

Experiences Using Case-Based Reasoning to Predict Software Project Effort

Gada Kadoda, Michelle Cartwright, Liguang Chen, and Martin Shepperd
Empirical Software Engineering Research Group
Department of Computing
Bournemouth University
Talbot Campus
Poole, BH12 5BB, UK
Email: {gkadoda, mcartwri, lchen, mshepperd}@bournemouth.ac.uk

January 27, 2000

Abstract

This paper explores some of the practical issues associated with the use of case-based reasoning (CBR) or estimation by analogy. We note that different research teams have reported widely differing results with this technology. Whilst we accept that underlying characteristics of the datasets being used play a major role we also argue that configuring a CBR system can also have an impact. We examine the impact of the choice of number of analogies when making predictions; we also look at different adaptation strategies. Our analysis is based on a dataset of software projects collected by a Canadian software house. Our results show that choosing analogies is important but adaptation strategy appears to be less so. These findings must be tempered, however, with the finding that it was difficult to show statistical significance for smaller datasets even when the accuracy indicators differed quite substantially. For this reason we urge some degree of caution when comparing competing prediction systems and only modest numbers of cases.

Keywords: Software project, effort prediction, cost estimation, case based reasoning, analogy

1. Background

Over the years a variety of techniques have been proposed to help solve the problem of making accurate, yet early, software project predictions. Many of these techniques involve the use of historical data in order to develop prediction systems. An example is statistical regression. Other approaches involve the use of general models or prediction systems that are parameterised to account for differences between project environments. Examples include COCOMO and proprietary methods such as SLIM.

Unfortunately no prediction technique has proved consistently accurate, even when we relax the accuracy criterion to merely require that a technique generates *useful* predictions. Worse still some techniques have proved consistently unsuccessful. For this reason there has been growing interest in recent years in exploring a variety of machine learning (ML) techniques either as a complement or an alternative to existing techniques. Examples the use of artificial neural nets [24], rule induction [15] and case based reasoning or analogy [21]. Although, on

occasions researchers report impressive results, what is often less visible is the amount of hidden effort to configure a ML system. Typically there is a large search space — since many decisions must be made — and little theory to guide the would-be predictor. Consequently, searching for an effective prediction system frequently degenerates into an exercise of trial and error.

The Empirical Software Engineering Research Group at Bournemouth have been involved in the development of case based reasoning (CBR) techniques and tools to build software effort prediction systems for five years. Over this period we have had some success in producing more accurate results than traditional regression based techniques [2, 21, 22]. Subsequently, other research groups have reported more mixed experiences. We believe there are a variety of reasons why this may be so. First, and foremost, effectiveness of any prediction technique depends upon characteristics of the dataset. For example, regression will do well if the majority of datapoints fall upon some hyperplane. Conversely, CBR may well be favoured when discontinuities exist in any underlying relationship between effort and other independent variables. However, we also believe there are a variety of decisions that must be made when utilising CBR techniques. Such decisions include feature subset selection, the number of analogies to utilise and choice of adaptation strategy. This paper is intended to provide some experimental data to assist with the more effective use of CBR techniques for building prediction systems. It also considers some of the issues involved in making comparisons between competing prediction systems.

The remainder of the paper is structured as follows. The next section describes CBR in more detail and details the various experiences of different research teams. We then consider the design decisions that must be made. Next we outline our experimental method and provide some background on the dataset. This is followed by an analysis of the impact of feature subset selection, the relationship between number of analogies and cases, different analogy selection techniques and finally competing adaptation strategies. We conclude by making some tentative recommendations for other researchers using CBR.

2. Related Work

The idea of using analogies as a basis for estimating software project effort is not new. Boehm [4] suggested the informal use of analogies as a possible technique almost 20 years ago. The idea was reiterated by Cowderoy and Jenkins [7] in 1988 but again with no formal mechanism for selecting analogies. The next development was from Vicinanza *et al.* [17, 19, 23] who suggested that developments from the machine learning community in the form of CBR might be usefully adapted to help make better software project predictions. Case based reasoning has four distinct aspects:

- characterisation of cases
- storage of past cases

- retrieval of similar cases to use as analogies
- utilising the retrieved case to solve the target case problem, sometimes known as case adaptation

In the situation of effort prediction CBR might be deployed as follows. We have n projects or cases, each of which needs to be characterised in terms of a set of p features. In addition, we must also know the feature that is to be predicted. Features can either be continuous (e.g. experience of the project manager), discrete (e.g. the number of interfaces) or categorical (e.g. development environment). In practice, many approaches treat discrete features as if they were continuous. Historical project data is collected and added to the case base. When a prediction is required for a new project this case is referred to as the target case. The target case is also characterised in terms of the p features. Incidentally this imposes a constraint on the feature set in that it should only contain features for which the values will be known at prediction time. The next step is to measure similarity between the target case and other cases in the p -dimensional feature space. The most similar cases or projects are then used, possibly with adaptation to generate a prediction for the target case. Once the target case has completed it can be added to the case base. A more general account of CBR may be found in Leake [14].

The approach of Vicinanza *et al.* is to use domain specific similarity measures (mainly based on size e.g. LOC and function points) coupled with rule based adaptation. The authors extract the rules by analysing a protocol of an expert estimating project effort for hypothetical projects. This has been implemented in tool known as cESTOR. They found their technique outperforms COCOMO and FPs for Kemerer's [11] dataset augmented with an additional 7 projects. One disadvantage of this approach is that it is very specific to the particular dataset since the rules and similarity measures are defined in terms of features available. It is unclear how this approach could be easily generalised.

At about the same time Bisio and Malabocchia [3] also experimented with CBR technology and again used the COCOMO dataset to evaluate their approach. They report accuracy levels in the region of MMRE = 40 to 50%, however, the system was only able to make predictions for 46 out of 63 projects.

The approach we adopted contrasts somewhat with that of Vicinanza *et al.* in that we were seeking to develop a more general means of building prediction systems. Our belief is that collecting historical data is sufficiently challenging as it stands without the additional requirement of having predetermined feature sets. We prefer to allow estimators the freedom to utilise those features that they believe best characterise their projects and are most appropriate to their environments. Consequently, we use Euclidean distance in p -dimensional feature space as a means of measuring similarity between cases. Note that categorical features are treated as either identical or completely dissimilar. Likewise, we adopt a simple analogy adaptation strategy of the mean of the k nearest neighbours, where k is an integer such that $0 < k \leq n$. When $k=1$ then the technique is a simple nearest neighbour method. As k tends towards n so the

prediction approach tends towards merely using the sample mean. The choice of k is determined by the estimator.

Our analogy based approach to project estimation is implemented in a software tool known as ANGEL. Using ANGEL we were able to compare the levels of accuracy we obtained with a straightforward stepwise regression (SWR) procedure. We found that for all nine (independent) datasets we studied, ANGEL generated better results than SWR with the exception of one dataset where results were equally good. Further details may be found in Shepperd and Schofield [21]. Subsequently a number of other research groups endeavoured to replicate these findings but with rather mixed results. Niessink and van Vliet [18] also reported that CBR outperformed SWR. Likewise Finnie *et al.* [10] found that CBR outperformed regression analysis (MMRE=36.2% compared with MMRE=62.3%). By contrast, Briand *et al.* [5] and Stensrud and Myrtveit reported the reverse, namely that regression based analysis generated more accurate models than using CBR. Why should this be the case? As we have already stated there is likely to be a strong interaction between the accuracy of a given prediction system and underlying characteristics of the dataset it is applied to. Briand *et al.* report very high adjusted R-squared values for their regression based prediction system. The obvious conclusion is that majority of datapoints fall close to a hyperplane. Such circumstances will favour regression since it will sensibly interpolate and extrapolate. Conversely, an analogy based approach tends to explore what *existing* datapoint, or clump of datapoints, is most similar to the target case.

