claim.

The experiment demonstrates large dataset, which prevents the use of ANGEL. It also shows that Analogy-X is capable of detecting abnormal project cases in a large dataset.

5. Experiment 2 – Multiple Dependencies

When it comes to case-based reasoning (or analogy) estimation, we found no examples of duration estimation. It may be that advocates of CBR assume that the similar projects will by definition be similar both with respect to effort and duration. This approach does ensure that effort and duration estimates are jointly feasible because they derive from the same project(s). However, it also assumes that the features included in CBR, which are optimized to effort estimation accuracy, are also suitable for duration prediction. Alternatively researchers may assume that duration prediction should be treated as a separate activity with feature selection optimized to duration estimation accuracy. However, this may overlook issues such as the joint feasibility of effort and duration estimates for a specific project.

An extremely important aspect of Mantel's method used in Analogy-X is that it does not impose any restrictions on the number of variables in either similarity matrix. This raises the possibility of testing whether projects similar with respect to various project and product features are similar with respect to both effort and duration (see Figure 2). This means, we can investigate whether a particular dataset can act as the basis for joint effort and duration predictions.

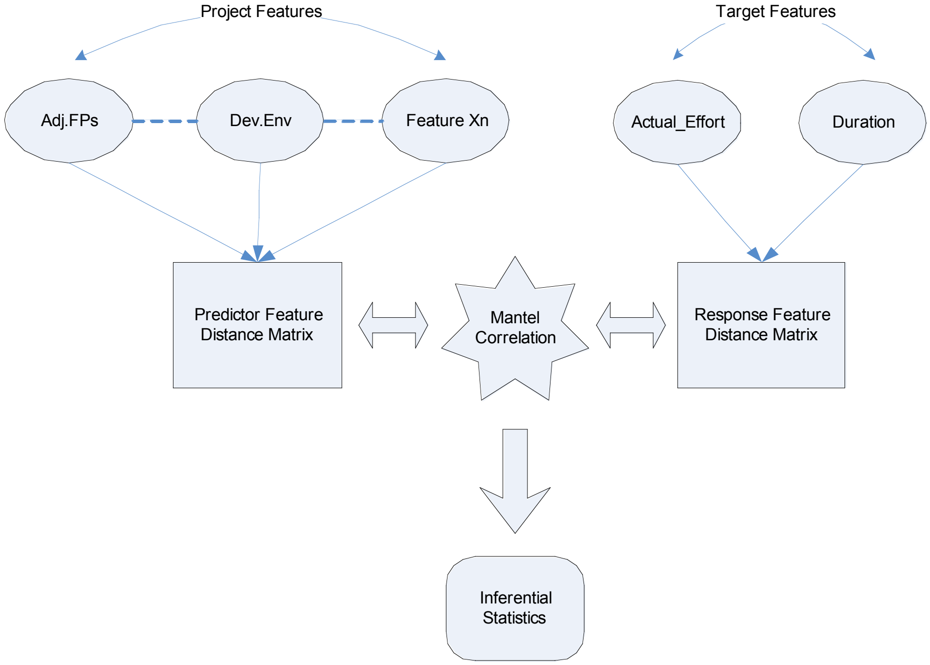

Figure 2 Multiple Dependencies in Analogy-X

5.1 Analysis Procedures

We propose the following procedure to investigate effort and duration estimation using CBR.

Step A: Determine the viability of joint effort and duration estimation using CBR

A-1. Construct a distance matrix with elements based on effort and duration.

A-2. Use Mantel's correlation and randomization test plus a stepwise feature selection process to test whether joint effort-duration estimation is feasible.

A-3. If joint estimation is not feasible, go to step B

A-4. If joint effort-duration estimate is feasible for the dataset, you may stop at this point, however, you may also want to check whether joint estimation is less accurate than estimating effort and duration separately.

Step B: Assess impact of separate estimation of duration and effort

B-1. Apply the Analogy-X method to effort and duration separately.

B-2. If Mantel's R_M is not significantly different from zero for either effort or duration, the data set is inappropriate for estimation based on CBR.

B-3. If the Mantel's R_M is significant for either effort or duration but not both, estimating both factors using CBR is inappropriate.

B-4. If the Mantel's R_M is significant for both outcomes, review the selected project factors.

- If they are the same, then for this dataset projects used to estimate effort can also be used to estimate duration.
- If the selected projects factors are different for duration and effort estimation, and Step A produced a significant Mantel's correlation, the Step A similarity matrix should be used for estimation to ensure that the combined duration and effort estimates are jointly feasible.
- If the selected projects factors are different for duration and effort estimation, and Step A did not produce a significant Mantel's correlation, effort and duration should be estimated separately. However, we cannot assume that the combined estimates are jointly feasible for a single project because the effort and duration estimates for a new project may be extracted from different projects.

