Impact Analysis of Missing Values on the Prediction Accuracy of Analogy-based Software Effort Estimation Method AQUA

Jingzhou Li
Software Engineering Decision Support Laboratory, University of Calgary, Calgary AB, Canada, T2N 1N4
jingli@ucalgary.ca

Ahmed Al-Emran
Software Engineering Decision Support Laboratory, University of Calgary, Calgary AB, Canada, T2N 1N4
aalemran@ucalgary.ca

Guenther Ruhe
Software Engineering Decision Support Laboratory, University of Calgary, Calgary AB, Canada, T2N 1N4
ruhe@ucalgary.ca

Abstract

Effort estimation by analogy (EBA) is often confronted with missing values. Our former analogy-based method AUQA is able to tolerate missing values in the data set, but it is unclear how the percentage of missing values impacts the prediction accuracy and if there is an upper bound for how big this percentage might become in order to guarantee the applicability of AQUA. This paper investigates these questions through an impact analysis.

The impact analysis is conducted for seven data sets being of different size and having different initial percentages of missing values. The major results are that (i) we confirm the intuition that the more missing values, the poorer the prediction accuracy of AQUA; (ii) there is a quadratic dependency between the prediction accuracy and the percentage of missing values; and (iii) the upper limit of missing values for the applicability of AQUA is determined as 40%.

These results are obtained in the context of AQUA. Further analysis is necessary for other ways of applying EBA, such as using different similarity measures or analogy adaptation methods from those used in AQUA. For that purpose, the experimental design in this study can be adapted.

1. Introduction

Software effort estimation by analogy (EBA) [1, 2, 3] is a data-driven estimation method. It compares the project under consideration (target project) with similar projects in a historical data set through their common attributes, and determines the effort of the target project as a function of the known efforts from the most similar historical projects. EBA can be used for effort estimation for objects at levels of project, feature, or requirement, given corresponding historical data sets. There are three basic steps for EBA for a given object under estimation:

Step 1: Retrieve analogs (or similar objects) of the given object from the historical data set through a set of common attributes using certain similarity measures.
Step 2: Determine the closest analogs of the given object.
Step 3: Predict the effort of the given object by adapting the effort information of the closest analogs, which is referred to as analogy adaptation.

In practical application, EBA is often confronted by the fact of missing values. For traditional EBA methods [1, 2, 4] that use distance-based similarity measures (e.g. Euclidian distance), the data sets are not allowed to contain missing values. Techniques that can help get rid of the missing values, such as deletion and imputation [5, 6, 7], are necessary for the applicability of these EBA methods. In the collaborative filtering based EBA method [8], similarity calculation just considers the available values. A special value *NULL* is defined to represent missing value in [3] together with a set of operations such that the missing values can be tolerated in the similarity calculation.

Although techniques dealing with missing values can be applied, the impact of the amount of missing values in a data set on the eventual prediction accuracy is unknown in detail. Intuitively, the prediction accuracy of an EBA method is expected to decrease when the percentage of missing values increases. However, we need empirical evidence to support the assertion. Further questions are related to the type of dependency and a possible upper bound for the percentage of missing values such that the prediction is still possible in principle.

This paper is dedicated to the empirical analysis of the impact of missing values on the prediction

0-7695-2886-4/07 $20.00 © 2007 IEEE
DOI 10.1109/ESEM.2007.10

accuracy of an EBA method called AQUA and proposed in [3]. Section 2 briefly introduces the AQUA method and the way of dealing with missing values. The research questions and the related hypotheses are formulated in section 3. Section 4 presents the experimental design that covers the data sets, evaluation criteria of prediction accuracy, the process of randomly introducing missing values as well as the experiment process. The major results are presented and discussed in section 5. Threats to internal and external validity of this study are discussed in section 6. Finally, we present the conclusion out of this paper and our future work in section 7.

2. EBA method AQUA in a nutshell

2.1 AQUA overview

AQUA is an EBA method that uses historical data to predict the effort for a new object that may be requirement, feature, or project. In AQUA, the historical data set *DB* is defined as a triple $DB = <R, P, V>$. We use the following notation:

R is the set of objects $R = \{r_1, r_2, ..., r_n\}$,

$P = A \cup \{Effort\}$,

where $A = \{a_1, a_2, \ldots, a_m\}$ is the set of attributes to describe the objects, *Effort* is a specific attribute characterizing the effort necessary to realize the respective object;

$Effort(r_i)$ represents the effort to develop object r_i,

$V = \{a_j(r_k)\}$ is the domain of attribute values of all objects in R,

$a_j(r_k)$ represents the value of attribute $a_j \in P$ for object $r_k \in R$,

$S = \{s_1, s_2, \ldots, s_t\}$ denotes the set of objects to be estimated. S shares the same attributes A with R.

The problem of effort estimation by analogy is then stated as:

Problem-EBA: For all $s_g \in S$, the effort of s_g, $Effort(s_g)$, is to be estimated based on the values of *Effort* from a set of most similar objects $r_i \in R$ to s_g that are retrieved through a common set of attributes using certain similarity measures.

The set of most similar objects is called Top-N analogs of s_g, denoted by $R_{topN}(s_g)$. The estimation using effort information from the analogs is also known as analogy adaptation.