However, there is a second issue that arises from the review of the results of the various research groups using CBR. It is that CBR is a more complex technology and involves more design decisions than would at first be apparent. This is the topic of this paper. We explore how to utilise CBR in a more effective manner. The problems that either we, or the other research groups, perceive fall into the following categories:

- feature subset selection
- scaling
- similarity measure
- how many analogies to use (i.e. finding a suitable value for k)
- analogy adaptation

It is well known that searching for feature subsets can enhance the performance of CBR systems [1]. The reason is not hard to determine. If we have an arbitrary set of p features it cannot be guaranteed that all will contribute equally to solving the problem in hand. Unfortunately, the problem cannot necessarily be solved merely by causal argument since some features may act as proxies. Suppose we wish to predict project effort. Further suppose that one of the features is a categorical feature indicating the type of pet, if any, that the project manager keeps. It is possible that this feature is redundant. It is also possible that it acts as a proxy for personality or temperament, which in turn is linked to managerial style. ANGEL overcomes this problem by applying a brute force search of all

possible feature subsets. Unfortunately this is an NP-hard search problem of the form $n2^p - 1$. This means that where p is large as in the case of the Briand dataset this approach is not feasible. Briand *et al.* endeavoured to overcome this problem by using a t-test procedure to identify features that exhibit a significant relationship with effort. We are not convinced of the efficacy of this approach due to potential interaction between features so that a stepwise procedure may not perform well. Another approach which has been deployed successfully for rule induction [8] is to use a search heuristic — in their case a simulated annealing algorithm — to find a good but not necessarily optimal feature subset. This remains the subject of future research.

A related problem area is that of scaling. Again this is known to be a major problem for CBR systems [16]. Scaling is essential since otherwise any analysis is dependent upon the choice of units for non-categorical features. Clearly we do not want LOC to be a thousand times more influential than KLOC! The normal solution, and one that ANGEL adopts, is to standardise all dimensions so that the minimum observed value is assigned one arbitrary value, typically zero, and the maximum observed value is assigned another value, typically one. This has the effect of allocating equal influence to all features¹. This obviously interacts with the optimal feature subset problem. Indeed it has the interesting effect that collinearity can sometimes be beneficial since it can be exploited to allow an underlying dimension to become more influential. Again we have no efficient means of searching for an optimal scaling, and indeed it is an even harder search problem than feature subset selection. One could conceive of it as a generalisation of feature subset selection since allocating a scale of zero distance is equivalent to excluding a feature. ANGEL does allow four different scales to be selected, however, our experiences suggest that it is hard to exploit this facility effectively. The problem is too hard without automated support. Again this is another area for further research.

Choice of a similarity measure is another design decision. Most researchers, including ourselves have elected to use Euclidean distance. By contrast Vicinanza *et al.* use a size based similarity measures which suggests an implicit model of project effort (in other words you look for something of a similar size, or more accurately estimated size, and then use the other features for adaptation purposes). Jeffery and Waterden also use a different similarity measure. The choice of measure is important since it will influence which analogies are found but we do not know which measure is “best”. For a fuller discussion of different measures see [13]. Nevertheless we can see that again different decisions may contribute to the quite conflicting results that have been obtained.

The next design decision is how many analogies to search for. Yet again there is no clear rule and ANGEL allows the estimator to choose. In our published results we used between one and three analogies. Briand *et al.* use a single analogy. Which is “best”? How can we know? This area has not been

¹ Another common strategy is to use standard deviation as the unit, see for example Michie *et al.* [Michie, 1994 #1002].

systematically explored and in particular the interaction that one might expect between k and n .

The final area of decision when using CBR for effort prediction is adaptation of the analogies once these are retrieved. Briand *et al.* simply use the nearest neighbour. In our published results we used the mean of the k analogies which differs from Briand *et al.* when $k > 1$. Vicinanza *et al.* use a more sophisticated approach that is rule based. Unfortunately the rules tend to be specific to a dataset. Another possibility is to incorporate distance information so that within the k analogies the more similar will be more influential. We know little about the impact of these different strategies but again it is plausible that they may contribute in part to differences in performance.

To summarise this section, we note that different research groups have had varying experiences using CBR to predict software project effort. We posit that there are two explanations. The first is that underlying dataset characteristics will be influential in favouring or inhibiting different techniques for building prediction systems. CBR is no exception. Elsewhere we are conducting research into this topic [20]. The second explanation is that design decisions when building a CBR prediction system are also influential upon the results. This paper goes on to explore the impact of feature subset selection, how many analogies to utilise and how to choose them and three different analogy adaptation strategies.

3. Method

The following analysis is based on a project effort dataset collected from a Canadian software house [9]. We chose this dataset since it is publicly available, it is reasonably large (81 cases of which 77 are complete) and combines both continuous and categorical features (8 continuous or discrete and 1 categorical). These factors were important since we required some flexibility for our experimentation into the interaction between different CBR strategies and the dataset characteristics. The only pre-processing was to discard the incomplete cases

In order to explore the impact of the size of n (the number of cases) we randomly selected cases to populate smaller datasets which we named Des17, 37, 57 and 77. Note that Des37 contains Des17 plus 20 additional cases and so forth. This was to try to reduce the impact of outlier cases.

Dataset	Mean	Median	Min	Max	Skewness
Des17	4608.88	3829	805	13860	1.140
Des37	5370.14	3472	546	23940	1.780
Des57	4754.23	3192	546	23940	1.792
Des77	4833.91	3542	546	23940	1.998

Table 1: Summary Statistics for the Dependent Variable (Effort)

Table 1 indicates that apart from the smallest dataset (Des17) each dataset contains the minimum and maximum observation and is positively skewed. The datasets are rather heterogeneous with an almost 20-fold variation in the dependent variable (effort) even for Des17.

In order to generate predictions we adopted a jack-knifing procedure for each dataset, as a means of avoiding very small validation sets. Consequently for each dataset there are n predictions, each based upon the remaining $n-1$ cases. For each prediction there is a corresponding residual, or error, being the difference between actual and predicted.

Dataset	Mean Absolute residual	MMRE
Des17	2681	112.4%
Des37	3869	139.8%
Des57	3304	135.9%
Des77	3008	119.30%

Table 2: Sample Mean Based Prediction Accuracy

In all our experiments we use prediction based upon the sample mean as our benchmark or null hypothesis. In situations where one is not predicting better than can be attained through the sample mean, it is questionable as to whether one is predicting at all! Table 2 indicates the threshold values for the four versions of the Desharnais dataset using two different accuracy indicators, namely mean absolute residual or error and mean magnitude of relative error (MMRE) due to Conte *et al.* [6]. Our preference is for mean absolute error since it is a symmetric² measure, however, it is dataset dependent so we also provide the more conventional MMRE indicator to allow some comparability with other results. We would caution the reader though, that the two indicators can provide contradictory results on occasions. In such circumstances we consider mean absolute residual to be a better guide. A more detailed discussion of different prediction accuracy indicators can be found in [12].

