Sensitivity of results to different data quality meta-data criteria in the sample selection of projects from the ISBSG dataset

Marta Fernández-Diego
Universidad Politécnica de Valencia
Camino de Vera, s/n
46022 Valencia, Spain
(+34) 96 387 76 85
marferdi@omp.upv.es

Mónica Martínez-Gómez
Universidad Politécnica de Valencia
Camino de Vera, s/n
46022 Valencia, Spain
(+34) 96 387 76 85
momargo@eio.upv.es

José-María Torralba-Martínez
Universidad Politécnica de Valencia
Camino de Vera, s/n
46022 Valencia, Spain
(+34) 96 387 76 85
jtorral@omp.upv.es

ABSTRACT

Background: Most prediction models, e.g. effort estimation, require preprocessing of data. Some datasets, such as ISBSG, contain data quality meta-data which can be used to filter out low quality cases from the analysis. However, an agreement has not been reached yet between researchers about these data quality selection criteria.

Aims: This paper aims to analyze the influence of data quality meta-data criteria in the number of selected projects, which can have influence in the models obtained. For this, a case study has been selected to gain a more complete understanding of what might be important to focus in future research.

Method: Data quality meta-data selection criteria of some works based on ISBSG dataset which propose prediction models were reviewed first. Considerable attention has been paid to two data quality meta-data variables in ISBSG dataset Release 11 which are Data Quality Rating and Unadjusted Function Point Rating. Secondly, this paper considers data from 830 projects which have been collected from the ISBSG dataset after a preliminary screening. This first screening leads mainly to a subset of projects with comparable definitions in size and effort. Then data quality meta-data criteria are applied in order to infer their influence.

Results: Overall, it seems that data selection criteria, regardless data quality meta-data concerns, involve an important reduction in sample size. From 5052 projects, only 830 are really considered. Then 262 projects remain for analysis if the maximum quality rate is applied for both data quality meta-data variables. But, since the initial data preparation focuses the problem of missingness for a certain purpose, data quality criteria seem not to be the clue for the analysis results. However, some variability has been observed.

Conclusions: Whilst this analysis is supported by a case study, it is hoped that it contributes to a better understanding of the subject. In fact, results found suggest that in those studies where the selection criteria of projects are not very strictly applied, these data quality criteria must be carefully taken into account.

PROMISE2010, Sep 12-13, 2010. Timisoara, Romania

Categories and Subject Descriptors

D.2.9 [**Software Engineering**]: Management – *productivity*.

General Terms

Management, Measurement, Economics.

Keywords

Datasets, data quality meta-data, functional size, effort, prediction models, empirical research, software projects.

1. INTRODUCTION

Since data quality in software projects datasets, used both in research and in professional practice, is a fundamental determinant of empirical results, it has become a matter of concern for the community of researchers in software engineering ([4], [23], [24], [46]and [2]). In this sense what data should be used for analysis seems an important question but surprisingly not very developed in the context of effort estimation models according to Shepperd in [46]. From this question two aspects can be derived, which datasets are more reliable and how to choose the specific projects. On one hand, Kitchenham in [23] warned researchers about the problems caused by unbalanced datasets which are that the impact of factors can be concealed and that spurious impacts can be observed. Furthermore, simulated datasets where the true underlying model that generated the data values is known were used in [39] to evaluate different analysis techniques. On the other hand, the problem of missing values is pointed out in [24] by Kitchenham and Mendes related with data preprocessing, in such a way that researchers do not always make explicit how they choose the specific projects they used in their studies.

However, data quality is a large concept that deals with various dimensions, completeness being only one of them. Moreover Liebchen and Shepperd published a systematic literature review [27] under a specific data quality view, accuracy or absence of noise. In a previous work [26], three noise correction methods were compared. Another data quality dimension was considered in [30], timeliness. Among the set of historical projects, the concern was to enable the inclusion of recent projects that are more up-to-date, while excluding projects not representative at all of current practices. Another aspect to consider is that of outliers or atypical observations to the extent that may affect the results of segmented parametric software estimation models [11].

Some datasets, such as ISBSG, contain variables assessing projects or cases data quality, i.e. data quality meta-data, and can be used to filter out low quality cases from the analysis. These meta-data collect the perception of ISBSG on projects data quality [6], which is mainly based in data availability of dataset variables [27].

In [47] a systematic comparison of several missing data techniques is described. Unfortunately, the common approach of handling missing data is to delete the case. However, how to deal with missing data is beyond the scope of this investigation.

Given the importance of replicating and confirming previous results [24], this paper seeks to analyze the sensitivity of the results obtained to different data selection criteria, adopting one specific view of data quality, data quality meta-data. Also concerning sensitivity analysis, a complementary approach to the one adopted in this paper is conducted in [15]. Essentially the authors assume that the data received for planning project will be incomplete and inaccurate and see to find which aspects of their approach are most sensitive to these data problems.

Since one of our interests was criteria adopted by other researchers, the method followed consists first in a partial literature review carried out in order to identify which studies explicitly consider data quality meta-data in the preprocessing of data and their criteria in that regard. Then an empirical analysis is performed to test the effect of data quality meta-data criteria in the number of selected projects, which can have influence in the behavior of models obtained. For this, a case study has been selected to gain a more complete understanding of what might be important to focus in future research.