This procedure is not intended to show joint effort-duration estimation a better approach however it provides a method to evaluate the effect of using joint estimation for the dataset under investigation.

5.2 The CSC Dataset

The dataset used in this study was colleted by Computer Sciences Corporation (CSC). The data set comprised 145 projects. The data set comprised 145 projects. The data collected on each project included:

Table 4: Project Features of CSC dataset

Independent Variable	Description	Type
Client	Client / Site	Categorical
ProjType	Project Type	Categorical
FPs	Adjusted Function Points	Numerical
RawFP	Raw Function Points	Numerical
CompFactor	Complexity Factor	Numerical
Input	Number of Inputs	Numerical
Output	Number of Outputs	Numerical
Internal	Number of Internal Files	Numerical
Interface	Number of Interfaces	Numerical
Inquiry	Number of Inquiries	Numerical
Dependent Variable	**Description**	**Type**
Duration	Actual Duration	Numerical
Effort	Actual Effort	Numerical

The full dataset is reported in [4]. The dataset is a subset of CSC projects for which all effort and size data was available. The projects were undertaken between 1994 and 1998. In this context a "project" represents a specific maintenance change to an existing application or a new product development.

5.3 Effort and Duration Estimation

Step A: Determine the viability of joint effort and duration estimation using CBR

The stepwise analysis of Analogy-X is applied to the continuous variables of the original CSC dataset (all 145 cases) to test whether joint estimation (*Effort* and *Duration*) is feasible. Table 5 identifies the Mantel's *R* obtained for each continuous variable separately.

Table 5: Mantel correlation for each project factor separately (*p*-value based on 1,000 permutations)

dist(Variable)	Mantel-R	p-value
RawFP	**0.7598**	**0.001**
FPs	0.7572	0.001
Output	0.7561	0.001
Inquiry	0.7548	0.001
Input	0.7548	0.001
Internal	0.7357	0.001
CompFactor	0.0673	0.112
Interface	-0.0212	0.525

In this case, *RawFP* was the variable for which Mantel's R was greatest. Furthermore no combination of *RawFP* with another variable improved the Mantel correlation. The new Jackknife estimator of *RawFP* is $\hat{R}_M$ = 0.7573, with an upper 95% confidence interval limit UCL = 0.8422.

To ensure the variable selected (*RawFP*) is not an artifact of any abnormal cases, we use the sensitivity analysis procedure of Analogy-X to identify abnormal cases based on all 145 cases, given selected variable, *RawFP*. Without applying such a procedure, spurious correlation may be introduced. In particular certain data point(s) in the dataset may heavily influence the analysis.

Table 6: Sensitivity Analysis for CSC Dataset (Joint Effort-Duration)

Case i	Mantel R	Mantel Ri	p_value	LMi	|Zi|
102	***0.759756***	***0.424850***	***0.000999***	***0.334906***	***11.91347417***
9	0.759756	0.778371	0.000999	0.018615	0.70953093
3	0.759756	0.776124	0.000999	0.016368	0.62929838
38	0.759756	0.769832	0.000999	0.010076	0.40463295
51	0.759756	0.766481	0.000999	0.006725	0.28498039
…	…	…	…	…	…

Table 6 shows the largest five Leverage statistics for the CSC dataset. Sensitivity analysis shows that one case (102) is extremely abnormal, and is a influential data point in the dataset. After its exclusion from the dataset the Mantel correlation is reduced from 0.759756 to 0.424850, the *p*-value shows the correlation is still statistically significant. This implies that the relationship observed on the full dataset (145 cases) is an artifact of case 102 but there is still an underlying predictive relationship. Thus, analogy is appropriate estimation method for this dataset upon the removal of case 102.

Once case 102 is removed from the dataset, and the stepwise analysis reapplied to the reduced dataset (144 cases) the revised Mantel's R-values for each of the numerical variables separately is shown in Table 7.

Table 7: Mantel Correlation for each project factor separately based on the reduced dataset (144 cases, p-value based on 1,000 permutations)

Dist(Variable)	Mantel-R	p-value
FPs	***0.4371***	***0.001***
Raw*FP*	0.4249	0.001
Output	0.3596	0.001
Inquiry	0.3243	0.002
Input	0.2666	0.001
Internal	0.2118	0.003
Interface	0.0597	0.159
CompFactor	-0.0169	0.591

In this case, *FP* exhibited the largest Mantel's R. Furthermore no combination of *FP* with another variable including categorical variables (*ClientCode* and *ProjectType*) improved the Mantel correlation. The new Jackknife estimator of *RawFP*'s is $\hat{R}_M$ = 0.4371, with an upper confidence interval limit UCL = 0.4513. These results confirm that joint effort-duration estimate is feasible using adjusted function points as the only variable for similarity assessment.