In order to get the Top-N similar objects from *DB*, s_g is compared with all the objects in R through certain similarity measures. Similarity measures between two objects over a set of attributes are defined in terms of local and global similarity measures [9]. Local similarity measure *Lsim* is defined as measuring the similarity between two objects related to an individual attribute $a_j \in A$:

$Lsim: M_j \times M_j \rightarrow [0, 1] \cup \{NULL\}$

where M_j is the type of attribute a_j.

The types of attributes that supported by AQUA are defined in [3].

The global similarity measure between $s_g \in S$ and $r_i \in R$ is defined as a function of local similarity measures:

$Gsim: S \times R \rightarrow [0, 1] \cup \{NULL\}$, and

$Gsim(s_g, r_i) = f(Lsim(a_1(s_g,), a_1(r_i)), Lsim(a_2(s_g,), a_2(r_i)), \ldots, Lsim(a_m(s_g,), a_m(r_i)))$.

While function f can be defined in many forms, *weighted mean* is used in AQUA:

$$Gsim(s_g, r_i) = \sum_{k=1}^{m} (w_k * Lsim(a_k(s_g), a_k(r_i)) \quad (1)$$

where w_i is the normalized weight of attribute a_i and $\sum_{k=1}^{m} w_k = 1$.

Given the closest N analogs of s_g, $R_{topN}(s_g)$, the predicted value for the effort of s_g, represented by $\widehat{Effort}(s_g)$, is then adapted as:

$$\widehat{Effort}(s_g) = \frac{\sum_{r_k \in R_{topN}(s_g)} (Effort(r_k) * Gsim(s_g, r_k))}{\sum_{r_k \in R_{topN}(s_g)} Gsim(s_g, r_k)} \quad (2)$$

2.2 Dealing with missing values in AQUA

In order to tolerate missing values in *DB*, a special value *NULL* is used to replace a missing value in AQUA such that the local and global similarity measures can be calculated in the presence of *NULL* values. The following operations on *NULL* are defined for local similarity measure *Lsim*, global similarity measure *Gsim*, arithmetic addition and multiplication:

(P_1) $Lsim(b, NULL) = NULL$
(P_2) $Lsim(NULL, b) = NULL$
(P_3) $Lsim(NULL, NULL) = NULL$
(P_4) $w*NULL = NULL*w = NULL$
(P_5) $w + NULL = NULL + w = w$
(P_6) $NULL + NULL = NULL$

where b is the value of a valid type of attribute defined in AQUA, $w \in [0, 1]$; operations "*" and "+" are arithmetic multiplication and addition respectively. Operations P_1, P_2, and P_3 define that *Lsim* is *NULL* if and only if either or both of the two participating attributes have *NULL* values.

Operations P_4, P_5, and P_6 define that *Gsim*, as defined in equation (1), is *NULL* if and only if either or both of the participating *Lsim* are *NULL*. $\widetilde{Effort}(s_g)$, as defined in equation (2), is set to *NULL* if and only if $R_{topN}(s_g)=\emptyset$, which means that s_g does not have any similar objects in the data set *DB*.

It can be seen from the above discussion that the effect of the *NULL* is to ignore the participating attributes that have missing values in searching similar objects. Therefore, the more *NULL* in the data set, the less number of attributes participating in searching analogs through similarity measures. Now we are interested to know how the prediction accuracy of AQUA is affected by the amount of missing values in the data set and what percentage of missing values in a data set is acceptable in terms of the applicability of AQUA.

We define that the percentage of missing values in data set *DB* is calculated as *MisVal*%:

$$MisVal\% = \#Missing\_Values/(nm) \quad (3.1)$$

$$MisVal = MisVal\%*100 \quad (3.2)$$

where n is the number of objects in R and m is the number of attributes in A; *#Missing_Values* is the number of *NULL* values in *DB*.

3. Research objectives

3.1 Research questions

The main goal of this research is to investigate the impact of missing value on the accuracy of prediction results for EBA method AQUA. In more detail, we formulate three research questions and the related hypotheses. The research questions are formulated using the goal-oriented template [10]:

Research Question 1	
Analyze (objects of interest)	Seven data sets of varying size and different percentage of missing values
In order to (purpose)	understand the impact of missing values
With respect to (focus)	Accuracy of prediction
From the point of view of (perspective)	Project manager
For the environment (context)	EBA method AQUA

Research Question 2	
Analyze (objects of interest)	Seven data sets of varying size and different percentage of missing values
In order to (purpose)	Understand the impact of missing values
With respect to (focus)	Type of the dependency between the percentage of missing values and the accuracy of prediction
From the point of view of (perspective)	Project manager
For the environment (context)	EBA method AQUA

Research Question 3	
Analyze (objects of interest)	Seven data sets of varying size and different percentage of missing value
In order to (purpose)	Understand the impact of missing values
With respect to (focus)	Applicability of the EBA method AQUA in dependence of the percentage of missing values.
From the point of view of (perspective)	Project manager
For the environment (context)	EBA method AQUA

3.2 Research hypotheses

We formulate the research hypotheses in correspondence to the three research questions stated above. Each hypothesis is presented as a pair of *alternative hypothesis* and *null hypothesis*. The null hypothesis is directly tested; while the alternative hypothesis asserts the opposite of the null hypothesis. The alternative hypothesis is supported if the null hypothesis is refuted [11].

The hypotheses corresponding to the three research questions are referred to as H1, H2, and H3, respectively. Their corresponding null hypotheses are labeled as H10, H20, and H30. The alternative hypotheses are denoted by H11, H21, and H31, respectively.