The remainder of the paper is organized as follows. Next section very briefly describes the ISBSG data quality variables used for analysis in this paper before a literature review is carried out in order to identify which studies explicitly consider data quality meta-data in the preprocessing of data. After that, experimental results of our analysis are given and finally the paper concludes discussing the significance of these results.

2. REVIEW OF DATA QUALITY META-DATA CRITERIA

In this section, a partial literature review was carried out in order to identify which studies, related with prediction models and based on ISBSG dataset, explicitly consider data quality meta-data in the preprocessing of data. Primarily, both ISBSG data quality meta-data variables are presented after a brief introduction on the ISBSG dataset.

2.1 ISBSG data quality variables

The International Software Benchmarking Standards Group (ISBSG) [18] designed and maintains two international public repositories (Software Development & Enhancement with over 5000 software projects and Maintenance & Support with over 470 software applications) in order to improve management of IT resources by both business and government.

Besides the Industry Data Suite Release 11 of Software Development & Enhancement and in order to make educational institutions aware of its activities, data and products and encourage their use, the ISBSG gives access to a subset of repository data for academic research purposes. This subset is used for research in this paper.

In relation with data validation and rating, ISBSG implements two fields in its datasets. Each project submitted to the ISBSG repository is validated against specific quality criteria and rated in four categories. As pointed out in [27], the classification is mainly guided by the completeness of the case, i.e. projects, which means that low quality data are interpreted as possessing high levels of missing values.

The ISBSG definitions of these data quality variables are presented next:

2.1.1 Data Quality Rating

Data Quality Rating contains the quality applied to the project data, as evaluated by the ISBSG quality reviewers. It indicates the reliability of the recorded data. The values admissible for this variable are:

'A = The data submitted was assessed as being sound with nothing being identified that might affect its integrity.

B = The submission appears fundamentally sound but there are some factors which could affect the integrity of the submitted data.

C = Due to significant data not being provided, it was not possible to assess the integrity of the submitted data.

D = Due to one factor or a combination of factors, little credibility should be given to the submitted data.'

However, this variable has not always had these four grades, A, B, C and D. In ISBSG dataset Release 7, it had only three grades, A, B and C, with a not an exactly equal definition. In [38], data quality distribution of the ISBSG Release 7 project data is illustrated in a figure. From ISBSG dataset Release 8, Data Quality Rating variable can take the four grades previously described.

2.1.2 Unadjusted Function Point Rating

This field is applied to the Functional Size data, i.e., the Unadjusted Function Point (UFP) count, as evaluated by the ISBSG quality reviewers. Thus, this variable can be referred to as UFP Rating. It has the same four grades A, B, C, and D to denote the following:

'A= The unadjusted function point count was assessed as being sound with nothing being identified that might affect its integrity.

B= The unadjusted function point count appears sound, but integrity cannot be assured as a single figure was provided.

C= Due to unadjusted function point or count breakdown data not being provided, it was not possible to provide the unadjusted function point data.

D= Due to one factor or a combination of factors, little credibility should be given to the unadjusted function point data.'

In ISBSG dataset Release 8, Gencel and Demirors [12] used for the first time this variable as one of the attributes to filter the dataset, taking advantage of ISBSG rating code of A, B, C, or D which are applied to both the Data Quality and Function Point Count data by the ISBSG quality reviewers. As it will be seen in the following subsection, in a latest paper [13], Gencel et al. chose more strict criteria for this variable.

2.2 Method and results of the review

For the search, the following bibliographic databases were used to make a general search for relevant articles: ACM Digital Library (ACM journals, newsletter articles and conference proceedings), IEEE Xplore (all IEEE online publications), ScienceDirect (Elsevier Reference Works) and Web of Knowledge.

The search term "ISBSG" was inputted into the four search engines. This resulted in 38, 16, 13 and 43 results respectively. Naturally, duplicated articles were eliminated. Also, depending on the number of results and the options of the advanced search, the search was sometimes refined by looking jointly for the terms "ISBSG" and "quality". This was supplemented by a scan of the remainder to determine if they are concerned with quality as well, by searching the term "quality" in the text. At this stage, a hand search was done to identify specifically those studies related with prediction models and based on ISBSG dataset which explicitly consider data quality meta-data in the preprocessing of data.

From the literature review, table 1 synthesizes the information extracted from the identified papers. The corresponding labeling for each column is as follows: (1) Paper reference; (2) Year of publication; (3) ISBSG Release; (4) Data Quality Rating; (5) UFP Rating; (6) Topic related with prediction models; (7) Use of Size/Effort variables; (8) Number of projects selected for analysis. Note that when using more than one sample size, this information is reflected in this last column separated by a slash.