Step B: Assess the impact of separate estimation of effort and duration

In Step A, we demonstrated the application of Analogy-X to support multiple outcome variables. Result shows the single variable *FPs* can be used to predict both effort and duration jointly. In this section, we apply the same Analogy-X procedures to predict effort and duration separately and compare the results with joint effort-duration prediction.

Applying the stepwise analysis to the CSC dataset (all 145 cases), the *Inquiry* distance matrix has a greatest Mantel's correlation with response distance matrix of *Duration*.

Table 8: Mantel Correlation for each project factor separately based on the CSC dataset (145 cases, p-value based on 1,000 permutations)

Dist(Variable)	Mantel-R	p-value
Inquiry	***0.5837***	***0.001***
Raw*FP*	0.5739	0.001
Output	0.5699	0.001
*FP*s	0.5687	0.001
Input	0.5506	0.001
Internal	0.5194	0.001
CompFactor	0.0379	0.238
Interface	-0.0338	0.646

Table 8 shows variable *Inquiry* is the most influential predictor and no combination of *Inquiry* with another variable improved the Mantel's correlation. The Jackknife estimator of Inquiry is $\hat{R}_M$ = 0.5825, with an upper confidence interval limit UCL = 0.6394.

To ensure the variable *Inquiry* is not an artifact of any abnormal cases, we use the same sensitivity analysis performed in Step A. (See Table 9)

Table 9: Sensitivity Analysis for CSC dataset (Duration as Response Variable)

Case i	Mantel R	Mantel Ri	p_value	LMi	|Zi|
102	***0.583714***	***0.244863***	***0.000999***	***0.338851***	***11.8770316***
9	0.583714	0.609817	0.000999	0.026103	0.95957984
38	0.583714	0.597533	0.000999	0.013819	0.52751187
30	0.583714	0.596680	0.000999	0.012966	0.49750910
41	0.583714	0.594386	0.000999	0.010672	0.41682171
…	…	…	…	…	…

Based on the statistics show in Table 9, and similar to what we found in step A, the exclusion of case 102 causes the Mantel correlation to be reduced from 0.583714 to 0.244863 in the case of predicting duration.

The removal of case 102 changed the best feature subset found using Analogy-X stepwise procedure as explained below.

Table 10: Mantel Correlation for each project factor separately based on the reduced dataset (144 cases, p-value based on 1,000 permutations)

First Iteration: (1st Variable)

Dist(Variable)	Mantel-R	p-value
Inquiry	***0.2449***	***0.001***
RawFP	0.2231	0.001
FPs	0.2038	0.002
Output	0.2035	0.001
Input	0.1088	0.047
Internal	0.0876	0.075
Interface	-0.0356	0.696
CompFactor	-0.0334	0.741

The Jackknife estimator of *Inquiry* is $\hat{R}_M$ = 0.2449, with an upper confidence interval limit UCL = 0.2670.

Furthermore, the combination of features *Inquiry* and *Output* produce an improved Mantel correlation of 0.2874, which is greater than the UCL of the Jackknife estimator of *Inquiry* (see Table 11). Therefore we include *Output* in the final feature set. No other variable had a significant impact on Mantel's correlation.

Table 11: Mantel Correlation for each project factor separately based on the reduced dataset (144 cases, p-value based on 1,000 permutations)

Second Iteration (Inquiry + 2nd Variable):

Dist(Variable)	Mantel-R	p-value
Output	***0.2874***	***0.001***
RawFP	0.2609	0.001
FPs	0.2555	0.001
Input	0.2230	0.001
Internal	0.2028	0.002
Interface	0.1955	0.005
CompFactor	0.1617	0.002

The final Jackknife estimator of Mantel's *R* using the two features *Inquiry* and *Output* is $\hat{R}_M$ = 0.2874, with an upper confidence interval limit UCL = 0.3048.

Based on the duration distance matrix, Analogy-X identified *Inquiry* and *Output* as the two most influential features. Thus, optimizing estimation for duration leads to a different selection of features to optimizing for joint effort-duration estimation. We used both sets of features for analogy-based duration prediction. Then we compared prediction accuracy using a paired *t*-test of the absolute residuals to assess whether the two set of predictions were significantly different (as illustrated in Table 12).

Table 12: Absolute Residual of Prediction using Duration as the response and using Joint Effort-Duration as the response

Project	Actual Duration (Target)		Duration only prediction	Difference Absolute Residuals		Duration Estimation from Joint Effort-Duration Prediction	Difference Absolute Residuals
1	107		88	19		462	355
2	144		95	49		186	42
3	604		120	484		183	421
4	226		82	144		155	71
5	328		186	140		192	134
6	294		238	56		109	185
…	…		…	…		…	…
144	92		207	115		127	35
Mean	**201.32**		**197.39**	117.46		**208.85**	125.57
Median	**170.00**		**164.00**	81.00		**181.50**	70.50

Our result shows that the overall predictions using these 2 strategies are (Mean=197.39, Median=164.00) person months and (Mean=208.85, Median=181.50) person months respectively. The paired *t*-test performed on the predictions produced by these two strategies shows a *p*-value of 0.3973, which means that the prediction accuracy obtained by using the *Inquiry + Output* distance matrix or by using the *FPs* distance matrix are not significantly different. Using either feature subset produces very similar predictions.

