Journal of
Information Technology & Software Engineering

Bala and Abran, J Inform Tech Softw Eng 2016, 6:1
DOI: 10.4172/2165-7866.1000171

Research Article Open Access

Use of the Multiple Imputation Strategy to Deal with Missing Data in the ISBSG Repository

Abdalla Bala and Alain Abran*

École de Technologie Supérieure (ÉTS)-University of Québec, Montréal, Québec, Canada

Abstract

Multi-organizational repositories, in particular those based on voluntary data contributions such as the repository of the International Software Benchmarking Standards Group (ISBSG), may be missing a large number of values for many of their data fields, as well as including some outliers. This paper suggests a number of data quality issues associated with the ISBSG repository which can compromise the outcomes for users exploiting it for benchmarking purposes or for building estimation models. We propose a number of criteria and techniques for preprocessing the data in order to improve the quality of the samples identified for detailed statistical analysis, and present a multiple imputation (MI) strategy for dealing with datasets with missing values.

Keywords: Multi-imputation technique; ISBSG data preparation; Identification of outliers; Analysis effort estimation; Evaluation criteria

Introduction

The data repository of the International Software Benchmarking Standards Group (ISBSG) is a multi-organizational dataset of software projects from around the world which is used for benchmarking purposes and in software effort estimation [1]. The ISBSG Group was set up by national software measurement associations to develop and promote the use of measurement to improve software processes and products for both businesses and governmental organizations. Release 12 of this repository includes information on over 6,000 software projects [1].

Researchers using the ISBSG repository in multivariable statistical analysis face a number of challenges, including the following:

- outliers in some of the numerical data fields, and
- numerous values missing for a significant number of variables.

This makes using the repository for research purposes quite challenging when a large subset of data fields must be analyzed concurrently as parameters in statistical analysis. In addition, since the data are contributed voluntarily, their quality varies, and this should be taken into account prior to statistical analysis.

With conventional statistical methods, all the variables in a specified model are presumed to have been collected and made available for all cases. The default action for virtually all statistical tools is simply to delete cases with any missing data on the variables of interest, a method known as *listwise deletion* or *complete case analysis*. As well, missing values are often ignored for convenience. While this simple treatment might be acceptable with a large dataset and a relatively small amount of missing data, biased findings can result if the percentage of missing data is significant, as information on the incomplete cases will have been lost. With relatively small datasets, it is poor practice to merely ignore missing values or to delete incomplete observations in these situations. More reliable imputation methods must be found, in order to ensure that the analyses in which they are used are meaningful. The most obvious drawback in listwise deletion is that it often removes a large fraction of the sample, which results in a serious loss of statistical power. Awareness of the importance of treating missing data in appropriate ways during analysis has been growing [2], and consequently techniques for dealing with missing multivariate data have been proposed, including the use of the multiple imputation (MI) technique [3].

This paper investigates the use of MI to deal with missing values in the ISBSG repository, and also considers the implications of the presence of outliers in numerical data fields. The paper raises a number of data quality issues associated with the ISBSG repository, and proposes a number of criteria and techniques for preprocessing the data in order to improve the quality of the samples identified for detailed statistical analysis, as well as presenting a multiple imputation (MI) strategy for dealing with missing values.

Related Work

A number of researchers have used the ISBSG repository for research purposes, but only a few have examined techniques designed to tackle the data-related issues that often arise in large multi-organizational repositories of software engineering data, such as the quality of the data, the presence of statistical outliers, and the problem of missing values. This section presents related work on these issues by those who have used the ISBSG repository in their research.

An approach is presented for building size-effort models based on the programming languages used [4]. The authors provide a description of the data preparation filtering method they applied to identify these languages, and use only relevant data in their analysis. From the 789 records in ISBSG Release 6, they removed records with very small project effort and those for which no data on the programming language were available. They then removed records for programming languages with too few observations to form adequate programming language samples, which left them with 371 relevant records for their analyses. Finally, they built estimation models for every programming language with a sample size of over 20 projects, and analyzed those samples, excluding 72 additional outliers for undisclosed reasons.

A quality rating filter is applied to investigate the links between team size and software size, and development effort [5]. Then, the authors removed records for which software size, team size, or work

***Corresponding author:** Alain Abran, École de Technologie Supérieure (ÉTS)-University of Québec, Montréal, Québec, Canada, Tel:+1 (514) 396; E-mail: alain.abran@etsmtl.ca

Received February 07, 2016; **Accepted** February 18, 2016; **Published** February 29, 2016

Citation: Bala A, Abran A (2016) Use of the Multiple Imputation Strategy to Deal with Missing Data in the ISBSG Repository. J Inform Tech Softw Eng 6: 171. doi:10.4172/2165-7866.1000171

effort values were missing. This procedure led to a reduction in the size of the original set of 1,238 project records (ISBSG Release 7) to 540 for investigation purposes.

Only projects with a data quality rating of A and B are used for the analysis of Release 8. They then applied additional filters for the software sizing method, as well as for development type, effort recording, and the availability of all the function point counting components (i.e. unadjusted function point components and 14 general system characteristics) [6]. This reduced the original collection of 2,027 records to a set of 184 records for further processing.

The issues of data quality and completeness in the ISBSG repository are discussed [7]. The authors describe the process they used in attempting to maximize the amount of data retained for modeling software development effort at the project level. For instance, in their study, they retained the projects that had been sized using IFPUG/NESMA Function Point Analysis (FPA), arriving at a dataset comprising 2,862 projects (out of 3,024 in Release 9), but with and without considering other quality criteria.

ISBSG Release 9 is used to investigate and report on the consistency in the effort data field during each development activity. A number of other major issues in data collection and analysis are identified, including the following: more than one field referring to the same type of information, and fields that contradict one another, both of which lead to inconsistencies. In this case, data analysts must either make an assumption on which field is the correct one, or drop the projects containing contradictory information. In the study reported by Déry [8], data missing in many fields led to much smaller usable samples with less statistical scope for analysis, making extrapolation, when desirable, a challenge. The authors do not treat the missing values across activities directly within the dataset, but indirectly by inference from average values within subsets of data containing similar activity groupings without missing values.

ISBSG Release 10 is used to analyze the relationship between software size and development effort [9]. For these authors, data preparation involves only software functional size in IFPUG/NESMA function points and effort in total hours, with no additional filtering. Consequently, a large proportion of the available records were retained for modeling purposes in the case they described – 3,433 out of 4,106 projects, but without considering other quality criteria.