H11	The prediction accuracy of AQUA decreases when the percentage of missing values in a data set increases.
H10	The prediction accuracy of AQUA does not decrease when the percentage of missing values in a data set increases.

H21	The dependency of the prediction accuracy of AQUA on the percentage of missing values follows a form of function approximately.
H20	The dependency of the prediction accuracy of AQUA on the percentage of missing values does not follow any form of functions.

If hypothesis H1 is supported, we would like to further test hypothesis H3.

H31	In terms of the percentage of missing values, there is an upper limit for the applicability of AQUA.
H30	There is not an upper limit of the percentage of missing values for the applicability of AQUA.

4. Experimental design

4.1 Data sets

Seven data sets were used for this study. Table 1 gives the summary of these data sets, where "#Objects" represents the number of objects in the data set, "#Attributes" represents the number of attributes excluding attribute Effort, "%Missing values" represents the percentage of missing values, and "%Non-Quantitative attributes" represents the percentage of non-quantitative attributes. The unit of attribute Effort is different in different data sets. Since we use relative error, other than absolute error, to measure accuracy in this analysis, the units do not affect the final analysis results.

Table 1. Summary of the data sets for analysis

Name	#Objects	#Attri-butes	%Missing values	%Non-quantitative attributes	Source
USP05-FT	121	14	2.54	71	[3]
USP05-RQ	76	14	6.8	71	[3]
ISBSG04-1	285	24	27.75	63	[12]
ISBSG04-2	158	24	27.24	63	[12]
Mends03	34	6	0	0	[4]
Kem87	15	5	0	40	[13]
Desh89	81	10	0.006	20	[14, 15]

Because ISGSG04 has 2024 projects, it is difficult to run multiple cross validation over the huge data set. We divided ISBSG04 into smaller subsets according to the range of effort by balancing both the number of values and the range of values in the subsets. Due to space and time, we selected two subsets as representatives of small/medium projects with effort between [500, 1000] (ISBSG04-1) and large projects with effort between [10,000, 20,000] (ISBSG04-2), respectively. Other subsets will be tested in future.

4.2 Measurement of prediction accuracy

For our purposes, Leave-One-Out Cross-Validation (LOOCV) [16] is applied on the historical data set R in all cases. With LOOCV, one object, say s_g, is estimated using others as analogs each time, until all the objects in R are estimated in the same way. After a LOOCV process, each object in R has (n-1) analogs and corresponding global similarity measures, from which the Top-N analogs, $R_{topN}(s_g)$, are chosen to generate $\widetilde{Effort}(s_g)$.

MRE (Magnitude of Relative Error) measures the prediction accuracy for each object s_g; *MMRE* (Mean MRE) and *Pred*(l) (prediction accuracy at level l) [17] are normally used to measure the prediction accuracy for one iteration of the LOOCV process.

Different from other analogy-based effort estimation methods that use a fixed number of analogs for analogy adaptation, thresholds of both global similarity measures (T) and the number of analogs (N) used for analogy adaptation are considered in AQUA when determining $R_{topN}(s_g)$. That means the N analogs in $R_{topN}(s_g)$ must satisfy

$Gsim(r_i, s_g) \geq T^*$, i=1..N.

For a given pair of values (N^*, T^*) of N and T, the corresponding prediction accuracy $Accuracy^*$ of s_g is obtained from a single run of LOOCV. (N^*, T^*, $Accuracy^*$) is called a Point-wise Accuracy (*PAC*) in the three dimensional space (N, T, *Accuracy*) when conducting LOOCV by varying N and T in certain ranges.

The *Accuracy* of a *PAC* is a vector of criteria: *MMRE*, *Pred*, *Strength*, and *MPSW*, which are used to measure the prediction accuracy from different perspectives and will be defined briefly in what follows. For detailed definition and discussion about these criteria, readers are directed to [3] and [18].

For a given data set *DB* as defined in section 2.1, multiple iterations of LOOCV are applied to R with given ranges and steps for varying both N and T. All *PAC* produced from iterations of LOOCV compose the accuracy distribution data base, denoted by *AccuDistr*(*DB*).

In what follows we provide some key definitions used in the analysis of prediction accuracy.

Definition 1. *MRE*(r_k, N, T) — Magnitude of Relative Error [17]

$$MRE(r_k, N, T) = \frac{\left| Effort(r_k) - \widetilde{Effort}(r_k) \right|}{Effort(r_k)} \quad (4)$$

for a given object $r_k \in R$ under estimation, where $Effort(r_k)$ is the actual effort and $\widetilde{Effort}(r_k)$ is the predicted effort of object r_k. □

Definition 2. *MMRE(N, T)*—Mean Magnitude of Relative Error [17]

$$MMRE(N, T) = \frac{1}{n} \sum_{r_k \in R} MRE(r_k) \quad (5)$$

for a given pair of values of N, T for all the n objects in R in a single run of LOOCV. □

Definition 3. *Pred*(α, N, T)—prediction at level l [17]

$$Pred(\alpha, N, T) = \tau / \lambda \quad (6)$$

where λ is the total number of objects that are estimated in a single run of LOOCV with a given pair of values of N, T; and τ is the number of objects with *MRE* less than or equal to l. □

l=0.25 is normally used for actual evaluation in literature and is used in this paper; and *Pred*(0.25) represents *Pred*(0.25, N, T) when N and T are given or not considered.

Definition 4. *Strength*(N, T) [3]

Support(N, T) is the number of objects in R that can be estimated with a given pair of values of N, T.