Table 1. Classification of retrieved papers

1	2	3	4	5	6	7	8
[33]	2010	10	A | B	A | B	Duration	Y/N	759
[13]	2009	10	A | B	A | B	Size	Y/N	14
[14]	2009	10	A | B		Substitution cost	Y/Y	176 / 48
[30]	2009	10	A | B		Effort	Y/Y	228
[28]	2009	10	A | B		Effort	Y/Y	909
[37]	2009	7	A | B		Effort	Y/Y	540
[45]	2009	9	?		Effort	Y/Y	373
[1]	2008	9	A | B		Effort	Y/Y	600
[3]	2008	7	A | B		Defect	Y/N	91
[12]	2008	8	A | B	A | B | C	Size	Y/Y	103
[16]	2008	10	A | B		Team size	Y/Y	19 / 6
[17]	2008	7	A | B		Effort	Y/Y	591
[22]	2008	9	A		Effort	Y/Y	502
[38]	2008	7	A | B		Effort	Y/Y	540
[44]	2008	9	?		Effort	Y/Y	99
[48]	2008	8	A | B		Size	Y/Y	63
[5]	2007	4	?		Duration	Y/Y	312
[21]	2007	10	A | B | C		Productivity	Y/Y	3322
[36]	2007	7	A | B | C		Effort	Y/Y	217
[29]	2006	6	A | B		Effort	Y/Y	89 / 12
[34]	2005	8	A | B		Effort	Y/Y	339
[43]	2005	7	A | B		Effort	Y/Y	52
[42]	2005	7	Recommended		Effort	Y/Y	166
[32]	2005	9	A | B		Effort	Y/Y	672 / 184
[20]	2001	6	A		Effort	Y/Y	324
[19]	2000	5	?		Effort	Y/Y	145

In total 26 papers were retrieved, 17 related with effort estimation, 3 with size estimation, 2 with duration estimation, 1 with team size estimation, 1 with substitution cost estimation, 1 with defect estimation and 1 with productivity.

Concerning Data Quality Rating, most papers, 17 in total, adopted the strategy of only using data graded as A or B. However, two of them ([22] and [20]) only accepted A quality projects, one paper [21] only discarded D quality projects and another one [36] worked with all grades, A, B and C of Release 7. Four papers ([45], [44], [5] and [19]) mentioned data quality in the filtering criteria, but they did not explicit how the filtering was performed in this regard. Finally, one paper [42] considered the data quality rating following the recommendation of ISBSG, which is supposed to accept projects with A and B rating only.

Note that only three papers out of 26 worked explicitly with UFP Rating variable in the preprocessing of data. Two of them ([12] and [13]) share the same first author and differ in the criteria, and the other one [33] is the latest.

In the papers reviewed, the common approach of handling missing data was to delete the case. Interestingly, two papers ([34] and [42]) proposed another approach which was imputing missing data, thus avoiding deletion of observations. Also in [36], the authors considered data quality as an interaction factor in their analysis, thus taking binary values of high quality (A or B) and low quality (C) data.

Beyond this systematic search, three other interesting papers ([35], [8] and [7]) were found involved with data quality meta-data in the ISBSG dataset. The two first did not propose any prediction model. In [35], the ratios of outliers which have data considered of very poor quality (D) were presented. However, the authors did not conclude that Data Quality Rating might explain such outliers' behavior. In [8], an analysis was performed to study the variability in the effort for projects for which data quality is considered to be high (A or B). The third paper [7] proposed a process to maximize the amount of data retained for modeling software development effort, resulting in a subset still including 77 projects of ISBSG Release 9 rated D. Before that, it provided an interestingly review of the treatment of the repository in relation to data retention, considering data quality meta-data too. In addition, it referred also to UFP Rating, but as a data quality indicator, not directly related to software effort estimation.

Finally, other authors worked explicitly with UFP Rating variable in the preprocessing of data. In [41], only projects graded as A or B for both Data Quality Rating and UFP Rating were selected, while in [25] only A quality projects for both variables were accepted. In [31], [10] and [9], authors pushed for a compromise in this regard: A or B for Data Quality Rating and A for UFP Rating.

To summarize, it appears that there is some variability in selection criteria concerning data quality meta-data.

3. EXPERIMENTAL RESULTS

Once a partial literature review has been carried out in order to identify which studies explicitly consider data quality meta-data in the preprocessing of data and we are aware of their criteria in that regard, an empirical analysis is performed to test the effect of data quality meta-data criteria in the number of selected projects, which can have influence in the behavior of the models obtained.

This section is organized in such a way that the strength of the association for general data quality, i.e. Data Quality Rating variable, by functional size quality, i.e. UFP Rating variable, is first studied. Then the influence of data quality meta-data criteria in the number of selected projects is analyzed and finally, some fluctuation in the behavior of the models obtained is pointed out in the last subsection through a case study.

3.1 Strength of the association for general data quality by functional size quality

This subsection starts by first showing the immediate consequences derived from being more or less selective in relation with data quality meta-data reported. Table 2 presents the number of projects remaining once the data quality criteria for both variables are applied. From 5052 projects contained in the ISBSG dataset Release 11, if it is decided for example to work with the highest grade (A) for both data quality variables, the sample falls down to 676 projects.

Table 2. Data quality meta-data distribution in the ISBSG dataset Release 11

		Data Quality Rating			
		A | B | C | D	A | B | C	A | B	A
UFP Rating	A | B | C | D | ND	5052	4907	4744	928
	A | B | C | D	4512	4389	4243	840
	A | B | C	4497	4385	4243	840
	A | B	3369	3288	3184	836
	A	2328	2277	2202	676

ND = No Data

Since the categories of both of these variables are ordered, measures can be utilized to determine the direction (obviously positive relationship in this case, the variables change in the same direction) and quantify the strength of the association.