To summarize, the data preparation techniques proposed in these studies are defined mostly in a heuristic manner. Their authors describe the techniques in their own terms and using their own structure, without applying any common practices involving the description and documentation of the requirements for pre-processing the ISBSG raw data prior to detailed data analysis.

A summary of these works is presented in Figure 1, including the ISBSG release used, the number of projects in the release, whether or not the issue of missing values was addressed, and, finally, whether or not statistical outliers were observed, and either removed or excluded from further analysis (Table 1).

Multiple Imputation Technique (MI) for Handling Missing Values

Overview of the multiple imputation process

In MI, each missing value is replaced with a pointer to a vector of ***m*** values taken from ***m*** possible scenarios or imputation procedures based either on the information observed, or on historical or subsequent analyses.

This is an attractive solution to missing data problems, as it provides a good balance between the quality of the results and ease of use. The performance of MI in a variety of missing data situations has been studied by Graham, et al. and Schafer, et al. [10,11] but not with software engineering datasets specifically. The technique has been shown to produce parameter estimates that reflect the uncertainty associated with estimating missing data. Furthermore, it has been shown to provide satisfactory results in the case of a small sample size or a high rate of missing data [12].

MI does not attempt to estimate each missing value by simulating it, but rather by representing a random sample of the missing values. This process results in valid statistical inferences that properly reflect the uncertainty that the missing values generate.

It has the advantage of using complete data methodologies for the analysis, as well as the ability to incorporate the data collector's knowledge [3]. Three steps are needed to implement MI - (Figure 1):

Create the imputed datasets: The first step is to create values (also referred to as *imputes*) to be substituted for the missing data. In order to achieve this, an imputation procedure must be identified that will allow imputes to be created based on the values found across the dataset for the same variable in the dataset. This involves the creation of imputed datasets, which are plausible representations of the data: the missing data are filled in ***m*** times to generate ***m*** complete datasets.

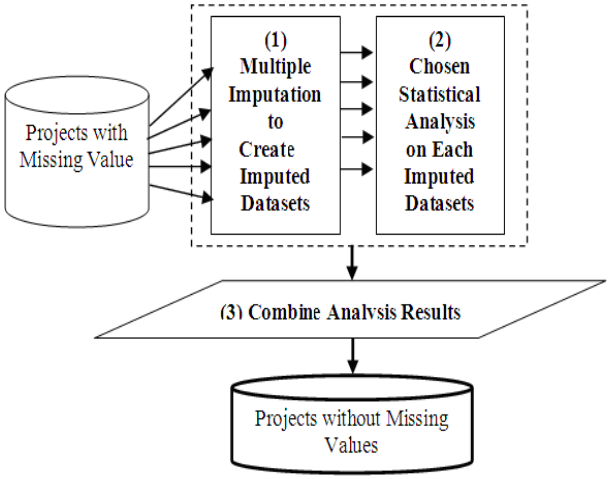


Figure 1: Multiple imputation process.

Paper	ISBSG Release	Number of projects in the release	Missing values	Outliers
Abran et al. [4]	6	789	Observed and investigated	Observed and removed
Pendharkar et al. [5]	7	1,238	Observed and removed	Undetermined
Xia et al. [6]	8	2,027	Observed and removed	Undetermined
Deng and MacDonell [7]	9	3,024	Removed	Undetermined
Dery and Abran [8]	9	3,024	Removed	Observed and removed
Jiang et al. [9]	10	4,106	Observed and investigated	Observed and removed

Table 1: Summary of ISBSG studies dealing with missing values and statistical outliers.

Analyze the imputed datasets: Note that standard statistical analysis is conducted separately for each imputed dataset. This analysis proceeds as if there were no missing data, except that it is performed on each imputed dataset. In other words, ***m*** complete datasets are analyzed using standard statistical procedures.

Combine the analysis results: Once the analyses have been completed for each imputed dataset, all that remains is to combine these analyses to produce one overall set of estimates. The results from the analysis of the ***m*** complete datasets are combined to produce inferential results once the imputed datasets have been created [3].

Applying MI in SAS software

SAS software is a comprehensive statistical software system that integrates utilities for storing, modifying, analyzing, and graphing data. Most SAS statistical procedures exclude observations with any missing variable values from the analysis. These observations are called *incomplete cases* [13].

The SAS MI procedure consists of three steps (PROC MI, PROC REG, and PROC MIANALYZE) for creating imputed datasets that can be analyzed using standard procedures. The specifics of this SAS MI procedure, in which multiple imputed datasets are created for incomplete *p*-dimensional multivariate data, are the following, (Figure 1).

- Appropriate variability is incorporated across the ***m*** imputations (in PROC MI).
- The multiple imputed datasets are analyzed using regression procedures (in PROC REG).
- Once the ***m*** complete datasets have been analyzed using standard statistical procedures, PROC MIANALYZE is applied to generate valid statistical inferences about these parameters by combining the results from the ***m*** complete datasets.

Data Preparation on ISBSG R9

In order to determine the impact of missing values, we use the data fields studied and the data preparation reported by Déry et al. [8] to evaluate the results of the MI technique for dealing with missing data. In accordance with their methodology, we present below the data preparation process for the effort variable with missing values by development activity using the ISBSG repository.

Data preparation 1

Prior to analyzing the data preparation process using the ISBSG repository, it is important to understand how fields are defined, used, and recorded. Déry [8], reported ISBSG repository R9 is used for the analysis reported in this paper, which contains 3,024 projects. In preparing the samples from this ISBSG dataset, two verification steps must be performed: data quality verification, and data completeness verification. The variables that may potentially have an impact on project effort are selected in this paper using the same criteria as explained by Déry [8] for preprocessing the data – see Table 2:

- The first two variables deal with software size measured in Function Points (FP), and the functional sizing method selected for this study is the IFPUG standard – ISO 20926.
- The next six variables are associated with the total project effort in hours (i.e. Summary Work Effort), as well as the project effort in each of the ISBSG-defined project activities (i.e. Plan, Specify, Build, Test, and Implement).
- As not all the projects in the ISBSG repository were sized using the same functional sizing method, only the 2,718 projects sized with the IFPUG method were retained for the analyses reported here.
- After filtering for data quality (A and B), the number of projects was reduced to 2,562, prior to identifying the missing values in the fields of interest (Figure 2).