Strength(N, T) is then defined as the ratio of *Support* to the total number of objects in R. □

A single criterion to measure the overall accuracy of a *PAC*, called *MPSW*, is defined by considering *MMRE*, *Pred*, and *Strength* that measure different aspects of the prediction accuracy.

Definition 5. *MPSW*(N, T) [18]

$$MPSW(N,T) = \eta_1(1 - MMRE\_N(N,T)) + \eta_2 Pred(0.25, N, T) + \eta_3 Strength(N, T) \quad (7)$$

Because *MMRE* may be greater than 1, the normalized *MMRE*, *MMRE_N*, is used. The normalization of *MMRE* is based on all the *PAC* under consideration. Factors η_i are normalized to 1 as well. □

Typically, *MMRE* is the most frequently-used criterion for measuring the prediction accuracy. Therefore, a stronger weight η_1 = 0.4 was given to *MMRE*, while η_2 = 0.3 for *Pred*(0.25) and η_3 = 0.3 for *Strength*.

To compare the overall accuracy across multiple data sets, the average of the *MPSW* of all the involved *PAC* in *AccuDistr*(*DB*), called *MPSV*, is used.

Definition 6. *MPSV*(DB) [18]

$$MPSV(DB) = \frac{1}{p} \sum_{i=1}^{p} MPSW_i(N, T) \quad (8)$$

where $MPSW_i(N, T)$ is the *MPSW* of the i^{th} *PAC* and p is the total number of *PAC* in *AccuDistr*(*DB*). □

The greater the *MPSW* and *MPSV*, the better the prediction accuracy of a *PAC* and the overall *AccuDistr*(*DB*).

Among these criteria, *MMRE*, *Pred*, *Strength*, and *MPSW* constitute the *Accuracy* vector and measure individual *PAC*, while *MPSV* measures the overall accuracy of AQUA with respect to *AccuDistr*(*DB*).

Because other existing EBA methods are normally validated in the case of full support, i.e. *Support* equals to the number of objects in DB, we thus consider *MPSW* at full support, denoted by *MPSW_FS*, as a baseline. There may be more than one *PAC* at full support. *MPSW_FS* is taken as the best *MPSW* at full support, which is the first PAC at full support when *AccuDistr*(*DB*) is ordered lexicographically by *Support* (Desc), *MPSW* (Desc), *Pred* (Desc), *T* (Asc), *N* (Asc), *MMRE* (Asc), where Desc and Asc mean descending and ascending order respectively.

4.3 Introducing missing values randomly

In order to see how the prediction accuracy of AQUA is affected by missing values, we introduce missing values in the data set *DB* randomly with an increment of percentage of missing values of the total number of attribute values in *DB*. Following the same naming convention in equation (3.1) and (3.2), the initial percentage of missing values in the data set is denoted by $MisVal_{ini}$%, the step of the increment by $MisVal_{stp}$%, the maximum percentage of missing values by $MisVal_{max}$%. A percentage of missing values $MisVal_{inc}$% in a round of increment *inc* is calculated as $MisVal_{inc} = MisVal_{inc-1} + MisVal_{stp}$, where $MisVal_0 = MisVal_{ini}$. Starting from 1, *inc* is increased until $MisVal_{inc} = MisVal_{max}$. For all the data sets, we set $MisVal_{stp}$=5, and $MisVal_{max}$=90 in this experiment, while $MisVal_{ini}$ is calculated from the given data set *DB*. If $MisVal_{ini} > 0$, $MisVal_1$ is set to the $MisVal_{stp}$ next to $MisVal_{ini}$.

When introducing percentage of $MisVal_{stp}$ missing values, firstly, all the cells in a two-dimensional table, with objects the rows and attributes the columns, are numbered using a positive integer. Secondly, a set of random positive integer numbers is generated in order to introduce $MisVal_{inc}$% missing values. The *NULL* value is thus assigned to a cell according to the random number that represents the location of a value in the data set. If there is already a *NULL* in the cell, next random number is used; this process is repeated until next non-*NULL* cell is reached.

4.4 Experiment process

After introducing a percentage of missing values in the data set *DB*, multiple LOOCV by varying N and T are then conducted to see the prediction accuracy of AQUA. Consequently, a series of accuracy measures, i.e. vector *Accuracy* and *MPSV*, of AUQA are obtained after each increment of missing values are tested. As a measure of the overall prediction accuracy of AQUA, *MPSV* will be used

for each increment of missing values to see how the overall accuracy will be influenced by the increasing percentage of missing values. In addition, *MPSW_FS* will be used as an auxiliary measure of the overall prediction accuracy.

MMRE_N in equation (7) is normalized across increments of missing values so that they are still comparable.

The experiment process is then described in the following pseudo code. Scatter plots of *MPSV* and *MPSW_FS* versus *MisVal* over the seven data sets in Table 1 will be analyzed in next section respectively.