In Table 3, the cross tabulation shows no clear pattern for both data quality variables. If any exists, it may be that projects with better general data quality are also with better functional size quality, which means that as general quality increases, functional size quality increases too.

Before proceeding, 540 projects with no data in Unadjusted Function Point Rating variable were removed from the analysis, resulting in 4512 projects. For some of them - in fact 186 projects - measured in Lines of Code (LOC), this field does not make any sense; furthermore, 345 projects measured in COSMIC with no functional size quality indication were also removed.

Table 3. Cross tabulation for Data Quality Rating by UFP Rating

		Data Quality Rating				Total
		A	B	C	D	
UFP Rating	A	676	1526	75	51	**2328**
	B	160	822	29	30	**1041**
	C	4	1055	38	31	**1128**
	D	0	0	4	11	**15**
Total		**840**	**3403**	**146**	**123**	**4512**

In fact, the approximate significance value of Gamma is equal to 0.000. Since this is less than 0.05, we can conclude that there is a statistically significant relationship between both variables. Also, the value of Gamma=0.525 tells us that we will make 52.5% fewer errors predicting general data quality when functional size quality is taken into account, which points out a moderate positive association between both variables. The inverse statement is also logically true. Finally, the value of Kendall's tau-b=0.205 with an approximate significance value also equal to 0.000 reinforces that there is a statistically significant relationship between both variables.

3.2 Influence of data quality meta-data in the sample size

3.2.1 Initial data preparation

Although the initial dataset comprised 5052 projects, a number of observations were removed in order to analyze the sensitvity of data quality requirements for a specific application, productivity trend over time, which was selected as our case study.

Due to the fact that the ISBSG dataset is a large heterogeneous one, a data preparation process was required before applying any analysis. To get a minimum of homogeneity in the samples to be analyzed, the following rules were first applied, summarized in table 4 and adapted from Lokan and Mendes [28]. The first one ensures that all projects are dated specifically with the implementation date, the following two provide comparable definition for size, and finally the last three supply comparable definition for effort and assure whole life cycle projects.

Actually, the selection process has been applied regarding only the matter of study, i.e., evolution of productivity over time, without considering data quality ratings which will be applied later on.

After the initial filtering, by following these steps, a subset of 830 projects resulted from the initial 5052 projects, all with comparable definitions for size and effort, and with a date reference in order to analyze the evolution of productivity over time.

Second column in the table 4 indicates the projects remaining after each rule; the third column shows the number of projects removed because of each rule. Thus a total of 4222 projects were eliminated from the analysis, representing 83.5% of total.

Table 4. Selection criteria for productivity concerns

Selection criteria	Projects remaining	Projects removed
Project implementation date known[1]	4407	645
IFPUG version 4.0 or later[2]	2277	2130
Unadjusted Function Points known[3]	1565	712
Development team effort known[4]	1329	236
Effort across the whole life cycle[5]	941	388
Web projects removed[6]	830	111

3.2.2 Data quality selection

After initial data preparation, we can now proceed to examining the productivity evolution over time. However, our main concern in this paper is to infer the influence of data quality on the results obtained. Thus, in this subsection further selection criteria were considered, actually those related with data quality.

Table 5 indicates the influence of data quality ratings in the count of the remainder of projects from the preliminary screening in subsection 3.2.1.

Table 5. Remained projects applying different data quality ratings

		Data Quality Rating			
		A | B | C | D	**A | B | C**	**A | B**	**A**
UFP Rating	**A | B | D**	830	820	802	316
	A | B	828	820	802	316
	A	545	537	521	262

After the preliminary screening, i.e. for the 830 projects, none of them lack of value in variable UFP Rating (empty string actually). This makes sense since at this stage we are dealing with projects all of them measured with IFPUG. Clearly it is the most representative (75.2% of total projects) functional size measurement method in ISBSG dataset and only 6 of these projects lack of value in variable UFP Rating, contrary to COSMIC projects where all of them remain in this sense without evaluation.

Moreover, it can be observed in table 5 that grade C for UFP Rating variable has also disappeared. Note that for all 1128 projects so classified (see table 3), the functional size variable, i.e. the unadjusted function point count, lacks of value, and projects were removed after the preliminary screening.

[1] V19~=""

[2] CHAR.INDEX(V66,'IFPUG 4')~=0|(V66="IFPUG"&V4>"1994")

[3] V6~=""

[4] V9~=""

[5] V11=V9

[6] CHAR.LENGTH(V50)=0

Overall, it seems that the data selection criteria, regardless data quality rating concerns, suppose an important filtering, since from 5052 projects only 830 are really considered. Then 262 projects remained for analysis applying the maximum quality rate for both variables. In this sense, data quality criteria seem not to be the clue for the analysis results. Actually the initial data preparation focuses the problem of missingness for a certain purpose. For those studies where the data selection criteria, regardless data quality concerns, are very strictly applied, data quality criteria may not add value, since ISBSG data quality classification is principally guided by the completeness of the project [27].

However, some variability has been pointed out.

3.3 The case study: Productivity evolution over time

Since the following analysis considers both qualities, in order to simplify the notation, capital letters have been chosen for Data Quality Rating grades and lower case letters for UFP Rating grades.