Another problem is determining how much significance to attach to differences in accuracy indicators. Suppose prediction system A yields an MMRE of 50% and prediction system B yields an MMRE of 48%. What, if anything may we conclude from these observations? Our approach here is to test to see if the residuals are samples drawn from the same underlying population. Since we use absolute values the residuals are strongly positively skewed consequently we use non-parametric tests. Where the data is naturally paired, for example comparing $k=1$ and $k=2$ on the same dataset we use the Wilcoxon Signed Rank Test, otherwise the Mann-Whitney U Test. We set the significance level $\alpha=0.05$. Since

² By symmetric we mean an accuracy indicator that doesn't penalise under and over estimates differently.

we are making multiple comparisons it could be argued that we should make the Bonferroni adjustment (i.e. adjusted significance level = α/x where x is the number of comparisons being made). This could be somewhat conservative, nevertheless we return to this in our discussion of results.

4. Results

In this section we organise the results by the following three questions:

- what is the impact of feature subset selection?
- how many analogies should we use and how should they be selected?
- can we use dissimilarity or distance information to form an effective adaptation strategy?

(a) The impact of feature subset selection

The first question has proved very difficult to investigate thoroughly which is unfortunate since we believe to be of considerable significance. Is feature subset selection worthwhile? The difficulty is that we have no efficient means of finding the optimal subset. With a jack knifing procedure this search must be repeated n times. The problem is further compounded by the fact that ANGEL optimises on MMRE which we no longer believe to be a completely trustworthy accuracy indicator. In earlier work we overcame this problem by searching once for the entire dataset and then jack-knifing. This was justified on the grounds that we were making a comparison with SWR where we were only model fitting and there was no hold-out sample. We argued that this was still conservative with regards to CBR. However for the purposes of this investigation, this introduces a bias. Consequently, we are deferring this particular question until we are able to develop more automated analysis tools. All subsequent analysis is based upon the full feature set.

(b) Selecting analogies

This part of our investigation is concerned with the relationship between the number of analogies (k) and the size of the dataset (n). We consider two approaches to determining k . First, it can be set to some constant value. We explore values in the range 1-5. Second, it can be determined dynamically as the number of cases that fall within distance d , of the target case. This has the effect that sometimes no prediction is possible since no analogies fall within the specified distance.

Dataset size (n)	k=1	k=2	k=3	k=4	k=5
17	56.8%	49.8%	52.9%	54.0%	60.9%
37	78.0%	62.8%	72.3%	82.5%	84.7%
57	67.4%	70.1%	61.4%	61.9%	64.4%
77	60.9%	55.3%	52.5%	47.6%	49.3%

Table 3a: MMRE by Dataset Size and Number of Analogies

Dataset size (n)	k=1	k=2	k=3	k=4	k=5
17	2336	1742	1599	1662	1648
37	3013	2592	2809	2800	3054
57	2715	2401	2177	2263	2229
77	2598	2417	2192	1979	1989

Table 3b: Mean Absolute Residual by Dataset Size and Number of Analogies

Tables 3a and 3b summarise the accuracy levels obtained by varying the number of analogies k , and the size of the dataset n .

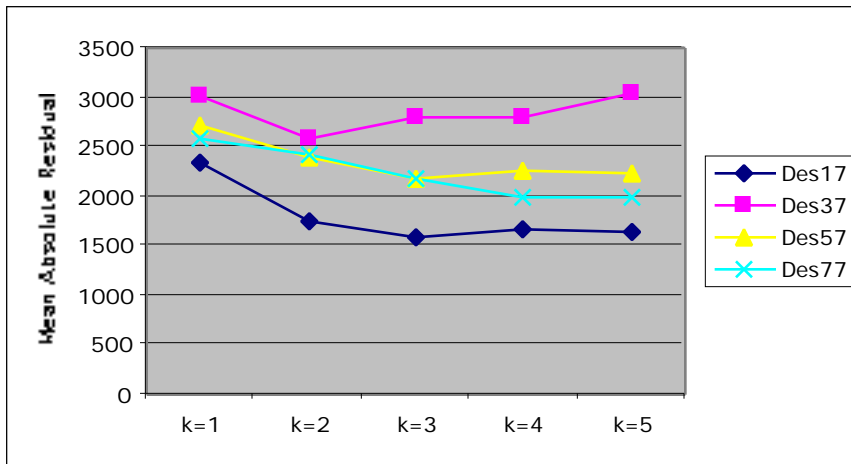


Figure 1a: Lineplots of k vs Mean Absolute Residual

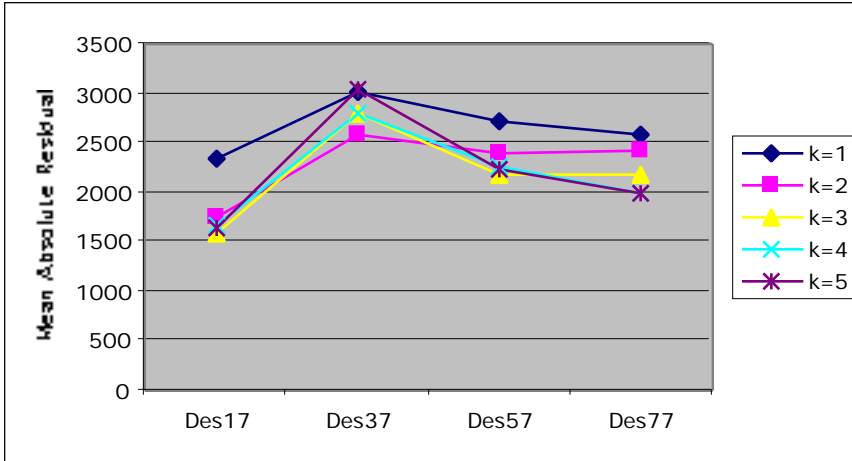


Figure 1b: Lineplots of n vs Mean Absolute Residual

Figures 1a and 1b summarise Table 3b in a graphical form. We note some tendency in Figure 1a for accuracy to improve between $k=1$ and $k=2$ and thereafter to flatten out. Figure 1b suggests that Des17 tends to lead to the most accurate predictions and Des37 to the least accurate. This slightly counter-intuitive finding — that larger case bases do not always lead to better predictions — can be explained in part due to the inclusion of an extreme outlier in Des37 but not in Des17. Des57 and Des77 show gradual improvement from Des37, however, heterogeneity in the dataset would seem to be the dominant effect.

Our analysis of the differences in accuracy has been somewhat subjective. The question remains as to how significant any of these differences are. As we indicated in the previous section we turn to the residuals to answer this question. For the purposes of this analysis we do not care about the direction of an error — an overestimate is deemed to be equivalent to an underestimate — consequently we use absolute residuals. We start with the smallest dataset Des17.