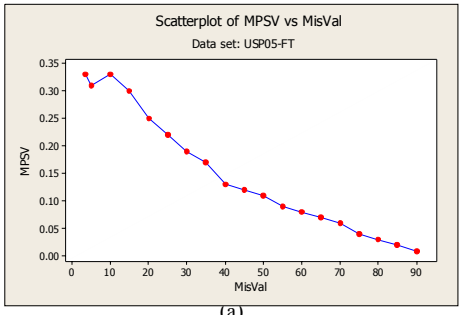
```
Data set DB;
N_max = minimum(Card(DB), 50); --Maximum threshold of N
Calculate MisVal_ini in DB;
FOR MisVal=MisVal_ini STEP BY MisVal_stp TO MisVal_max DO
    Introduce MisVal_stp% missing values into DB randomly;
    FOR N=1 STEP BY 1 TO N_max DO
        FOR T=0 STEP BY 0.1 TO 0.9 DO
            DO LOOCV for DB;
            Obtain the following variables:
            MMRE(N, T), Pred(N, T), Support(N, T),
            Strength(N, T), MPSW(N, T)
        END FOR; --T
    END FOR; --N
    MPSV(MisVal) = Average of MPSW(N, T);
    Find the best MPSW(N, T) at full support:
        MPSW_FS(MisVal);
END FOR; --MisVal
```

5. Experiment results

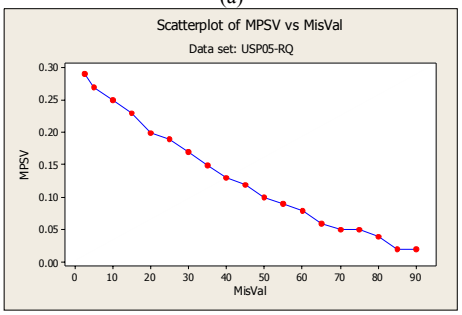
5.1 Testing of hypothesis H1

In order to study hypothesis H1, scatter plots of the aggregated criterion *MPSV* versus *MisVal*, the percentage of missing values in a data set, are presented in (a) to (g) of Figure 1 for all the seven data sets under investigation.

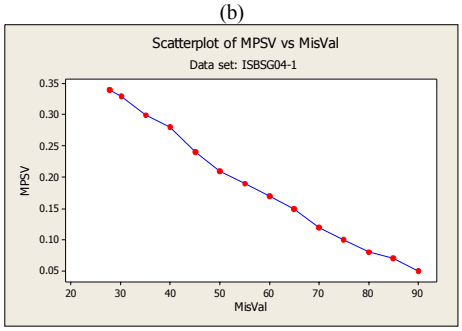
We can see from Figure 1 the general trends that accuracy is decreasing as *MisVal* increases. There is a small increase at one point over data set USP05-FT in (a). This increase of *MPSV* can be explained in terms of attribute selection in that a subset of attribute may produce better prediction accuracy than the whole set of attributes [18], because *NULL* values lead to the omission of the attributes that contains the *NULL* values in the calculation of global similarity measure *Gsim* in equation (1) according to operations (P_1) to (P_3). Nevertheless, this increase takes place only for a few subsets of attributes in a specific data set. In summary of the results from all the seven data sets, *MPSV* decreases as *MisVal* increases. Therefore, null hypothesis H10 is refuted and consequently H11 is supported.



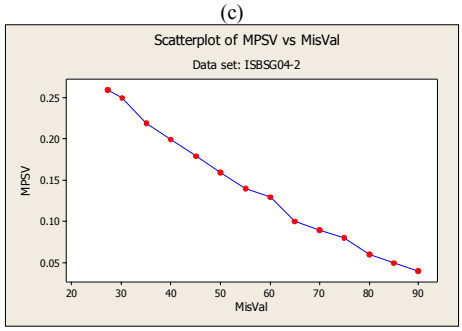
(a)



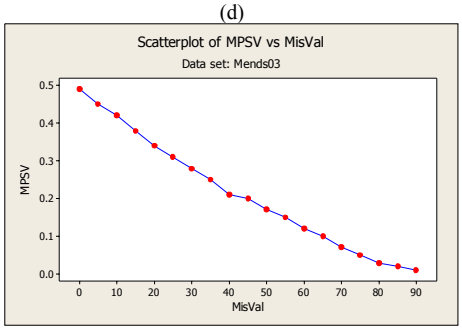
(b)



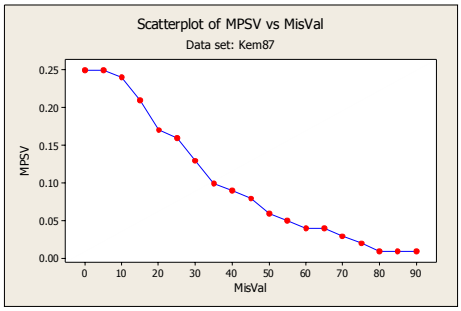
(c)



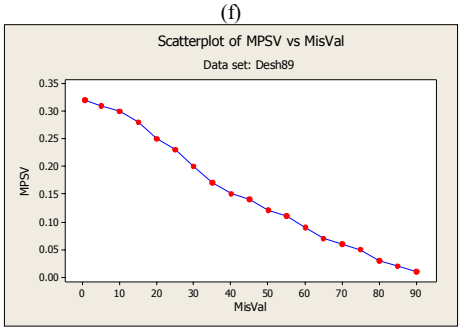
(d)



(e)

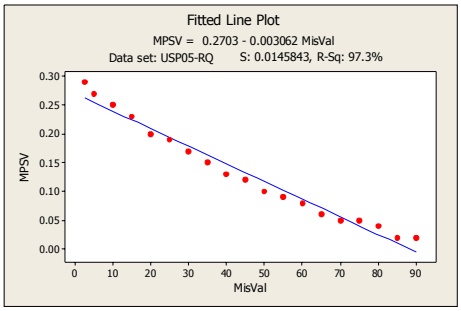


(f)

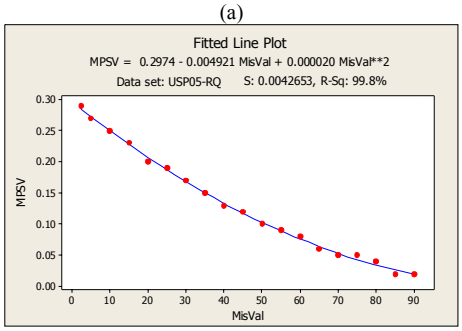


(g)

Figure 1. Scatter plots of *MPSV* vs. *MisVal* over seven data sets.