This subsection considers the following study cases derived from table 5 which are strongly different in the count of projects removed: a sample of 802 projects corresponding to (A | B) & (a | b) quality, a sample of 521 projects corresponding to (A | B) & a quality, a sample of 316 projects corresponding to A & (a | b) quality and a sample of 262 projects corresponding to A & a quality. Note that in our previous studies in [10] and [9], the analysis was performed with a sample of 521 projects, where data quality meta-data criteria were first applied.

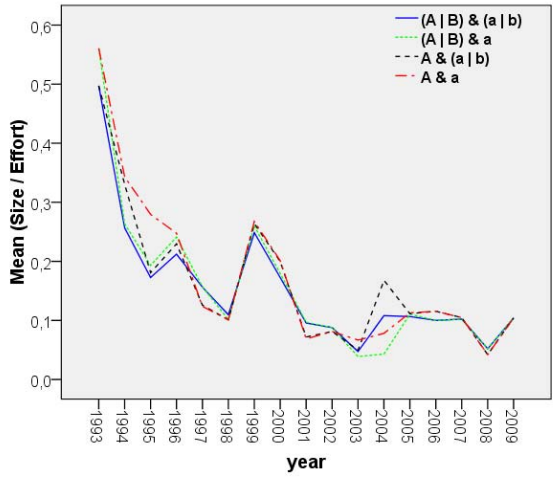
Moreover, these cases fit with higher general data quality (A and A | B), which is taken into account in the papers reviewed (even forcing it until A | B | C and in some cases to D) and a sweep of functional size quality. However, since A and A | B general data quality is retrieved for analysis, grades C and D of functional size quality automatically disappear. Thus the four possible combinations taking into account both higher qualities have been considered for analysis.

We begin first by investigating productivity measured as the ratio of size to effort (in functional size units per hour), in order to give some overall idea of the sensitivity of the mean productivity trend to different quality requirements. A visual inspection of figure 1 indicates some differences between the four curves. Each of them represents the evolution of mean productivity over time. Actually, the A & a quality curve appears to be the smooth approximation of the other ones.

Next, for the different cases of study, a linear regression model is now built relating effort with size, assuming a production equation of the form:

$$\ln(Effort) = C + B\ln(Size)$$

Figure 1. Evolution of mean productivity over time



In order to analyze the evolution over time of this production function and following the model proposed by Premraj et al. in [40], the equation can be split as:

$$\ln(Effort) = C + B_{93}\ln(Size_{93}) + B_{94}\ln(Size_{94}) + \cdots + B_{09}\ln(Size_{09})$$

The idea is to create as many variables as years have the time horizon. Each of these variables represents the size of the projects implemented in the same year and zero otherwise. Thus Byear coefficients are the regression coefficients of each independent variable.

Since the trend in productivity over time is not the aim of this paper, but the influence of the data quality data requirements in the trend on productivity over time, it was decided not to include any other possible explained variable influencing productivity and considered in [40], such as business sector, project type, company and process model particularly.

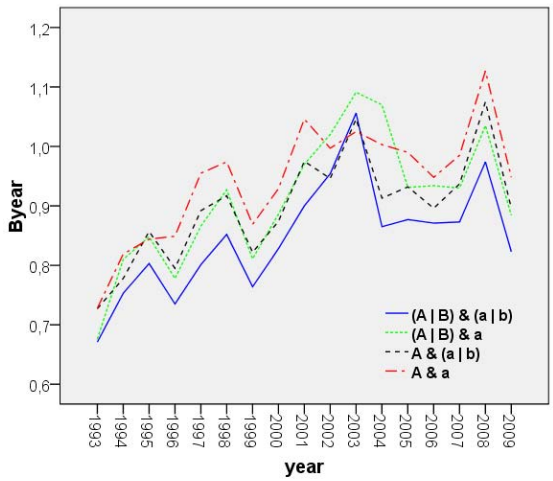
Table 6 summarizes some statistical information about the models, such as the constant value from regression analysis, the coefficient of determination and Durbin-Watson test used to determine whether the residuals are independent. Since values close to 2.0 for the Durbin-Watson statistic indicate that there is no serial correlation, diagnosis is appropriate.

Table 6. Statistical summary of the models

	(A | B) & (a | b)	(A | B) & a	A & (a | b)	A & a
Constant	3.216	2.883	2.820	2.538
R^2	0.557	0.516	0.557	0.568
Durbin-Watson	1.824	1.906	1.933	1.959

Finally, figure 2 shows the evolution of Byear over time for the four previous selected combinations, thus taking into account both general data quality and functional size quality. Note that a low value of Byear implies that less effort is required to implement a project with a fixed size, so the more productive a software project is. In this sense, this plot also shows the trend towards a decline in productivity over time, even if variable, already indicated in figure 1.

Figure 2. Evolution of Byear over time



From the regression model the 95% confidence limits were derived together with the value for each Byear. Table 7 shows the information for the two most extreme cases considered for quality levels (A | B) & (a | b) and A & a.