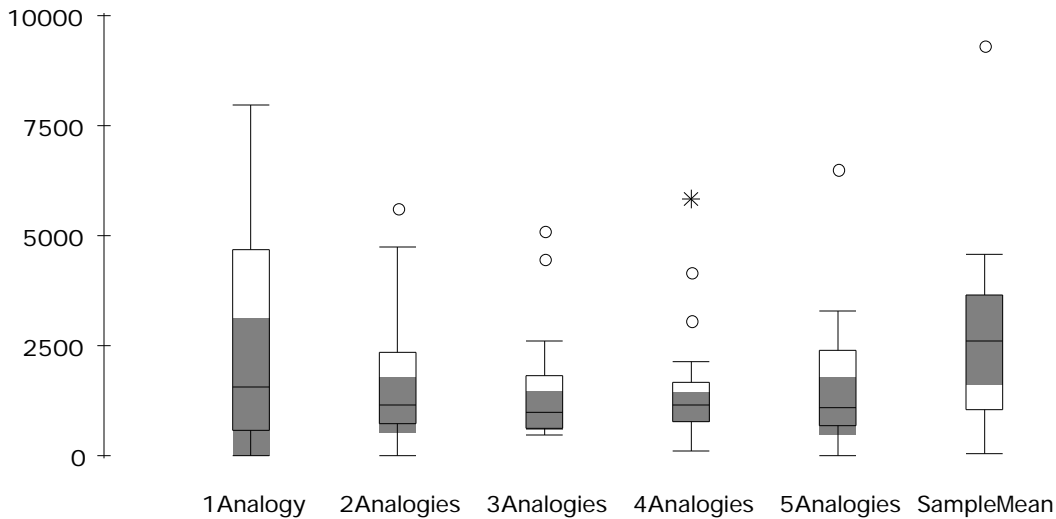


Figure 2: Boxplots of Absolute Residuals for k=1 to 5 for Des17

Figure 2 shows the distribution of residuals or errors using boxplots. We compare using ANGEL with $k=1$ to 5 with the control, i.e. using the sample mean as our prediction for every single project. The shaded area in each boxplot represents the 95% confidence limits for the median which are denoted by the horizontal line across the middle of each box. It can be seen that these overlap and we found no significant differences in medians using the Wilcoxon Signed Rank test. In other words, although there is some evidence that accuracy improves using more analogies we cannot show statistical significance ($\alpha=0.05$). Note that in all cases we generate more accurate predictions than merely using the sample mean, however, we are unable to demonstrate statistical significance. This suggests that a certain degree of caution should be exercised when comparing prediction systems over small datasets even if there appear to be difference in accuracy indicators such as MMRE and sum of the absolute residuals.

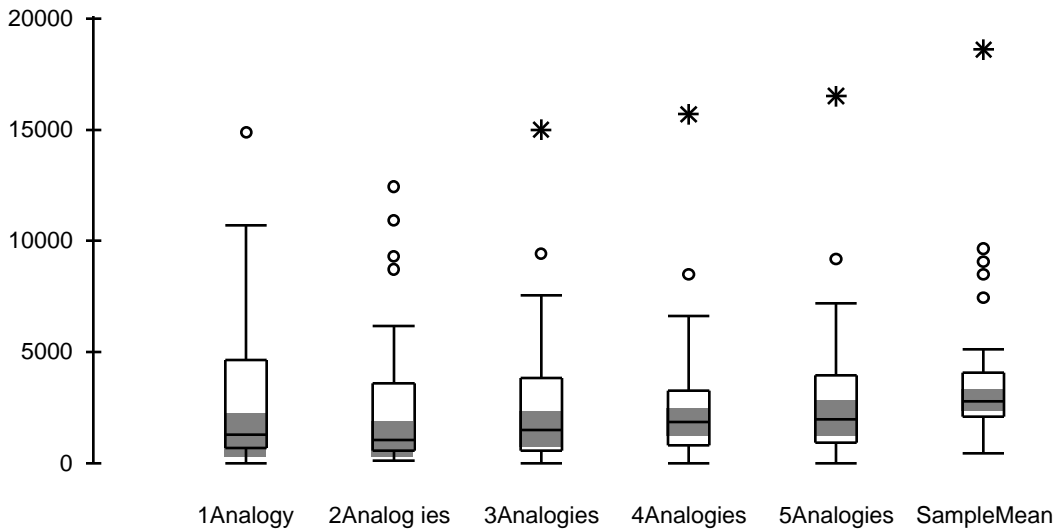


Figure 3: Boxplots of Absolute Residuals for k=1 to 5 for Des37

Figure 3 shows the distribution of absolute errors or residuals for Des37. With 37 cases we were able to show that we do have prediction systems since we can show that the median of the absolute residuals is significantly smaller ($\alpha=0.05$) for ANGEL with $k=1$ to 5, than merely using the sample mean. Unfortunately this analysis is unable to reveal any difference between $k=1$ through to $k=5$ as we were unable to reject the null hypothesis — that there is no difference between the medians of absolute residuals. This is the case even when the MMRE indicators vary quite considerably, e.g. 62.8% ($k=2$) cf 84.7% ($k=5$). Thus we see even with a dataset with 37 cases it can be difficult to make comparisons between prediction systems, other than to assert that we do indeed have predictors! In the past researchers, including ourselves, have attached some significance to a difference in accuracy of such magnitude.

Again, for Des57 we can show that all predictors are significantly better than using the sample mean. Between the predictors we find that $k=3$ is to be preferred to $k=1$.

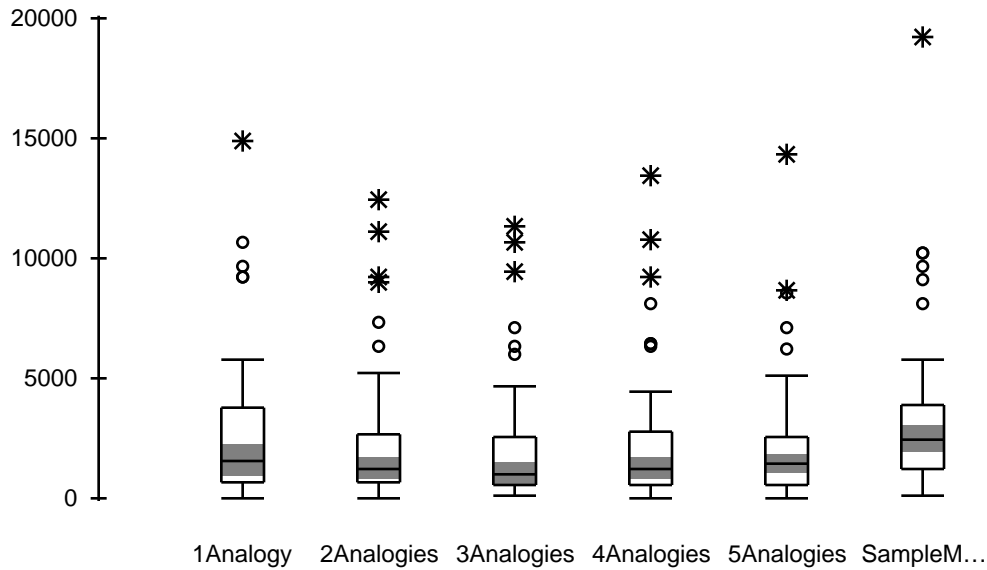


Figure 4: Boxplots of Absolute Residuals for $k=1$ to 5 for Des37

The difficulty in establishing significant differences between the predictors, even for 57 is cases is not hard to understand when observing the distributions of residuals in Figure 4. The shaded areas indicate the 95% confidence limits for the medians and it can be seen these show overlap except with the sample mean predictor.