Data preparation 2: Effort by project activity

The 2,562 projects selected in the previous section come from many organizations, each with its own effort recording standard. For instance, some organizations include the effort for all ISBSG-identified project activities, while others may not include the planning activity in their project effort reporting, and still others might not include the implementation activity. The ISBSG data collection form contains fields for recording information on the project activities, and other fields for recording the effort for each project activity, but none of these fields is mandatory. Therefore, of the 2,562 projects, only 847 have activity tags, and only 325 of these have detailed effort recorded by project activity – (Table 3), columns 2 and 3.

We now present the data preparation process reported by Déry [8], in which the information required for the analysis of the distribution of effort data across the development activities is identified.

In order to use the data for statistical analysis, at least two requirements must be met:

- There must be enough historical data.
- The data must be homogeneous enough to provide meaningful interpretations.

Table 3, illustrates the detailed effort by activity, along with the total project effort recorded in the ISBSG data repository, and, by corollary, the corresponding number of fields with missing values [8].

The numbers in the rows in Table 3 refer to the number of projects. The labels in the left-most column comprise the set of 1st letters of each

Data variable	Abbreviation	Units	Min in R9	Max in R9
1- Functional Size	FP	Function Points	0	2,929
2- Functional Sizing Method	IFPUG	-	-	-
3- Summary Work Effort	Effort	Hours	170	100,529
4- Effort in the Planning Activity	P	Hours	2	5,390
5- Effort in the Specify Activity	S	Hours	1	28,665
6- Effort in the Build Activity	B	Hours	30	48,574
7- Effort in the Test Activity	T	Hours	14	15,005
8- Effort in the Implement Activity	I	Hours	20	8,285

Table 2: ISBSG data fields used in this study.

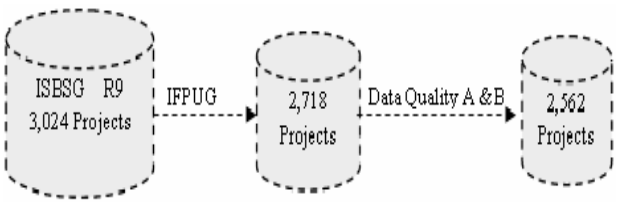


Figure 2: Preparation of the ISBSG R9 dataset.

activity[1] included in the project effort reported:

- The label PSBTI refers to the projects with an effort tag for each of the five project activities: Planning, Specification, Build, Test, and Implementation.
- The label PSBT refers to the projects with an effort tag for each of the following four project activities: Planning, Specification, Build, and Test (i.e. without any data on the implementation activity.)
- The label SBTI corresponds to the projects with an effort tag for each of the following four project activities: Specification, Build, Test, and Implementation (but without any data on the planning activity.)

However, of the 847 projects with activity tags (Table 3, column 2), only 325 have detailed effort by project activity concurrently (Table 3, column 3). Since only projects with effort data recorded by project activity have the detailed effort data by project activity required for the purposes of this research paper, this significantly reduces the size of the samples available for detailed analysis: for instance, for the PSBTI activity, of the 350 projects in this effort profile (Table 3, column 2, line 1), only 113 have detailed effort data by activity (Table 3, column 3, line 1).

Verification of the consistency of the detailed effort by activity with the total project effort recorded leads to only 76 projects that meet this consistency criterion for our analysis purposes (Table 3, column 4, line 1). In addition, 35 projects (Table 3, column 7, line 1) have to be deleted because of other inconsistencies in the data, such as the following:

- The project with the greatest amount of effort did not have the mandatory size in function points, which points to a lack of quality control of the data recorded for this project.
- There is an unusual effort pattern in some of the projects: in 34 of them, 98% of the effort was recorded, on average, in the specification activity, and less than 1% in each of the other 4 activities, pointing to a problem in the data collection process. Of course, for the purposes of our analysis, these projects must also be discarded.

Using the same data preparation criteria, the final count of projects in the sample of projects with the PSBT activity profile (Table 3, line 2) is 100 (Table 3, column 4, line 2), 38 of which had to be dropped from further analysis because of inconsistencies between detailed effort levels by activity and total effort.

Identification of outliers

Outliers are defined as observations in a dataset that appear to be inconsistent relative to those in the remainder of the dataset. The identification of outliers is often considered as a means to eliminate observations from a dataset in order to avoid undue disturbances in future analysis [14,15]. For this reason, appropriate methods for detecting them are needed.

Outlier identification is first and foremost a means to verify the relevance of the values of the input data: candidate outliers would typically be at least 1 or 2 orders of magnitude larger than the data point closest to these points, and so a graphical representation can be used to identify them. Statisticians have devised several ways to achieve this. The Grubbs test and the Kolmogorov-Smirnov test can be used to determine whether or not a variable in a sample has a normal distribution, and so verify whether or not that data point is a true statistical outlier [16]. These tests comprise an ESD (Extreme Studentized Deviate) method, in which the studentized values measure how many standard deviations each value is from the sample mean:

- When the P-value for the Grubbs test is less than 0.05, that value is a significant outlier at the 5.0% significance level.
- Values with a modified Z-score greater than 3.5 in absolute value may well be outliers.
- The Kolmogorov-Smirnov test is used to give a significant P-value (high value), which means that we can assume that the variable is distributed normally.

The Grubbs test is particularly easy to perform. The first step is to quantify how far the outlier is from the other values by calculating the Z ratio as the difference between the outlier and the mean divided by the standard deviation (SD). If Z is large, then the value is far from the other values.

After calculating the mean and standard deviation of all the values, including the outlier, this test calculates a P-value only for the value furthest from the rest. Unlike some of the other outlier tests, this test asks only whether or not that one value is an outlier. If it is, the outlier is removed and the test is run again.

Table 4 presents the overall results of the Grubbs test with the set of data N=103 projects with valid data (Table 3, column 5), and Table 5 presents the 2 significant outliers that will be removed from further statistical analysis.

The outlier tests were performed on the Functional Size and Summary Work Effort variables. The figures in the "Test no." column in Table 4 represent the number of iterations for the application of the Grubbs test for identifying the outliers, one at a time. The details of the 3 outliers identified by the Grubbs test are presented in Table 6.

Multiple Imputation Technique applied on ISBSG R9

This section presents an application of the three distinct **steps**of the MI statistical inferences on the ISBSG repository (Release 9)**.** This section is structured as follows:

- Section 5.1 presents Step 1: creating the imputed datasets.
- Section 5.2 presents Step 2: analyzing the imputed datasets.
- Section 5.3 presents Step 3: combining the analysis results.

[1]In the ISBSG repository prior to R5,design was not included as a development activity: a high level design activity had previously been included in the Specificationactivity and a low level design activity in the Build activity.