(a)



(b)

Figure 2. Fitted line plots of data set USP05-RQ

5.2 Testing of hypothesis H2

When testing hypothesis H2, regression is used to generate the fitted functions of the scatter plots in Figure 1. All the scatter plots are tested using linear, quadratic, and cubic regressions. Because of the consistent testing results from all the data sets, only the results of data set USP05-RQ are presented as a representative in Figure 2, in which (a) is the linear and (b) is the quadratic fitted functions of the fitted line plots respectively. In the fitted line plots, S represents the estimated standard deviation of the error in the model. The smaller the S, the better the fitted function is. R-Sq represents Coefficient of determination; indicates how much variation in the response is explained by the model. The greater the R-Sq, the better fitted the function.

In summary of the fitted functions over the seven data sets, the best fitted function is quadratic, because all the scatter plots are better fitted by quadratic functions than the linear functions in terms of smaller S and greater R-Sq. This is demonstrated by the comparison of the values about S and R-Sq between (a) and (b) over data set USP05-RQ. Meanwhile, there are not cubic functions generated due to the fact that the coefficients of the cubic terms are always zero.

Based on the results of the fitness testing that are illustrated in Figure 2, null hypothesis H20 is refuted, H21 is thus supported. The most fitted function is in quadratic form.

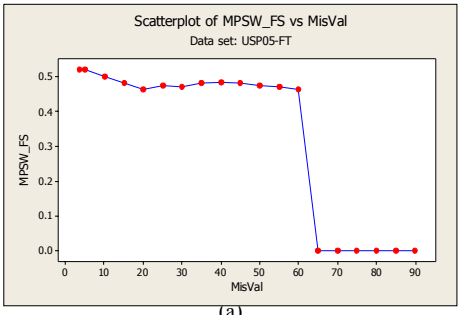
5.3 Testing of hypothesis H3

Based on the testing results of hypotheses H1 and H2, our concern now is about hypothesis H3: whether there is an approximate upper limit of the percentage of missing values, based on which we can decide if AQUA is still applicable or not in terms of overall prediction accuracy.

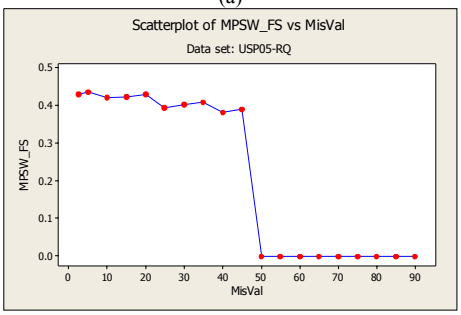
In addition to the scatter plots about *MPSV* in Figure 1, scatter plots about *MPSW_FS* are also observed, as presented in (a) to (g) of Figure 3.

It can be seen from Figure 3, the general tendency of *MPSW_FS* in dependence of *MisVal* is decrease, which also confirms the testing results of H1 and H2 using *MPSV*. The short increases in (a) to (g) are explained in the same way of that about *MPSV* in section 5.1 in terms of attribute selection.

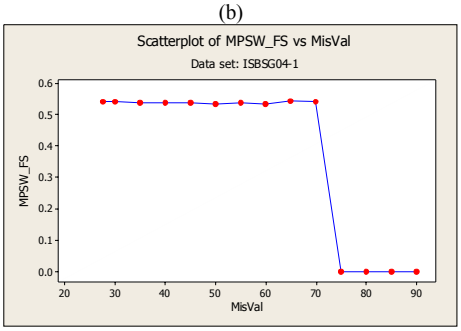
Besides, we observe a sharp decrease of *MPSW_FS* in every data set. This sharp change is caused by the fact that *MPSW_FS* does not exist after a certain percentage of missing values is introduced. In this case, only some of the objects can be estimated, the applicability of AQUA is thus considered quite limited. Based on this fact, we determine the sharp change as a turn point— AQUA is not applicable beyond the *MisVal* of the turn point. Therefore, H30 is refuted and H31 is thus supported.



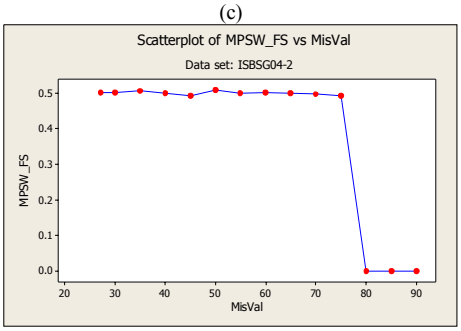
(a)



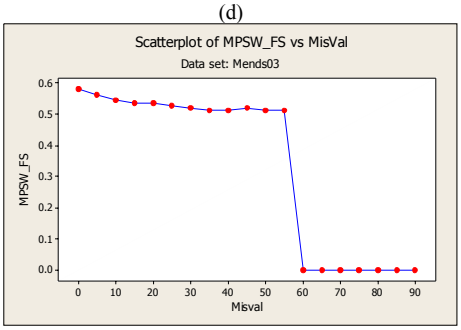
(b)