Table 7. Confidence intervals of the regression coefficients

	(A | B) & (a | b)			A & a		
	B_{year}	Lower Bound	Upper Bound	B_{year}	Lower Bound	Upper Bound
B_{93}	0.671	0.553	0.788	0.728	0.571	0.885
B_{94}	0.753	0.687	0.820	0.819	0.702	0.937
B_{95}	0.803	0.736	0.870	0.844	0.693	0.996
B_{96}	0.735	0.617	0.852	0.849	0.669	1.029
B_{97}	0.801	0.709	0.893	0.955	0.798	1.111
B_{98}	0.852	0.757	0.948	0.974	0.813	1.136
B_{99}	0.764	0.700	0.829	0.869	0.747	0.992
B_{00}	0.828	0.767	0.890	0.929	0.809	1.050
B_{01}	0.900	0.811	0.988	1.046	0.831	1.261
B_{02}	0.954	0.888	1.019	0.997	0.789	1.205
B_{03}	1.056	0.965	1.147	1.025	0.769	1.281
B_{04}	0.865	0.795	0.936	1.003	0.629	1.377
B_{05}	0.877	0.810	0.943	0.990	0.858	1.123
B_{06}	0.871	0.783	0.959	0.948	0.810	1.087
B_{07}	0.873	0.801	0.944	0.985	0.868	1.103
B_{08}	0.974	0.869	1.080	1.127	0.978	1.276
B_{09}	0.823	0.500	1.146	0.948	0.618	1.278

Even though the confidence intervals do overlap, the overlap in confidence intervals is small for some Byear. Because this is a conservative method of testing, the overlap is not conclusive.

Overall, two observations can be derived from the analysis of the four selected cases.

First, the distribution of projects over time is not constant, nor is the count of removed projects while increasing data quality requirements. This is confirmed by information presented in table 8. As it can be seen, a minimum number of projects per year has not been taken into account regarding the regression models, although all coefficients are statistically significant in the four models. However, a grouping of years by triennia is been investigated in order to avoid this limitation.

Table 8. Distribution of projects over time depending on data quality requirements

	(A | B) & (a | b)		(A | B) & a		A & (a | b)		A & a	
		%		%		%		%
1993	5	0.6	4	0.8	5	1.6	4	1.5
1994	45	5.6	42	8.1	25	7.9	24	9.2
1995	47	5.9	36	6.9	17	5.4	7	2.7
1996	10	1.2	6	1.2	8	2.5	5	1.9
1997	19	2.4	18	3.5	13	4.1	12	4.6
1998	16	2.0	15	2.9	11	3.5	11	4.2
1999	68	8.5	62	11.9	57	18.0	56	21.4
2000	86	10.7	74	14.2	62	19.6	57	21.8
2001	27	3.4	25	4.8	6	1.9	4	1.5
2002	144	18.0	141	27.1	3	.9	3	1.1
2003	46	5.7	4	0.8	7	2.2	1	0.4
2004	64	8.0	7	1.3	17	5.4	1	0.4
2005	158	19.7	20	3.8	27	8.5	19	7.3
2006	22	2.7	22	4.2	16	5.1	16	6.1
2007	32	4.0	32	6.1	30	9.5	30	11.5
2008	12	1.5	12	2.3	11	3.5	11	4.2
2009	1	0.1	1	0.2	1	0.3	1	0.4
Total	**802**	**100.0**	**521**	**100.0**	**316**	**100.0**	**262**	**100.0**

The major percentage decrease of projects occurs in the interval between 2001 and 2005, which includes year 2003; in this range strange crossings can be observed between the four curves in figure 2.

Second, the variation in productivity mean between each case should also be taken into account in order to explain the differences obtained between the models. Table 9 presents some basic statistics for size to effort productivity ratio, such as mean and standard deviation for each combination of Data Quality Rating by UFP Rating.

Table 9. Mean and standard deviation of productivity ratio depending on data quality requirements

UFP Rating	Mean			Standard deviation		
	Data Quality Rating					
	A	**B**	**Total**	**A**	**B**	**Total**
a	0.1950	0.1081	**0.1518**	0.30244	0.14929	**0.24261**
b	0.1330	0.0961	**0.1032**	0.13804	0.22713	**0.21321**
Total	**0.1844**	**0.1025**	**0.1347**	**0.28203**	**0.18955**	**0.23375**

Moreover, an ANOVA analysis has been performed to determine whether any differences among means are greater than would be expected by chance. Before performing the analysis, the productivity ratio was logarithmic transformed in order to approach to normal distribution. Actually both factors, Data Quality Rating and UFP Rating, are found to be statistically significant, but not the interaction between them. Since the assumption of homogeneity of variances is not met, an ANOVA analysis of the square residuals stresses the point that the only factor influencing variance is Data Quality Rating. Note that now the significance of Levene's test is above 0.05, which suggests that the equal variances assumption is no more violated.

4. CONCLUSIONS

First, this study presented criteria adopted by other researchers with respect to data quality meta-data in the preprocessing of data. After that, working on the latest release of ISBSG dataset, some variability has been pointed out in the behavior of the models obtained by varying these criteria. As a conclusion, the observed variability could arise for two reasons: either there is a reduction in the sample size and even a different proportion of removed projects over time, or a variation in average productivity which differs statistically between the samples considered.

Actually, the initial data preparation focuses the problem of missingness for a certain purpose, being in a certain extent redundant with data quality meta-data concerns and so the decision adopted in this regard could have little impact. However, results found suggest that in those studies where the selection criteria of projects, regardless data quality meta-data concerns, are not very strictly applied, these data quality criteria must be carefully taken into account.