	Number of Projects					
Project activities included (1)	With activity tags (2)	With detailed effort by activity (3)	All activity effort consistent with Summary Effort (4)	Projects with valid data (5)	Projects with missing values (6)	Data with some inconsistencies (7)
PSBTI	350	113	76	41	0	35
PSBT	405	200	100	62	62	38
SBTI	92	12	3	3	3	0
Total	847	325	179	106	65	73

Table 3: Detailed effort by activity in the ISBSG.

Test no.	Mean Total Effort	SD	No. of values	Outlier detected?	Significance level	Critical value of Z
1	5726	11032	106	Yes	0.05 (two-sided)	3.40
2	4823	5970	105	Yes	0.05 (two-sided)	3.40
3	4460	4692	104	Yes	0.05 (two-sided)	3.40
4	4173	3686	103	No	0.05 (two-sided)	3.39

Table 4: Descriptive Statistics for the Grubbs test on Total Effort (N=106).

Test no.	Total Effort of the candidate outlier	Z	Significant outlier?
1	100529	8.59	Yes. P < 0.05
2	42574	6.32	Yes. P < 0.05
3	34023	6.30	Yes. P < 0.05
4	15165	2.98	No, although furthest from the rest (P > 0.05).

Table 5: Outlier analysis using the Grubbs test on Total Effort.

No. of outliers	Functional Size	Summary Work Effort	Effort Plan	Effort Build	Effort Test	Effort Specify	Effort Implement
1	(0)	34023	1190	9793	17167	4489	1384
2	781	42574	5390	7910	15078	14196	(0)
3	2152	100529	(0)	28665	48574	15005	8285

Table 6: Description of the 3 outliers deleted.

Step 1: Creating the Imputed Datasets (Imputation)

In this step, the missing values from the ISBSG R9 are imputed with a PBSTI profile: random numbers are generated to provide the values that are missing from the selected data fields, that is:

- The Effort Implementation (EI) activity, and
- The Effort Planning (EP) activity.

The SAS software procedure PROC MI is used to generate 5 'completed' datasets[2] for the repository. The random numbers are imputed data based on the 'seed' values inserted manually to generate random numbers. The details of this step are presented in 5.1.1, and the analysis of variances in 5.1.2.

Effort profile following MI based on the seeds with the full sample of 106 projects: The seed values selected for the full sample of 106 projects are set to the minimum and maximum values in hours for the two corresponding fields (EI and EP) of the PBSTI profile that does not have missing value in R9, that is, the Effort Plan and Effort Implementation for the 41 projects with the PBSTI profile. Here, the minimum values for the Plan and Implement activities are (2, and 20) hours, and the maximum values are (5,390, and 8,285) hours – see the two rightmost columns in Table 2.

This leads to the following vectors of parameters for this imputation step: the vector of minimum values for the missing values of the Plan and Implement activity sets to be generated is (2, and 20 hours), and the vector of the maximum values is (5,390, and 8,285 hours) – (Table 2).

The positions in the vector correspond to the order that appears in the (var) statement in the SAS procedure. In the dataset used in this research, the variables *min* and *max* are based on each variable that is entered in the procedure.

Figure 3 displays the outcome of Imputation 1, which generated effort data for the 65 projects (out of the 106 projects) with missing values:

- The 62 projects with missing effort values in the Implement activity, and
- The 3 projects with missing effort values in the Plan activity– see the shaded areas in Figure 2.

For the first imputation, which involves the 65 projects with missing values, the imputation only occurs in the column with missing values.

Analysis of variance information and parameter estimates for the Implement effort and Plan effort estimation model following MI: This section presents the output results of the variance information and parameter estimates for MI based on 106 or 103 projects, before and after removal of the outliers respectively. These are used to generate valid statistical inferences about the dependent variables (Effort Plan and Effort Implement).

In addition, the MI parameter estimates will show the estimated mean of the 5 imputed datasets, which represent the mean of 5 imputations and the standard error of the mean for Effort Implement and Effort Plan estimation. The tables also display a 95% mean confidence interval and a t-test with the associated P-value, and are inferences based on the t-distribution. All that remains is to combine these analyses to produce one overall set of estimates.

The variance information is analyzed by identifying the differences within datasets(variances measure uncertainty due to missing data) and between datasets (variances measure additional uncertainty due to imputation), as follows [3]:

A. Estimate the parameter ($\hat{P}_j$), which is the mean across the *m* imputations.

The mean of $\hat{P}_j$ is then given by $\overline{P} = \sum_{j=1}^{m} \hat{P}_j$

B. Analyze the variances (within and between):

Within: the imputation variance $\overline{U}$ of the parameter $\overline{P}$ is the mean of the variances across the *m* imputations.

Between: the imputation variance B of the parameter $\overline{P}$ is the standard deviation of $\overline{P}$ across the *m* imputations.

The total variance of $\overline{P}$ is a function of $\overline{U}$ and B and is used to calculate the standard error used for test statistics.

The variability of $\hat{P}_j$ is divided into two components:

Within imputation variance $\overline{U}m = \frac{1}{m}\sum_{j=1}^{m} U_j$

Between imputation variances $Bm = \frac{1}{m\text{-}1}\sum_{j}^{m}(\hat{P}_j - \overline{P}m)^2$

Total variance $Tm = \overline{U}m + (1+\frac{1}{m})Bm$

C. Combine the Standard Error results:

Variance of $\overline{P}m$:

Var($\overline{P}m$)= $Tm = \overline{U}m + (1+\frac{1}{m})Bm$

$\overline{U}$ = Average of the 'within' variances

m= Correction for a finite number of imputations m

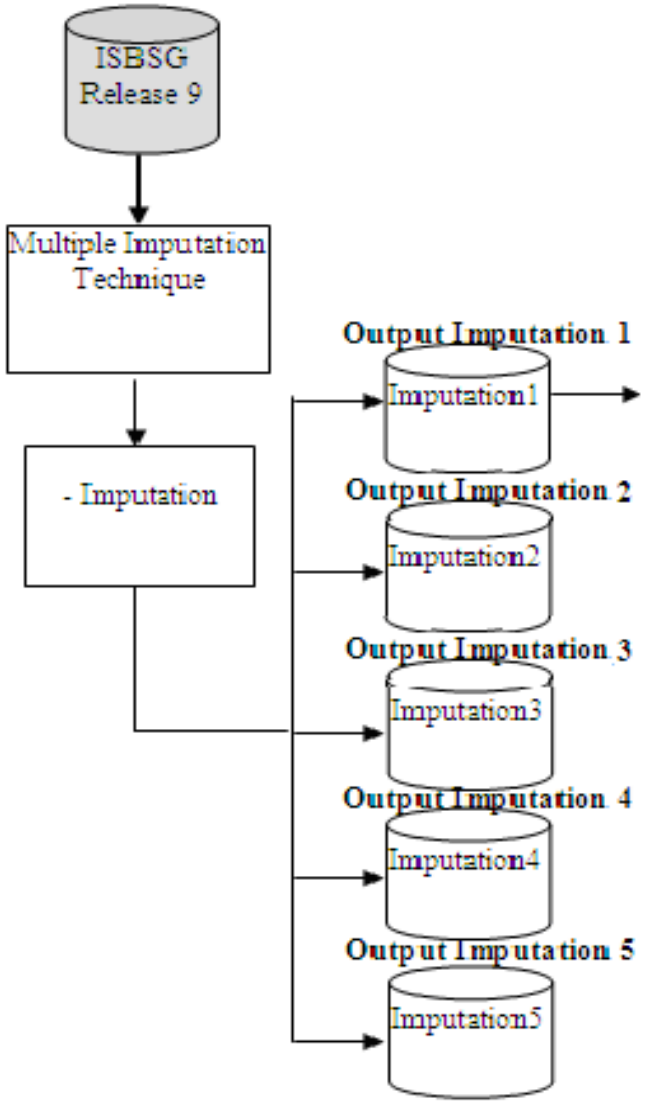
Bm = Variation in the ***m***results; Variance of the ***m***different parameters

α) Standard error (SE):

SE ($\overline{U}m$) = $\sqrt{Tm}$

Tables **7** and **8** display the variances between imputations (Bm) and within imputations $\overline{U}m$, and the total variances when combining completed data inferences respectively, after the completion of ***m*** imputations.

[2]By default, SAS creates 5 imputed datasets.



Imputation 1

Input	Plan	Specify	Build	Test	Implement
1	635	2086	479	7150	972
2	234	296	624	1124	1196
3	242	492	279	452	513
4	257	534	1292	2767	195
5	7	50	591	44	94
6	936	1248	4680	624	1130
7	468	1216	2714	2254	542
8	210	503	4790	1618	469
9	234	312	624	156	695
10	16	62	281	39	471
11	437	1092	1388	936	589
12	390	2605	2527	1824	472
13	52	1	286	107	141
14	776	70	1085	280	691
15	455	140	560	140	68
16	560	278	1274	552	436
17	20	20	815	21	242
18	632	1053	7036	4586	461
19	42	47	545	145	44
20	50	20	235	840	778
21	66	70	662	352	47
22	624	2964	2508	7800	869
23	78	390	117	156	458
24	16	16	311	31	432
25	35	70	150	50	72
26	468	2964	2028	3822	917
27	525	417	1195	867	372
28	59	9	256	114	121
29	196	10	54	20	92
30	160	120	470	280	46
31	1270	237	1165	945	1019
32	546	2028	5120	2574	903
33	156	936	624	499	337
34	554	582	1626	3542	117
35	504	1267	1624	2996	634
36	3915	584	1760	4340	2495
37	484	874	2604	1170	330
38	325	440	1369	977	337
39	77	45	354	149	230
40	2094	760	2883	5517	1424
41	60	100	910	250	303
42	95	15	294	615	272
43	554	635	755	2295	391
44	50	40	510	100	100
45	663	1635	1482	1650	75
46	1538	1844	1760	3935	1885
47	468	2496	2184	3744	1131
48	657	998	5889	1217	555
49	767	491	1807	1652	526
50	359	858	2343	234	457
51	468	1833	1560	1594	634
52	480	1746	1747	2329	555
53	513	1125	1212	5466	614
54	671	1326	1201	1607	248
55	545	658	1374	969	595
56	172	108	692	162	495
57	825	1652	2207	1052	1205
58	5590	7910	15078	14196	4584
59	78	47	390	47	305
60	140	212	475	14	64
61	624	1872	3276	5413	52
62	170	264	30	295	246
63	5085	23665	45574	15005	5135
64	377	168	2080	976	144
65	184	207	836	111	56

Figure 3: Sample result of the Multiple Imputation method – Step 1.