Using Analogy-X with *effort* as the only dependent variable, *FPs* is the only significant project features (after the removal of case 102). Thus, for the CSC dataset, *FPs* is suitable both for effort prediction and for joint effort-duration prediction. Because the feature used to search for analogues is the same, the effort predictions are the same in both cases, and no further comparison of these two different strategies is required.

6. Discussion and Conclusions

The two experiments presented in this paper are particularly important for CBR or Analogy based estimation, because conventional analogy-based applications with large datasets has proved problematic and multiple outcome variables has not been the subject of analogy research in the past.

(1) We have demonstrated our procedure using a large publicly available dataset. With 502 projects analogy-X successfully identified three anomalous cases and selected the most useful project features for analogy-based estimation. ANGEL was unable to complete analysis of this dataset in three hours whereas our method analyzed the dataset in about one hour. This confirms that our method is able to handle larger datasets than ANGEL but that it is itself fairly data intensive. A more detailed analysis of the limits of Analogy-X and ANGEL needs to be based on simulation studies where the impact of increasing the number of cases and variables can be studied systematically.

(2) We have also demonstrated our procedure for effort and duration on a publicly available company dataset. For this specific dataset, we found that the project feature used for joint effort-duration estimation differed from the project features selected for duration estimation. Although there was no significant difference between the predicative accuracy of duration prediction for the different feature sets, it is clearly possible that optimizing analogy for duration estimation might sub-optimize for joint effort-duration estimation. Our experiment demonstrates how our estimation procedure provides an empirical method for testing whether duration estimation using CBR is appropriate in the context of a specific data set. In particular, Analogy-X can identify projects that are similar both with respect to effort and duration, and produce an estimate that is feasible jointly for effort and duration.

7. Acknowledgements

National ICT Australia is funded through the Australian Government's Backing Australia's Ability initiative, in part through the Australian Research Council.

8. References

[1] F. J. Heemstra, "Software Cost Estimation," *Information and Software Technology,* vol. 34(10), p. 627, 1992.

[2] M. J. Shepperd and C. Schofield, "Estimating software project effort using analogies," *IEEE Trans. on Software Engineering,* vol. 23(11), pp. 736-743, 1997.

[3] B. W. Boehm, *Software Engineering Economics*. Englewood Cliffs, N.J.: Prentice-Hall, 1981.

[4] B. Kitchenham, S. L. Pfleeger, B. McColl, and S. Eager, "An empirical study of Maintenance and Development Estimation Accuracy," *Journal of Systems and Software,* vol. 64, pp. 57-77, 2002.

[5] B. Londiex, *Cost estimation for Software Development*: Addison-Wesley, Reading, 1987.

[6] J. W. Keung, "Providing Statistical Inferences to Case-based Software Cost Estimation," in *Computer Science and Engineering*. Ph.D Thesis, Sydney: UNSW, 2007.

[7] J. W. Keung, B. Kitchenham, and R. Jeffery, "Analogy-X: Providing Statistical Inferences to analogy-based Software Cost Estimation," *Submitted to IEEE Transactions on Software Engineering,* 2007.

[8] M. J. Shepperd and G. Kadoda, "Using simulation to evaluate prediction techniques," in *Software Metrics Symposium*, London, 2001.

[9] B. Efron and G. Gong, "A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation," *The American Statistician,* vol. 37(1), pp. 36-48, 1983.

[10] M. J. Shepperd, C. Schofield, and B. Kitchenham, "Effort estimation using analogy," in *18th Intl. Conf. on Software Engineering*, Berlin, 1996.

[11] N. Mantel, "The detection of disease clustering and a generalized regression approach," *Cancer Research,* vol. 27(22), pp. 209-220, 1967.

[12] B. F. J. Manly, *Randomization, Bootstrap and Monte Carlo Methods in Biology*: Chapman & Hall, 1997.

[13] B. F. J. Manly, *Multivariate Statistical Methods - A Primer*, 2nd ed.: Chapman & Hall, 1998.

[14] F. H. C. Marriott, "Barnard's Monte Carlo tests: How many simulations?," *Applied Statistics,* vol. 28, 1979.

[15] R. Jeffery, M. Ruhe, and I. Wieczorek, "Using public domain metrics to estimate software development effort," in *Proceedings of Seventh International Software Metrics Symposium*, London, England, 2001.