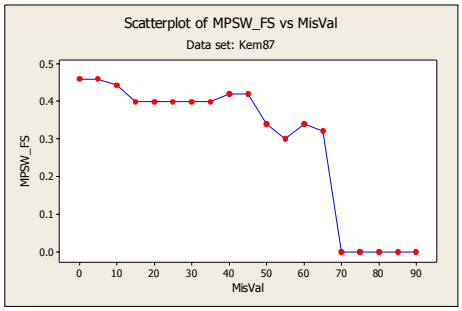
(c)



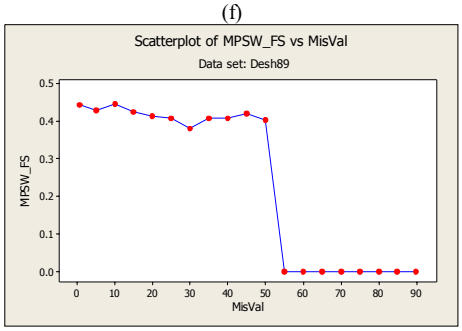
(d)



(e)



(f)



(g)

Figure 3. Scatter plots of *MPSW_FS*

In summary of all the scatter plots in Figure 3, the turn points in terms of *MisVal* are distributed between 45 and 75, while *MPSV* values distribute from 0.07 to 0.24 correspondingly, which means relative low values of *MPSV*. Because 45 is the earliest turn point for all the data sets, we determine that 40 is the upper limit in terms of *MisVal* in order for AQUA to be applicable for any data set, i.e. *MisVal*% ≤ 40%.

5.4 Further observation

From the scatter plots in Figure 1 and Figure 3, we also observe: the more number of attributes (*m*), as well as objects (*n*), in a data set, the better the overall prediction accuracy of AQUA as *MisVal* increases. A better prediction accuracy here means a latened sharp change in the scatter plot of *MPSW_FS*, and a greater value of *MPSV* in the scatter plot at a given turn point.

This can be seen from ISBSG04-1 and ISBSG04-2, which contain much more attributes and objects than other data sets, but have the latest sharp change. As for *MPSV*, if the turn point of the sharp change of ISBSG04-1 is taken as a reference, i.e. *MisVal*=70, the corresponding *MPSV* values of all the data sets are summarized in Table 2.

Among the seven data sets, ISBSG04-1 and ISBSG04-2 have the latest sharp change and they have the greatest *MPSV* values when considering *MisVal*=70. Better *MPSV* values are obtained for ISBSG04-1 and ISBSG04-2 if *MisVal*=75 from the turn point of ISBSG04-2 is taken as a reference.

We conclude from this observation that the more attributes, as well as objects, in a data set, the more applicable AQUA could be as the percentage of missing values increases.

Table 2. Data set size vs. *MisVal* vs. *MPSV*

Data set	Size (*m***n*)	*MisVal* of turn point	*MPSV* at *MisVal*=70
USP05-FT	14*121	60	0.06
USP05-RQ	14*76	45	0.05
ISBSG04-1	**24*285**	**70**	**0.13**
ISBSG04-2	**24*158**	**75**	**0.08**
Mends03	6*34	55	0.07
Kem87	5*15	45	0.03
Desh89	10*81	50	0.06

6. Threats to validity

6.1 Internal validity

This study assumes and tests a causal relationship between the prediction accuracy of AQUA and the percentage of missing values in the data set used by AQUA. Threat to internal validity mainly comes from the measurement criteria of the prediction accuracy of AQUA.

The criterion used in this study for measuring the overall prediction accuracy of AQUA is *MPSV*, which is the weighted average of *MPSW* that is an aggregation of normalized *MMRE*, *Pred*, and *Strength*. While using a single criterion to measure the overall prediction accuracy, the influence from each individual criterion may be neutralized by others' with the aggregation, and further the averaging. This situation degrades the usability of *MPSV* when it is used alone.

In fact, we have considered other criteria as well for this analysis, such as the total number of objects that can be estimated (sum of *Support*) and the total number of estimates (*MPSY* [18]) that satisfy the acceptable thresholds of *MMRE* and *Pred* (i.e. *MMRE*≤25% and *Pred*(0.25)≥75% as proposed by Conte et al. in [17]). The results from these criteria also support the alternative hypotheses. The threat of using a single criterion *MPSV* to the internal validity can be alleviated if these criteria are used along with *MPSV*. However, the results of these criteria are not presented in this paper due to space limitation.

6.2 External validity

We discuss the threat to the external validity of this study in terms of the generalization of the experiment results and the design of the experiment.

As formulated in the research questions, the context of this study includes (i) the data sets; (ii) the EBA method; and (iii) the way of dealing with missing values.

For hypothesis testing based on inferential statistics, the sampling space must be randomly selected. The seven data sets for this study are considered randomly selected and sufficiently representative based on the following observations: (i) they are collected at different periods of time by different organizations over the world; (ii) they are software projects in different application domains; (iii) they have different sizes in terms of number of attributes and objects; (iv) they have different percentages of missing values; (v) they have different types of attributes in terms quantitative and non-quantitative metrics.

However, the number of data sets used is only seven, which is quite small in terms of statistical inference based on which we draw the conclusions. A small size sampling space may cause bias in the results.