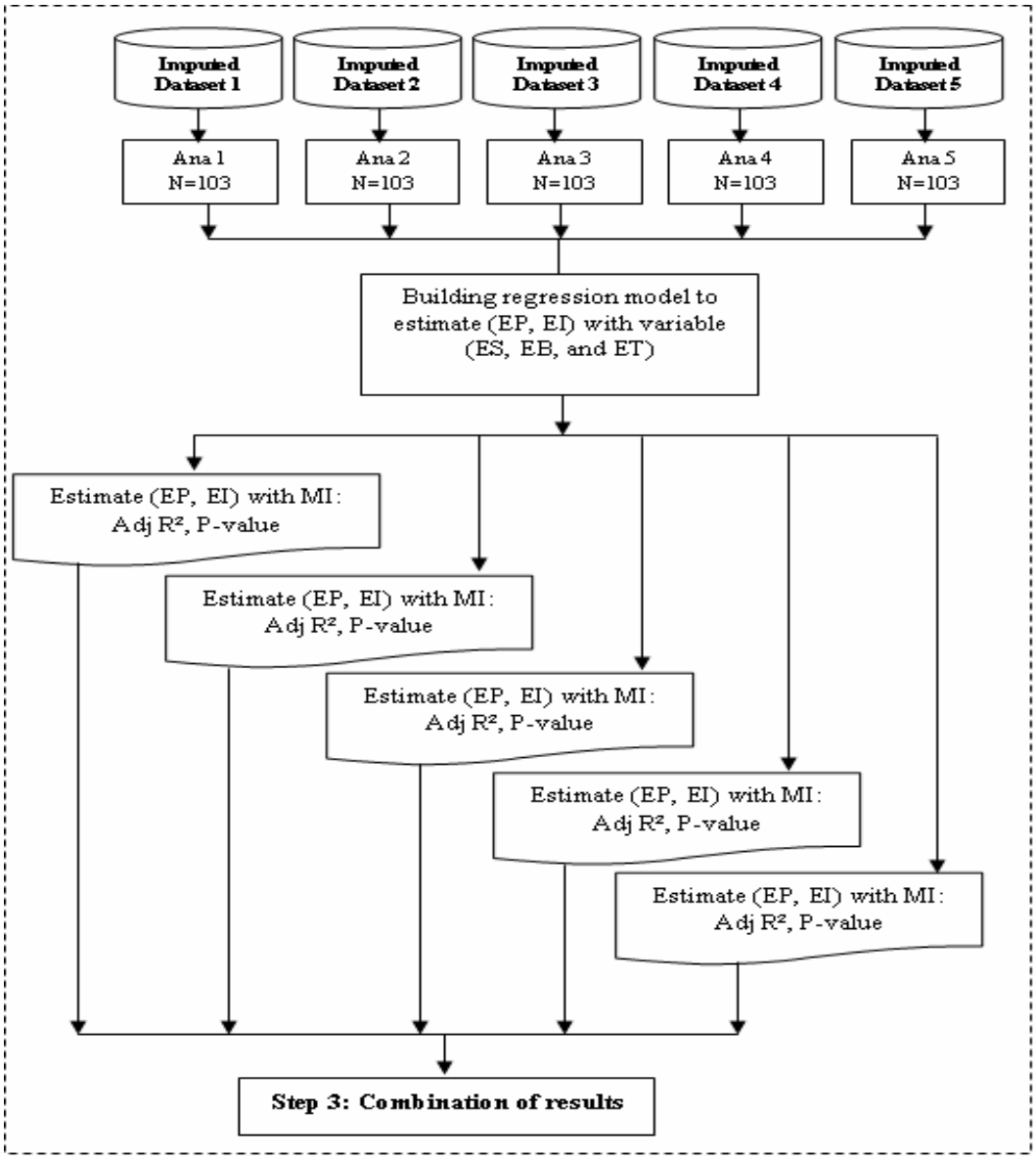


Figure 4: Building the regression analysis estimation models.

For instance, for the 5 imputed datasets with 106 projects in Table 7, the combined results of the Effort Implementation (EI) variable give a Mean of $\overline{Pm}$ = 541 hrs, a variance within imputations of $\overline{Um}$ = 8,455 hrs, a variance between imputations of Bm = 2,144 hrs, and M = 5 imputations.

Total variance Tm is = 8,455 + 1.2*2,144 = 11,028 hrs, and the SE result is = $\sqrt{11028}$ = 105 hrs.

Considering that the P-values in Tables 7 and 8 are all <0.1, we can conclude that, when the outliers are taken out upfront, the variance results of the standard error of the estimates have decreased from 106 hours to 60 hours for the Effort Planmodel and from 105 hours to 73 hours for the Effort Implement model. As well, the results are statistically significant at t-test and P-values with and without outliers for the Effort Plan and Effort Implement estimates (Table 9).

Analysis of average effort after MI based on seeds selected, excluding outliers: Tables 10-14 present, in parentheses, the averages of the values imputed based on seeds selected within the ranges of values that exclude outliers, that is, for Effort Plan in the SBTI profile and Effort Implement in the PSBT profile. In summary, in Table 15, the averages of the 5 imputations are as follows:

Imputation 1: Effort Plan = 20.8% for the SBTI profile & Effort Implement = 10.5% for the PSBT profile.

Imputation 2: Effort Plan = 12.4% for the SBTI profile & Effort Implement = 7.9% for the PSBT profile.

Imputation 3: Effort Plan = 20.4% for the SBTI profile & Effort Implement = 7.2% for the PSBT profile.

Imputation 4: Effort Plan = 6.7% for the SBTI profile & Effort Implement = 11.3% for the PSBT profile.

Imputation 5: Effort Plan = 19.9% for the SBTI profile & Effort Implement = 10.3% for the PSBT profile.

Variable	N=106 Projects, before removal of outliers								
	Mean $\bar{P}m$	Std Error	95% Confidence Limits		T- test	Variance			P-Value
						Between Bm	Within $\bar{U}m$	Total	
EP	573 hrs	106 hrs	364 hrs	783 hrs	5.42	99	11066	11184	<.0001
EI	541 hrs	105 hrs	328 hrs	753 hrs	5.15	2144	8455	11028	<.0001

Table 7: Variance information for parameter estimates of Effort Plan and Effort Implement (N=106 projects, before removal of outliers).

Variable	N=103 Projects, after removal of 2 outliers								
	Mean $\bar{P}m$	Std Error	95% Confidence Limits		T-test	Variance			P-Value
						Between Bm	Within $\bar{U}m$	Total	
EP	448 hrs	60 hrs	330 hrs	567 hrs	7.50	15	3562	3598	<.0001
EI	395 hrs	73 hrs	221 hrs	569 hrs	5.38	3030	1747	5383	<.0001

Table 8: Variance information for parameter estimates of Effort Plan and Effort Implement (N=103 projects, after removal of 3 outliers).

Variable	Before removal of outliers N=106 projects		After removal of 3 outliers N=103 projects	
	Significant T-test	Significant P-values	Significant T-test	Significant P-values
EP	Yes	Yes	Yes	Yes
EI	Yes	Yes	Yes	Yes

Table 9: Summary of parameter estimates for Effort Implement with and without outliers.

Profile	Project activity – % Effort					No. of projects
	Effort Plan	Effort Specify	Effort Build	Effort Test	Effort Implement	
PSBTI	10.3	23.9	36.8	21.1	8.0	40
PSBT	9.8	16.3	30.9	32.5	(10.5)	61
SBTI	(20.8)	6.5	50.5	18.7	3.4	2

Table 10: Average effort distribution by profile (1st imputation), N=103 projects, excluding outliers.

Profile	Project activity – % Effort					No. of projects
	Effort Plan	Effort Specify	Effort Build	Effort Test	Effort Implement	
PSBTI	10.3	23.9	36.8	21.1	8.0	40
PSBT	10.1	16.8	31.8	33.5	(7.9)	61
SBTI	(12.4)	7.1	55.9	20.7	3.8	2

Table 11: Average effort distribution by profile (2nd Imputation), N=103 projects, excluding outliers.

Profile	Project activity – % Effort					No. of projects
	Effort Plan	Effort Specify	Effort Build	Effort Test	Effort Implement	
PSBTI	10.3	23.9	36.8	21.1	8.0	40
PSBT	10.1	16.9	32.0	33.7	(7.2)	61
SBTI	(20.4)	6.5	50.8	18.8	3.5	2

Table 12: Average effort distribution by profile (3rd imputation), N=103 projects, excluding outliers.

Profile	Project activity – % Effort					No. of projects
	Effort Plan	Effort Specify	Effort Build	Effort Test	Effort Implement	
PSBTI	10.3	23.9	36.8	21.1	8.0	40
PSBT	9.7	16.1	30.6	32.2	(11.3)	61
SBTI	(6.7)	7.6	59.6	22.1	4.1	2

Table 13: Average effort distribution by profile (4th imputation), N=103 projects, excluding outliers.