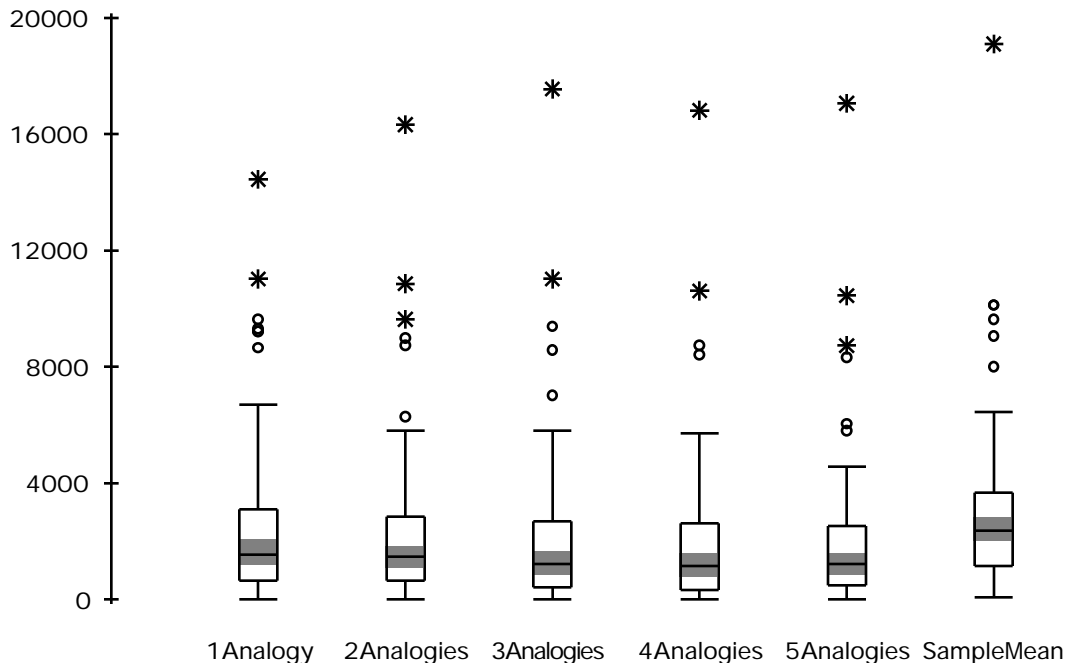


Figure 5: Boxplots of Absolute Residuals for k=1 to 5 for Des77

With a dataset of 77 cases the boxplots in Figure 5 show that even with the worst predictor (using $k=1$) we still have a median error that is significantly lower than using the sample mean approach ($p=0.0126$) using a Wilcoxon Signed Rank test. With a dataset of this size we are at last able to show some differences between the prediction systems. We are able to establish the following order of preference $k=4, k=5 < k=3 < k=1, k=2$ where $x < y$ denotes x has the smaller mean absolute residual value. It would seem that a larger number of analogies is more effective with a larger case base.

Thus far we have determined the value of k independently of the actual distribution of cases in p -space. The next set of results are derived using k analogies where k is the number of cases that fall with distance d of the target case. The distances were determined by studying what seemed sensible steps, bearing in mind too small distance would frequently result in no analogies being chosen, whilst too great distance would lead to something similar to a sample mean. In the end we selected $d=0.1$ to $d=0.3$ in steps of 0.05 where maximum dissimilarity is 1.0.

d	Median	Min	Max
0.10	0	0	3
0.15	2	0	14
0.20	8	0	23
0.25	15	0	37
0.30	21	0	41

Table 4: Summary of Number of Analogies Selected for Des77 Using a Distance-Based Approach

Table 4 gives some indication of the number of analogies being utilised for different values of d using Des77. Obviously the numbers fall substantially as the case base is reduced in size. Even with $d=0.30$ we were unable to find any analogies for two targets in Des77. This could be regarded as a good thing since it suggests that these two projects are profoundly different from other projects in the case base and one would expect a high degree of risk associated with any predictions.

Dataset size	Distance = 0.1	Distance = 0.15	Distance = 0.2	Distance = 0.25	Distance = 0.3
17	n.a. (100%).	n.a. (100%)	89.6% (71%)	71.0% (53%)	46.0% (18%)
37	53.7% (73%)	43.7% (46%)	52.5% (22%)	90.1% (5%)	95.9% (5%)
57	52.6% (68%)	55.9% (32%)	63.4% (9%)	69.5% (5%)	77.6% (4%)
77	54.2% (57%)	38.6% (26%)	56.1% (6%)	63.5% (3%)	66.6% (3%)

Table 5a: Estimation Accuracy by Dataset Size and Distance Based Analogy Selection using MMRE

Dataset size (n)	Distance = 0.1	Distance = 0.15	Distance = 0.2	Distance = 0.25	Distance = 0.3
17	n.a. (100%)	n.a. (100%)	1473 (71%)	1473 (53%)	1723 (18%)
37	1728 (73%)	1670 (46%)	1768 (22%)	2967 (5%)	2800 (5%)
57	2198 (68%)	1927 (32%)	2174 (9%)	2475 (5%)	2509 (4%)
77	2155 (57%)	1663 (26%)	1949 (6%)	2173 (3%)	2128 (3%)

Table 5b: Estimation Accuracy by Dataset Size and Distance Based Analogy Selection using Mean Absolute Residuals

Tables 5a and 5b summarise accuracy levels (using MMRE and mean absolute residuals) achieved using different values for d for the four different sized datasets. The percentages in brackets denote the proportion of estimates that could not be made due to no analogies being sufficiently similar.