Albeit generalizations are difficult, we believe that this study can provide a better understanding of what might be important to focus in future research. In fact, there is awareness of the need of reproducing these results for both variables often used in most prediction models: functional size (UFP count actually) and effort, separately. The aim would be to show how data quality meta-data criteria affect the sample under study, in relation with these two variables separately, in order to be able to better interpret this study.

Moreover, we have already seen that the interaction of both factors, Data Quality Rating and UFP Rating, is found to be statistically significant in the average effort, not their effect separately, and that the only factor influencing variance is UFP Rating. On the other hand, both simple and interaction effects influence the average size, but only Data Quality Rating and the interaction influence size variance.

5. ACKNOWLEDGMENTS

Our thanks to the reviewers for their helpful contributions.

6. REFERENCES

1. Ahmed, F., Bouktif, S., Serhani, A., and Khalil, I. Integrating Function Point Project Information for Improving the Accuracy of Effort Estimation. *2008 The Second International Conference on Advanced Engineering Computing and Applications in Sciences*, (2008), 193-198.
2. Berlin, S., Raz, T., Glezer, C., and Zviran, M. Comparison of estimation methods of cost and duration in IT projects. *Information and Software Technology 51*, 4 (2009), 738-748.
3. Bibi, S., Tsoumakas, G., Stamelos, I., and Vlahavas, I. Regression via Classification applied on software defect estimation. *Expert Systems with Applications 34*, 3 (2008), 2091-2101.
4. Boehm, B. Value-Based Software Engineering: Overview and Agenda. In *Biffl, S., Aurum, A., Boehm, B., Erdogmus, H., and Grünbacher, P. Value-Based Software Engineering*. Springer, 2006, 3-14.
5. Bourque, P., Oligny, S., Abran, A., and Fournier, B. Developing Project Duration Models in Software Engineering. *Journal of Computer Science and Technology 22*, 3 (2007), 348-357.
6. Bundschuh, M. and Dekkers, C. *The IT Measurement Compendium*. Springer, 2008.
7. Deng, K. and MacDonell, S.G. Maximising data retention from the ISBSG repository. *12th International Conference on Evaluation and Assessment in Software Engineering*, (2008).
8. Déry, D. and Abran, A. Investigation of the Effort Data Consistency in the ISBSG Repository. *15th International Workshop on Software Measurement*, (2005), 123-136.
9. Fernández-Diego, M., Maciel, J., Elmouaden, S., and Torralba-Martínez, J. Physical productivity evolution of software projects in the ISBSG dataset. *XIV International Congress on Project Engineering*, (2010).
10. Fernández-Diego, M., Maciel, J., Marcelo-Llácer, D., and Torralba-Martínez, J. Software projects size and economies of scale in the ISBSG dataset. *XIV International Congress on Project Engineering*, (2010).
11. Garre, M., Cuadrado, J., Sicilia, M., Charro, M., and Rodríguez, D. Segmented parametric software estimation models: Using the EM algorithm with the ISBSG 8 database. *27th International Conference on Information Technology Interfaces, 2005*, 181-187.
12. Gencel, C. and Demirors, O. Functional size measurement revisited. *ACM Trans. Softw. Eng. Methodol. 17*, 3 (2008), 1-36.
13. Gencel, C., Heldal, R., and Lind, K. On the Relationship between Different Size Measures in the Software Life Cycle. *2009 16th Asia-Pacific Software Engineering Conference*, (2009), 19-26.
14. Haaland, K., Stamelos, I., Ghosh, R., and Glott, R. On the Approximation of the Substitution Costs for Free/Libre Open Source Software. *2009 Fourth Balkan Conference in Informatics*, (2009), 223-227.
15. Harman, M., Krinke, J., Ren, J., and Yoo, S. Search based data sensitivity analysis applied to requirement engineering. *Proceedings of the 11th Annual conference on Genetic and evolutionary computation*, ACM (2009), 1681-1688.
16. Hericko, M., Zivkovic, A., and Rozman, I. An approach to optimizing software development team size. *Information Processing Letters 108*, 3 (2008), 101-106.
17. Huang, S., Chiu, N., and Liu, Y. A comparative evaluation on the accuracies of software effort estimates from clustered data. *Information and Software Technology 50*, 9-10 (2008), 879-888.
18. ISBSG. ISBSG dataset Release 11. *International Software Benchmarking Standards Group*, 2009. http://www.isbsg.org/.
19. Jeffery, R., Ruhe, M., and Wieczorek, I. A comparative study of two software development cost modeling techniques using multi-organizational and company-specific data. *Information and Software Technology 42*, 14 (2000), 1009-1016.
20. Jeffery, R., Ruhe, M., and Wieczorek, I. Using Public Domain Metrics To Estimate Software Development Effort. *Proceedings of the 7th International Symposium on Software Metrics*, IEEE Computer Society (2001), 16.
21. Jiang, Z., Naudé, P., and Comstock, C. An investigation on the variation of software development productivity. *International Journal of Computer and Information Science and Engineering 1*, 2 (2007), 72-81.
22. Keung, J. and Kitchenham, B. Experiments with Analogy-X for Software Cost Estimation. *Proceedings of the 19th Australian Conference on Software Engineering*, IEEE Computer Society (2008), 229-238.
23. Kitchenham, B. A Procedure for Analyzing Unbalanced Datasets. *IEEE Trans. Softw. Eng. 24*, 4 (1998), 278-301.
24. Kitchenham, B. and Mendes, E. Why comparative effort prediction studies may be invalid. *Proceedings of the 5th International Conference on Predictor Models in Software Engineering (PROMISE)*, ACM (2009), 1-5.
25. Koh, T., Selamat, M., and Ghani, A. Exponential Effort Estimation Model Using Unadjusted Function Points. *Information Technology Journal 7*, 6 (2008), 830-839.
26. Liebchen, G., Twala, B., Shepperd, M., and Cartwright, M. Assessing the Quality and Cleaning of a Software Project Dataset: An Experience Report. *10th International Conference on Evaluation and Assessment in Software Engineering*, (2006).
27. Liebchen, G.A. and Shepperd, M. Data sets and data quality in software engineering. *Proceedings of the 4th international workshop on Predictor models in software engineering (PROMISE)*, ACM (2008), 39-44.
28. Lokan, C. and Mendes, E. Investigating the use of chronological split for software effort estimation. *IET Software 3*, 5 (2009), 422-434.
29. Lokan, C. and Mendes, E. Cross-company and single-company effort models using the ISBSG database: a further replicated study. *Proceedings of the 2006 ACM/IEEE International Symposium on Empirical Software Engineering*, ACM (2006), 75-84.
30. Lokan, C. and Mendes, E. Applying moving windows to software effort estimation. *Proceedings of the 2009 3rd International Symposium on Empirical Software Engineering and Measurement*, IEEE Computer Society (2009), 111-122.
31. Maciel, J., Fernández-Diego, M., Sanz-Berzosa, M., and Torralba-Martínez, J. The recent evolution of the ISBSG (International Software Benchmarking Standards Group) software projects dataset. *XIV International Congress on Project Engineering*, (2010).