Table 16 combines, for each imputation round, the data from all the projects, including the 40 in the PBSTI profile which already had all the data and the 61 projects in the PSBT and 2 projects in the SBTI profiles which had missing data in one activity. Some variations can, of course, be observed across the 5 imputation steps: for instance, the distribution of effort in the 'Implement' activity varies from 7.5% to 10.1%, with an average of 8.9 across all 5 imputations – (Table 16).

Step 2: Analyzing the completed datasets

Analysis strategy: Once the MI techniques have replaced missing values with multiple sets of simulated values to complete the data, the regression analysis procedure PROC REG is used with each completed dataset to obtain estimates and standard errors, which adjusts the parameter estimates obtained from PROC MI for missing data.

In this step, the results of the regression analysis estimation models for the imputed values after removing the outliers are presented, this time trained with the 5 imputed datasets and 63 projects excluding outliers.

The objective in using this procedure is to obtain an analysis of the imputed dataset based on linear regression models, that is:

to estimate the dependent variables with the missing values (i.e. Effort Plan and Effort Implement),

on the basis of the independent variables (i.e. Effort Specify, Effort Build, Effort Test) that have observed values.

For the evaluation of the accuracy performances of the estimation models, this section presents the percentage of variation in the dependent variable explained by the independent variables of the model using the adjusted R^2 that accounts for the number of independent variables in the regression model.

Figure 4 illustrates how to build the regression analysis estimation models and obtain the analysis results to use them (in Step 3).

Step 2 is as follows:

- Use each completed dataset from Step 1;
- Execute PROC REG;
- Build an estimation regression model for each completed dataset from MI;
- Obtain an analysis of the imputed dataset based on linear regression models;
- Combine the analysis results obtained in this step for use in Step 3.

Implement the effort estimation model (using the 61 imputed Implement values): To build an estimation model of the Implement effort, a multiple regression analysis is performed using:

A) the dependent variable, Effort Implement, using:

Profile	Project activity – % Effort					No. of projects
	Effort Plan	Effort Specify	Effort Build	Effort Test	Effort Implement	
PSBTI	10.3	23.9	36.8	21.1	8.0	40
PSBT	9.8	16.3	30.9	32.6	(10.3)	61
SBTI	(19.9)	6.5	51.1	18.9	3.5	2

Table 14: Average effort distribution by profile (5th Imputation), N=103 projects, excluding outliers.

# Imputation	%Effort Plan in SBTI profile	%Effort Implement in PSBT profile
1st Imputation	20.8%	10.5%
2nd Imputation	12.7%	7.9%
3rd Imputation	20.4%	7.2%
4th Imputation	6.7%	11.3%
5th Imputation	19.9%	10.3%

Table 15: Comparison across the imputations without outliers (N=103 projects).

Imputation No.	Project activity – % of total Effort					Total
	Effort Plan	Effort Specify	Effort Build	Effort Test	Effort Implement	
1st Imputation	10.1	18.9	33.2	28.2	9.5	100%
2nd Imputation	10.2	19.3	33.9	28.8	7.9	100%
3rd Imputation	10.3	19.3	34.0	28.9	7.5	100%
4th Imputation	9.9	18.8	33.1	28.1	10.1	100%
5th Imputation	10.1	18.9	33.3	28.3	9.4	100%
Average of the 5 imputations	10.1	19.0	33.5	28.5	8.9	100%

Table 16: Profiles of average effort distribution for N=103 projects, excluding outliers.

- the *actual* Implement effort of the 40 projects from the PSBTI profile,
- the *imputed* Implement effort of the 61 projects from the PSBT profile,
- the *actual* Implement effort of the 2 projects from the SBTI profile;

B) the independent variables Effort Specify, Effort Build, and Effort Test.

For instance, in Table 17, the parameter estimates for the Effort Implement model in the first line are: 170, 0.01, 0.10, and 0.08. Therefore, the regression equation for predicting the dependent variable from the independent variables in the first imputation is:

Effort Implement = 170 hours + 0.01 x Effort Specify + 0.10 x Effort Build + 0.08 x Effort Test.

Table 17 also shows the coefficients of determination (i.e. R^2 and Adjusted R^2) for the regression model for each imputation. For instance, for the Model of Effort Implement, the adjusted R^2 obtained for each of the five imputations without outliers are (0.28, 0.09, 0.14, 0.35, and 0.39). Moreover, the regression analysis results for the estimation models present a statistically significant P-value in each of the 5 imputations of <0.0001.

Plan effort estimation models (built using the 2 imputed Plan values): To build an estimation model of the Plan effort, a multiple regression analysis is performed using:

A) the dependent variable, Plan effort, using:

- The *actual* Plan effort on the 40 projects for the PSBTI profile,
- The *actual* Plan effort of the 61 projects from the PSBT profile,
- The *imputed* Plan effort of the 2 projects from the SBTI profile;

B) the independent variables Specify effort, Build effort, and Test effort.

Table 18 presents the results of the estimation models for the dependent variable (Effort Plan) trained with the independent variables (Effort Specify, Effort Build, and Effort Test) for each of the five imputations and based on 103 projects (without outliers).

For instance, in Table 18, the parameter estimates for the Effort Plan model in the first line are (86, -0.09, 0.17, and 0.14), and the regression equation for predicting the dependent variable from the independent variables is:

Effort Plan= 86 hours - 0.09 x Effort Specify + 0.17 x Effort Build + 0.14 x Effort Test.

Table 18 also shows the coefficients of determination (i.e. R^2 and Adjusted R^2) for the regression model for each imputation. For instance, in Table 18, for the model of Effort Plan,the adjusted R^2 obtained for the five imputations without outliers are (0.33, 0.34, 0.34, 0.34, and 0.33) respectively. Moreover, the regression analysis results for the estimation models present a statistically significant P-value in each of the 5 imputations of <0.0001.

Step 3: Combining the inferences from the imputed datasets (combination of results)

Strategy and statistical tests used: Step 3 presents the results of the parameter estimates for the Effort Implement and Effort Plan estimation models previously trained on the full dataset with imputed values and N=103 projects after removing the outliers. In this step, the results of the regression analysis estimation in Step 2 are combined, taking into account differences within datasets (variation due to the missing data) and between datasets (variation due to imputation).