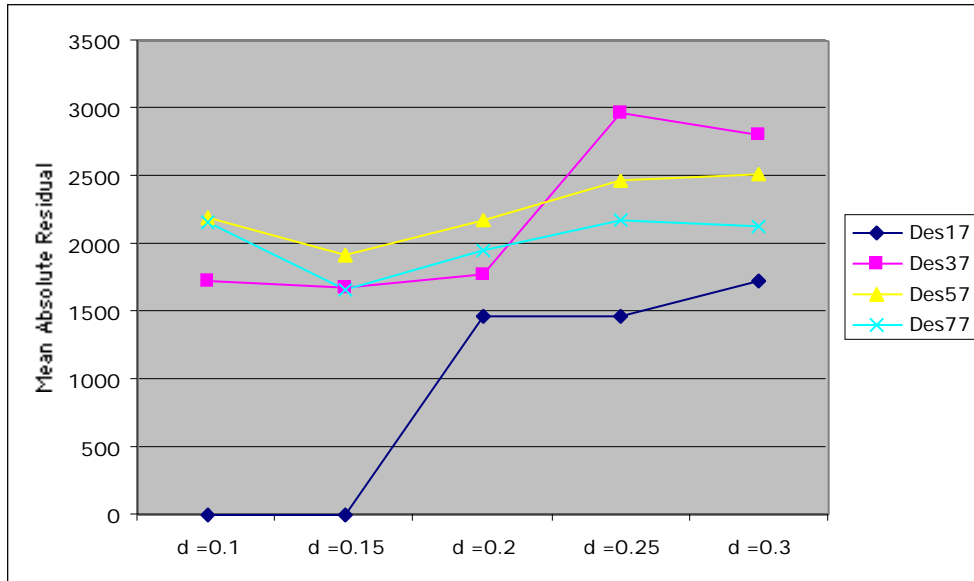


Figure 6a: Lineplots of Accuracy vs Distance d

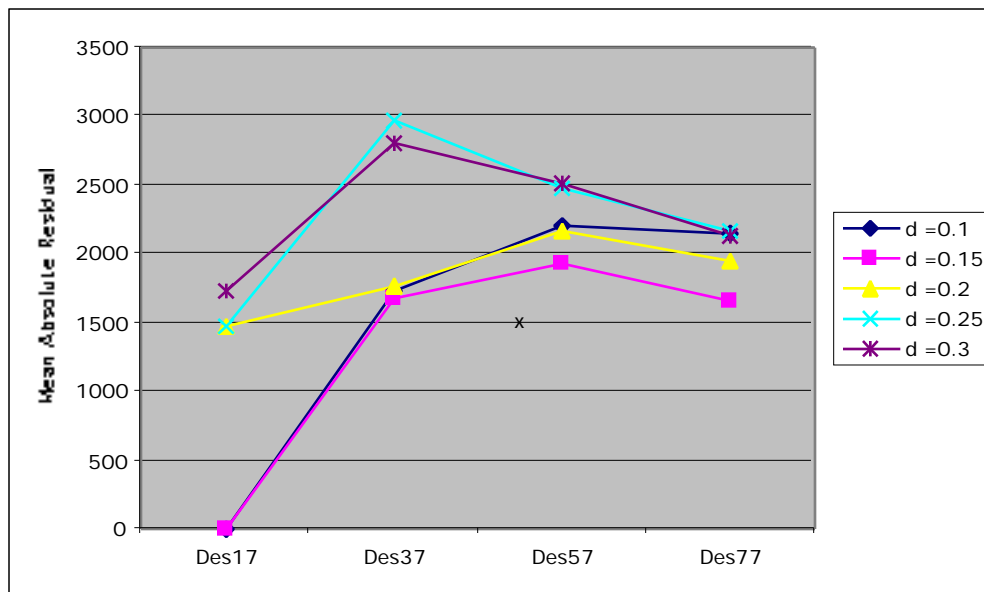


Figure 6b: Lineplots of Accuracy vs Dataset size

Figures 6a and 6b summarise graphically the results from Table 5b. We see in Figure 6a some tendency for accuracy to decrease as the distance d is increased, however, it must be appreciated that greater accuracy is at the expense of fewer predictions being made. This is not an aspect of prediction systems that has really been addressed, although if the costs of not being able to predict and predicting inaccurately are known it would be interesting to search for the optimal prediction system. In Figure 6b we also observe that the least good results are for Des37 — although as we have previously remarked it contains

some extreme outliers — and then improvement as the case base is increased in size. The better results for Des17 are probably an artefact of there being so few predictions possible.

We now turn to testing for the significance of any these differences.

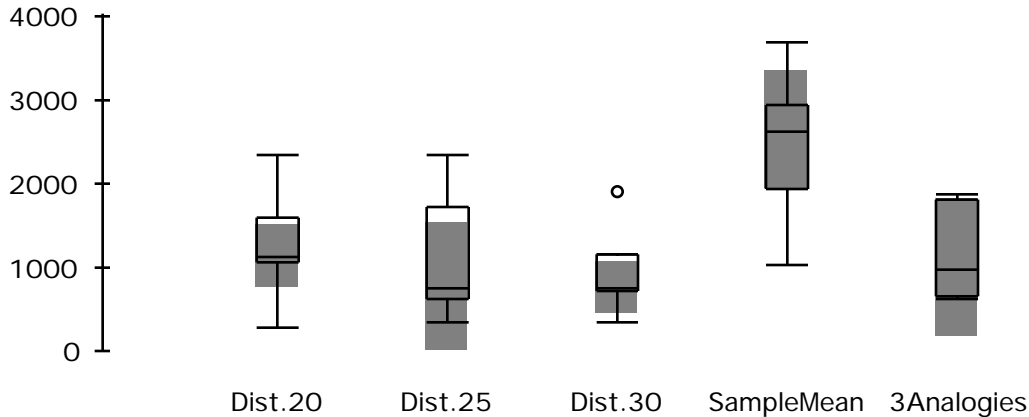


Figure 7: Boxplots of Absolute Residuals using Distance Based Analogy Selection for Des17

Figure 7 shows the distribution of residuals for analogy based prediction using all analogies that fall within a specified distance. Note that distances of 0.1 and 0.15 are excluded since no sufficiently similar analogies could be found within such a small dataset. Again there are no significant differences between the medians, however, we did generally find better performance when comparing this approach with a fixed number of analogies, albeit at the cost of being unable to make a prediction at all for some projects when no analogies fall within the specified distance. Figure 3 also shows $k=3$, which is the best alternative.

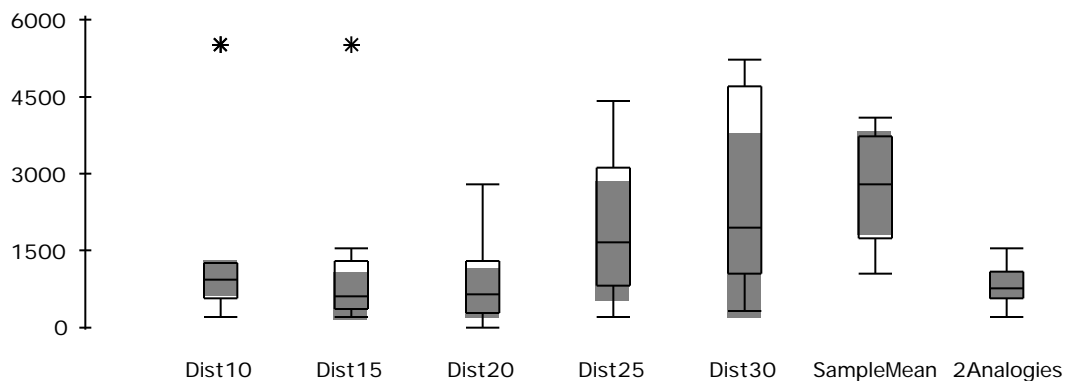


Figure 8: Boxplots of Absolute Residuals using Distance Based Analogy Selection for Des37

For Des37 we are only able to show that setting the distance at $d=0.2$ is more accurate than $d=0.3$ ($p = 0.0384$). Comparing with $k=2$, the best alternative we see

that $d=0.2$ is superior but not significantly so. Also this is at the expense of being unable to predict 8 cases out of 37.

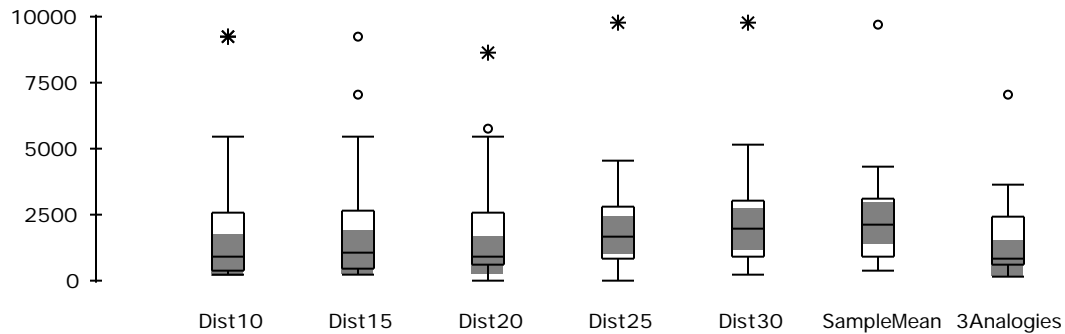


Figure 9: Boxplots of Absolute Residuals using Distance Based Analogy Selection for Des57

For Des57 we can show that all predictors are significantly better than using the sample mean other than for distance = 0.3 where we narrowly fail to reject the null hypothesis ($p=0.0597$). In general $k=3$ is more accurate than the distance based predictors although we can only show significance in relation to $d=0.3$.