32. Mendes, E., Lokan, C., Harrison, R., and Triggs, C. A replicated comparison of cross-company and within-company effort estimation models using the ISBSG database. *Software Metrics, 2005. 11th IEEE International Symposium*, (2005), 10 pp.-36.
33. Mittas, N. and Angelis, L. Visual comparison of software cost estimation models by regression error characteristic analysis. *Journal of Systems and Software 83*, 4 (2010), 621-637.
34. Moses, J. and Farrow, M. Assessing Variation in Development Effort Consistency Using a Data Source with Missing Data. *Software Quality Control 13*, 1 (2005), 71-89.
35. Paré, D. and Abran, A. Obvious Outliers in ISBSG Repository of Software Projects: Exploratory Research. *Metrics News 10*, 1 (2005), 28-36.
36. Pendharkar, P.C. and Rodger, J.A. An empirical study of the impact of team size on software development effort. *Inf. Technol. and Management 8*, 4 (2007), 253-262.
37. Pendharkar, P.C. and Rodger, J.A. The relationship between software development team size and software development cost. *Commun. ACM 52*, 1 (2009), 141-144.
38. Pendharkar, P.C., Rodger, J.A., and Subramanian, G.H. An empirical study of the Cobb-Douglas production function properties of software development effort. *Inf. Softw. Technol. 50*, 12 (2008), 1181-1188.
39. Pickard, L., Kitchenham, B., and Linkman, S. An Investigation of Analysis Techniques for Software Datasets. *Proceedings of the 6th International Symposium on Software Metrics*, IEEE Computer Society (1999), 130.
40. Premraj, R., Kitchenham, B., Shepperd, M., and Forselius, P. An empirical analysis of software productivity over time. *11th IEEE International Symposium On Software Metrics (METRICS 2005), IEEE Computer Society*, (2005), 37.
41. Santillo, L., Lombardi, S., and Natale, D. Advances in statistical analysis from the ISBSG benchmarking database. *Proceedings of 2nd Software Measurement European Forum*, (2005).
42. Sentas, P. and Angelis, L. Categorical missing data imputation for software cost estimation by multinomial logistic regression. *Journal of Systems and Software 79*, 3 (2006), 404-414.
43. Sentas, P., Angelis, L., Stamelos, I., and Bleris, G. Software productivity and effort prediction with ordinal regression. *Information and Software Technology 47*, 1 (2005), 17-29.
44. Seo, Y., Yoon, K., and Bae, D. An empirical analysis of software effort estimation with outlier elimination. *Proceedings of the 4th international workshop on Predictor models in software engineering (PROMISE)*, ACM (2008), 25-32.
45. Seo, Y., Yoon, K., and Bae, D. Improving the Accuracy of Software Effort Estimation Based on Multiple Least Square Regression Models by Estimation Error-Based Data Partitioning. *2009 16th Asia-Pacific Software Engineering Conference*, (2009), 3-10.
46. Shepperd, M. Software project economics: a roadmap. *2007 Future of Software Engineering*, IEEE Computer Society (2007), 304-315.
47. Twala, B., Cartwright, M., and Shepperd, M. Ensemble of missing data techniques to improve software prediction accuracy. *Proceedings of the 28th international conference on Software engineering*, ACM (2006), 909-912.
48. Xia, W., Capretz, L., Ho, D., and Ahmed, F. A new calibration for Function Point complexity weights. *Information and Software Technology 50*, 7-8 (2008), 670-683.