The EBA method used for testing the hypotheses is AQUA, in which equal weights of attributes and only one type of similarity measure for each attribute type are used. Other options of attribute weighting and selection, similarity measures, and adaptation strategies are not tested.

The *NULL* value and corresponding operations are defined in AQUA, hence, are used in this study. Other techniques dealing with missing values are not tested either.

Therefore, the results of this study are obtained in the context of EBA method AQUA. Changes to the context may lead to different results.

Nevertheless, the design of this study can be used to test the dependency of the prediction accuracy of other EBA methods on the missing values when other techniques are used to handle issues such as missing values, attribute weighting and selection, and similarity measures.

7. Conclusions and future work

This study provides new insights into EBA method AQUA with knowledge about the prediction accuracy of AQUA in dependence of the percentage of missing values in the data set. The following results were obtained:

(1) The tendency of the overall prediction accuracy of AQUA decreases as the percentage of missing values in a data set increases. This tendency follows a quadratic form of function.

(2) The upper limit of missing values for any data set in general is determined as 40%, in terms of the

applicability of AQUA.

(3) The more attributes, as well as objects, in a data set, the better applicability of AQUA could be as the percentage of missing values increases.

The experiment method of this study can be used for impact analysis of missing values for organizations that grow their data set periodically by adding new data that may contain missing values. The fitted quadratic functions can help predict the overall prediction accuracy of AQUA given a certain percentage of missing values. On the other hand, if missing values are reduced in some way, the quadratic functions can also help analyze how the prediction accuracy of AQUA might be improved over a given data set.

Our future research on this topic will be directed to the analysis of other techniques dealing with missing values, such as imputation, and compare with the use of *NULL*. Other EBA methods will also be explored over more data sets.

Acknowledgements

The authors would like to thank the Alberta Informatics Circle of Research Excellence (iCORE) for its financial support of this research.

References

[1] T. Mukhopadhyay, S. Vicinanza, and M.J. Prietula, "Examining the Feasibility of a Case-based Reasoning Model for Software Effort Estimation", *MIS Quarterly*, Vol. 16, No. 2, 1992, pp 155-171.

[2] M. Shepperd, C. Schofield, "Estimating Software Project Effort Using Analogies", *IEEE Transactions on Software Engineering*, Vol. 23, No. 12, 1997, pp 736-743.

[3] J.Z. Li, G. Ruhe, A. Al-Emran, and M.M. Ritcher, "A Flexible Method for Effort Estimation by Analogy", *Empirical Software Engineering*, Vol. 12, No. 1, 2007, pp 65-106.

[4] E. Mendes, et al., "A Comparative Study of Cost Estimation Models for Web Hypermedia Applications", *Empirical Software Engineering*, Vol. 8, No. 2, 2003, pp 163-196.

[5] K. Strike, et al., "Software Cost Estimation with Incomplete Data", *IEEE Transactions on Software Engineering*, Vol. 27, 2001, pp 890-908.

[6] M. Cartwright, M. Shepperd, and Q. Song, "Dealing with Missing Software Project Data", *Proceedings of the 9th International Symposium on Software Metrics*, 2003, pp 154-165.

[7] I. Myrtveit, E. Stensrud, and U.H. Olsson, "Analyzing Data Sets with Missing Data: An Empirical Evaluation of Imputation Methods and Likelihood-Based Methods", *IEEE Transactions on Software Engineering*, Vol. 27, 2001, pp 999-1013.

[8] N. Ohsugi, et al., "Applying Collaborative Filtering for Effort Estimation with Process Metrics", *PROFES'04: 5th International Conference on Product Focused Software Process Improvement*, LNCS 3009, 2004, Japan, pp 274-286.

[9] M.M. Richter, "On the Notion of Similarity in Case-Based Reasoning", *Mathematical and Statistical Methods in Artificial Intelligence* (Eds. G. Della Riccia, et al.), Springer Verlag, 1995.

[10] C. Lott, H. Rombach, "Repeatable Software Engineering Experiments for Comparing Defect-Detection Techniques", *Empirical Software Engineering*, Vol. 1, No. 3, 1996, pp 241-277.

[11] B. Johnson, L. Christensen, *Educational Research—Quantitative and Qualitative Approaches*, Allyn and Bacon, USA, 2000.

[12] ISBSG, Data R8, *International Software Benchmark and Standards Group*, www.isbsg.org, October 18, 2005.

[13] C.F. Kemerer, "An Empirical Validation of Software Cost Estimation Models", *Communication of the ACM*, Vol. 30, No. 5, 1987, pp 436-445.

[14] J.M. Desharnais, "Analyse statistique de la productivitie des projets informatique a partie de la technique des point des function", *Masters Thesis*, *University of Montreal*, 1989.

[15] S.J. Sayyad, T.J. Menzies, "The PROMISE Repository of Software Engineering Databases", *School of Information Technology and Engineering, University of Ottawa*, Canada, 2005. Available: http://promise.site.uottawa.ca/SERepository.

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

[17] S.D. Conte, H. Dunsmore, and V.Y. Shen, *Software engineering metrics and models*, Benjamin-Cummings Publishing Co. Inc., 1986.

[18] J.Z. Li, G. Ruhe, "A Comparative Study of Attribute Weighting Heuristics for Effort Estimation by Analogy", *Proceedings of the ACM/IEEE International Symposium on Empirical Software Engineering (ISESE'06)*, September 2006, Brazil, pp 66-74.