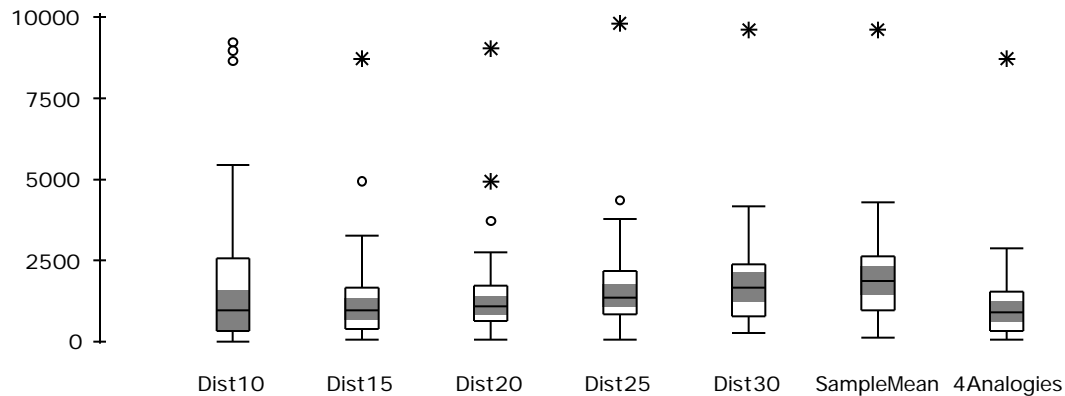


Figure 10: Boxplots of Absolute Residuals using Distance Based Analogy Selection for Des77

With the 77 cases we are able to show the following preferences $d=0.1, d=0.15 < d=0.2 < d=0.25, d=0.3$ at the $\alpha=0.05$ confidence level. However we see $k=4$ as being superior, significantly so compared with $d=0.25$ and marginally so with $d=0.3$ ($p=0.0593$). In addition, all prediction systems were also significantly better than using the sample mean.

To summarise:

- Des37 shows some anomalous behaviour probably due to the inclusion of extreme outliers

- larger case bases do not necessarily lead to more accurate predictions
- increasing the number of analogies k up to 3 generally improves a prediction, but thereafter makes little difference. This effect is more pronounced on the larger case bases.
- a distance based mechanism for selecting analogies tends to be more effective for the smaller case bases and a fixed value for k more effective for the larger case bases. This may be due to the distance values being less suitable for larger case bases and therefore finding too many analogies.

(c) Analogy Adaptation Strategies

This section considers the issue of how to adapt the analogies chosen in order to generate a prediction. Currently, where more than one analogy is selected, ANGEL will calculate the mean of the chosen analogies. This has the effect of treating all analogies as being equally influential on the outcome. Intuitively, we should expect the closest analogies to have more influence. For example, when choosing five analogies, three may be very close, whereas the other two are far less so. It would seem reasonable to weight mean in order to reflect the influence of the analogies chosen according to distance. Another weighting strategy would be to allow the higher ranked analogies to have more influence than lower ones. Both of these simple strategies make use of data produced by ANGEL (distance of an analogy from the target, rank of analogy in relation to the analogies chosen). The five closest analogies were considered. The distance between each of these and the target was recorded. We then calculated the reciprocal distance weighted mean of the five nearest analogies for each of the 77 predictions again using a jack knifing procedure. This analysis is based only on Des77 and again uses all features.

No. of analogies	mean of k analogies	inverse distance weighted mean	inverse rank weighted mean
1	60.9%	n.a.	n.a.
2	55.3%	61.7%	62.7%
3	52.5%	57.5%	59.2%
4	47.6%	52.7%	55.8%
5	49.3%	53.5%	55.3%
Distance =0.2	56.1%		

Table 6a: A Comparison of Different Analogy Adaptation Procedures using MMRE

No. of analogies	mean of k analogies	inverse distance weighted mean	inverse rank weighted mean
1	2598	n.a.	n.a.
2	2417	2592	2613
3	2192	2331	2417
4	1979	2142	2276
5	1989	2118	2232
Distance =0.2	1949		

Table 6b: A Comparison of Different Analogy Adaptation Procedures using mean absolute residual

The results are summarised in Tables 6a and 6b using the MMRE and mean absolute residual accuracy indicators.

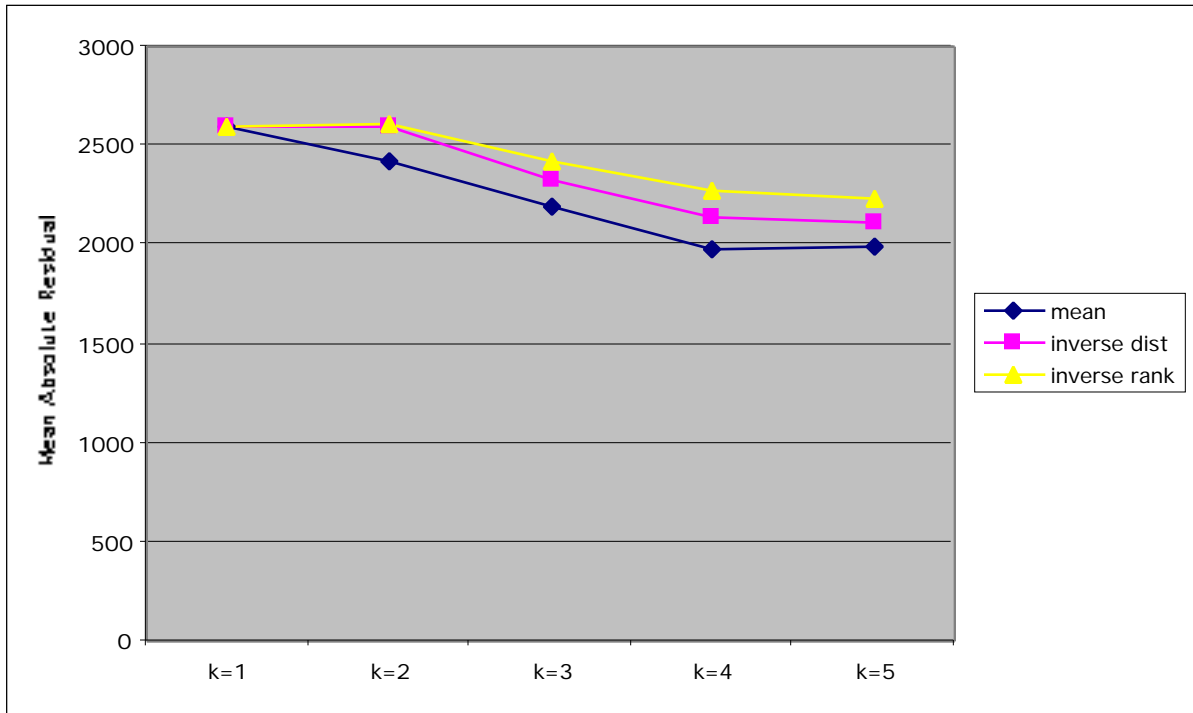


Figure 11: Predictive Accuracy of Three Analogy Adaptation Techniques

Figure 11 summarises graphically the data in Table 6b and shows that somewhat surprisingly that a straightforward mean is the most accurate technique. In other words, using distance information to modify influence does not contribute to the accuracy of a prediction. Note that for $k=1$ all three techniques are identical.

When testing for differences between the residuals using the Wilcoxon Signed Ranks test no significant differences were found leading us to conclude for

Des77, at least, the more sophisticated analogy adaptation strategies offer little value.

5. Discussion

In this paper we have looked at some of the reported experiences of using CBR technology for software project effort prediction. In particular, we have considered the question of why there have been such differing results. We have argued that these arise in part due to different characteristics of the datasets being examined. We also argue — and this is the major theme of our experimentation — that configuring a CBR prediction system is a non-trivial task. We identify a number of design decisions that include feature subset selection, scaling, selection and adaptation of analogies. Unfortunately the subset selection and scaling are computationally demanding and thus beyond the scope of this paper. We focus, however, upon different mechanisms for selecting analogies and different mechanisms for utilising the analogies once retrieved. We also considered the interaction between these factors and the size of the case base.

Our analysis suggests that decisions on how to configure the CBR system, in our case ANGEL, can have a substantial impact upon the level of accuracy obtained. For example, with the Des77 dataset we obtained accuracy levels that ranged from MMRE=38% to MMRE=66%. Such variations may in part explain why different research teams have generated conflicting results. Unfortunately there is little published data on effective strategies for configuring CBR systems. The result is a situation close to trial and error. We have tried to remedy this situation somewhat, however, in doing so have encountered another problem. Our tentative findings are that the presence of extreme outliers can have a major impact upon accuracy. Consequently increasing the size of the case base does not necessarily enhance accuracy. Three analogies seemed to be optimal although a fixed value for k was more effective for the larger case bases. Distance based case selection appeared more effective for the smaller case bases. Surprisingly we found case adaptation has little impact although we only examined this for Des77.

In some situations there have been relatively small differences in accuracy levels reported by either indicator (MMRE or mean absolute residual). In order to determine the significance of such differences we have formally tested the difference between sets of residuals where the null hypothesis is that they are drawn from the same underlying population. We found that with the smaller datasets we have insufficient datapoints even to show that the prediction system outperforms using the sample mean. Revealing small differences between competing prediction systems is even more difficult. In our case, we have few significant results other than with the Des77 dataset where through a jack knifing procedure we have a validation set of 77 predictions. This is quite disturbing since many datasets are considerably smaller than the Desharnais dataset that we used. One way forward could be to make greater use of

simulated datasets so that a large validation — though not necessarily training — set can be made available.

Acknowledgements

This research was partly funded by grant (Grant GR/L37298) from the UK Engineering and Physical Sciences Research Council and the Defence Evaluation Research Agency. The authors are also grateful to Jean-Marc Desharnais for making his dataset available.

References

- [1] Aarmodt, A. and E. Plaza, 'Case-based reasoning: foundational issues, methodical variations and system approaches', *AI Communications* , 7(1), 1994.
- [2] Atkinson, K. and M.J. Shepperd. 'The use of function points to find cost analogies', in *Proc. European Software Cost Modelling Meeting* . Ivrea, Italy: 1994.
- [3] Bisio, R. and F. Malabocchia. 'Cost estimation of software projects through case base reasoning', in *Proc. 1st Intl. Conf. on Case-Based Reasoning Research & Development* . Springer-Verlag, 1995.
- [4] Boehm, B.W., *Software Engineering Economics* . Prentice-Hall: Englewood Cliffs, N.J., 1981.
- [5] Briand, L., T. Langley, and I. Wieczorek. 'A replicated assessment and comparison of common software cost modeling techniques', in *Proc. ICSE (under review)* . 1999.
- [6] Conte, S., H. Dunsmore, and V.Y. Shen, *Software Engineering Metrics and Models* . Benjamin Cummings: Menlo Park, CA, 1986.
- [7] Cowderoy, A.J.C. and J.O. Jenkins. 'Cost estimation by analogy as a good management practice', in *Proc. Software Engineering 88* . Liverpool: IEE/BCS, 1988.
- [8] Debuse, J.C.W. and V.J. Rayward-Smith, 'Feature subset selection within a simulated annealing data mining algorithm', *J. of Intelligent Information Systems* , 9, pp57-81, 1997.
- [9] Desharnais, J.M., *Analyse statistique de la productivite des projets informatique a partie de la technique des point des fonction* . 1989, University of Montreal:
- [10] Finnie, G.R., G.E. Wittig, and J.-M. Desharnais, 'A comparison of software effort estimation techniques using function points with neural networks, case based reasoning and regression models', *J. of Syst. Softw.* , 39, pp281-289, 1997.

- [11] Kemerer, C.F. , 'An empirical validation of software cost estimation models', *CACM* , 30 (5), pp416-429 , 1987 .
- [12] Kitchenham, B.A., et al. , 'Assessing Prediction Systems', *IEEE Transactions on Software Engineering (submitted)* , 1999.
- [13] Kolodner, J.L., *Case-Based Reasoning* . Morgan-Kaufmann: 1993.
- [14] Leake, D., *CBR in context: the present and the future*, in *Case based reasoning: experiences, lessons and future directions*, D. Leake, Editor, AAAI Press: Menlo Park, 1996.
- [15] Mair, C., et al. , 'An investigation of machine learning based prediction systems', *J. of Syst. Softw.* , (accepted for publication), 1999.
- [16] Michie, D., D.J. Spiegelhalter, and C.C. Taylor, ed. *Machine learning, neural and statistical classification* . Ellis Horwood Series in Artificial Intelligence, ed. J. Campbell. Ellis Horwood: Chichester, Sussex, UK, 1994.
- [17] Mukhopadhyay, T., S.S. Vicinanza, and M.J. Prietula, 'Examining the feasibility of a case-based reasoning model for software effort estimation', *MIS Quarterly* , 16(June), pp155-71, 1992.
- [18] Niessink, F. and H. van Vliet. 'Predicting maintenance effort with function points', in *Proc. Intl. Conf. on Softw. Maint.* Bari, Italy: IEEE Computer Society, 1997.
- [19] Prietula, M.J., S.S. Vincinanza, and T. Mukhopadhyay, 'Software Effort Estimation With a Case-Based Reasoner', *J. Experimental & Theoretical Artificial Intelligence* , 8, pp341 - 363, 1996.
- [20] Shepperd, M.J. and G. Kadoda, *Comparing Techniques to Build Prediction System using Simulated Data*. ESERG Technical Report No. 00/04, Bournemouth University, ESERG, 2000.
- [21] Shepperd, M.J. and C. Schofield, 'Estimating software project effort using analogies', *IEEE Trans. on Softw. Eng.* , 23(11), pp736-743, 1997.
- [22] Shepperd, M.J., C. Schofield, and B.A. Kitchenham. 'Effort estimation using analogy', in *Proc. 18th Intl. Conf. on Softw. Eng.* Berlin: IEEE Computer Press, 1996.
- [23] Vicinanza, S., M.J. Prietula, and T. Mukhopadhyay. 'Case-based reasoning in effort estimation', in *Proc. 11th Intl. Conf. on Info. Syst.* 1990.
- [24] Wittig, G. and G. Finnie, 'Estimating software development effort with connectionists models', *Information & Softw. Technol.* , 39(7), pp469-476, 1997.