The MI regression analysis procedure (PROC MIANALYZE)

Imputation no.	N=103 Projects, without outliers						
	Effort Implement Model, N=61						
	Intercept	Effort Specify	Effort Build	Effort Test	Adjusted R^2	R^2	P-value
1	170	0.01	0.10	0.08	0.28	0.30	<0.0001
2	189	-0.002	0.08	0.03	0.09	0.11	<0.0001
3	194	0.07	0.09	-0.04	0.14	0.17	<0.0001
4	168	0.0004	0.06	0.16	0.35	0.37	<0.0001
5	138	-0.008	0.09	0.12	0.39	0.41	<0.0001

(N=103 projects, without outliers).

Table 17: Regression analysis estimation model for Effort Implement based on the 5 imputed datasets.

Imputation no.	N=103 Projects, without outliers						
	Effort Plan Model, N=3						
	Intercept	Effort Specify	Effort Build	Effort Test	Adjusted R^2	R^2	P-value
1	86	-0.09	0.17	0.14	0.33	0.35	<0.0001
2	76	-0.08	0.18	0.14	0.34	0.36	<0.0001
3	82	-0.09	0.18	0.14	0.34	0.36	<0.0001
4	72	-0.08	0.17	0.14	0.34	0.36	<0.0001
5	85	-0.09	0.17	0.14	0.33	0.35	<0.0001

(N=103 projects, without outliers).

Table 18: Regression analysis estimation model for Effort Plan based on the 5 imputed datasets.

is used for combining the MI results. This step combines ***m*** sets of estimates and standard errors to obtain a single estimation model, standard error, and the associated confidence interval or significance test P-value.

The parameter estimates for MI display a combined estimate and standard error for each regression coefficient (parameter). The inferences are based on *t*-test distributions, as well a 95% confidence interval and a t-statistic with the associated P-value.

The P-value is the number attached to each independent variable in an estimation model, and represents that variable's significance level in the regression result. It is a percentage, and explains how likely it is that the coefficient for that independent variable emerged by chance and does not describe a real relationship.

A P-value of 0.05 means that there is a 5% chance that the relationship emerged randomly and a 95% chance that the relationship is real. It is generally accepted practice to consider variables with a P-value of less than 0.1 as significant.

There is also a significance level for the model as a whole, which is the F-value. This value measures the likelihood that the model as a whole describes a relationship that emerged at random, rather than a real relationship. As with the P-value, the lower the F-value, the greater the chance that the relationships in the model are real.

In addition, the t-statistic value is used to determine whether or not an independent variable should be included in a model. A variable is typically included in a model if it exceeds a predetermined threshold level or 'critical value'. The thresholds are determined for different levels of confidence: e.g. to be 95% confident that a variable should be included in a model, or, in other words, to tolerate only a 5% chance that a variable doesn't belong in a model. A t-statistic greater than 2 (if the coefficient is positive) or less than -2 (if the coefficient is negative) is considered statistically significant.

The strategy for combining results (Step 3) is as follows – see also section 5.1.2 on the analysis of variances [3]:

A. Combine the results, taking into account differences within datasets (variances, uncertainty due to missing data) and between datasets (variances, additional uncertainty due to imputation).

B. Estimate the parameter ($\overline{U}$).

C. Calculate the variances: within and between imputations (i.e. $\overline{U}$ and B) and the total variance of $\overline{P}$ as a function of $\overline{U}_m$ and B.

D. Combine Standard Error results to obtain the standard error: SE ($\overline{U}m$) = $\sqrt{Tm}$.

Average parameter estimates for MI of the full imputed dataset (N=103 projects): This section presents the variance information and parameter estimates from MI for the full 5 imputed datasets after removal of the outliers: the results of the 5 imputed dataset estimates are combined, and the averages of the parameter estimates are obtained using the results of the five estimation models in Step 2. This makes it possible to generate valid statistical inferences for the estimated analysis of dependent variables with missing values (i.e. Effort Plan and Effort Implement) on the observed values of the independent variables Effort Specify, Effort Build, and Effort Test. For instance, in Step 2, the results of 5 individual imputations for the intercepts were 170, 189, 194, 168, and 138 – see Table 17.

Calculation of variances: Table 19 shows the regression analysis of the EI parameter estimate and Table 20 the regression analysis of the EP parameter for the combined imputations.

After combining the results, the average intercept estimate for Effort Implement without outliers is 172 hours – see Table 19 (with a Standard Error of 60 hours), and the average estimation for the intercept for Effort Plan is 80 hours – (Table 20), with a Standard Error of 75 hours.

The Standard Error in Table 20 is obtained as follows:

- Intercept estimate:
- Within variance $\overline{U}m$ = 5,512 hrs, between variance Bm= 41 hrs,
- Total variance Tm = 5,512 + 1.2*41 = 5,561 hrs,
- Standard Error: SE = $\sqrt{5561}$ = 75 hrs.

Regression analysis of the Effort Implement (EI) parameter estimate: Table 19 also shows that the P-value of EB has a significant impact on effort (Effort Implement): the P-value is 0.01, with t-statistic of 2.60.

The effect of EB on the EI parameter is 2.69 (Table 19), which is higher than 2, and a P-value of 0.01, which is less than 0.1. Therefore, the EB parameter is statistically significant with EI.

The estimated effect of EI on ES, EB, and ET is 0.01, 0.08, and 0.07 respectively, with a t-statistic equal to 0.22, 2.6, and 0.77, and a P-value of 0.82, 0.01, and 0.48 respectively. The values of the t-statistic are less than 2, and so the intercept coefficient is not statistically significant. This means that the regression analysis results do not show evidence that EP has any impact on ES, but that it does have an impact on EB or ET. Moreover, the regression analysis results of EI do not show evidence that EI has any impact on ES or ET, but that it does have an impact on EB.

This means that in Table 19 the independent variables of ES and ET are not a significant predictor of the dependent variable of EI, and the variation in the dependent variable is not significantly explained by the independent variables.

Table 21 presents a summary of these results of the average estimate model of Effort Implement after they have been combined, without outliers. The test of the null hypothesis P-value in Table 21 shows that, of the three variables (ES, EB, and ET), ES and ET have a less significant impact on the Effort Implement estimate, while the P-value of EB is much more statistically significant.

Regression analysis of the Effort Plan (EP) parameter estimate: Table 20 shows that the P-values of EB and ET have a significant impact on effort (Effort Plan): the P-values are <0.0001, 0.0002, with a t-statistic of 4.85 and 3.75 respectively. Table 20 also presents a t-statistic of less than 2 for ES and P-values greater than 0.05, which means that the independent variable ES is not a significant predictor of the dependent variable of EP, and the variation in the dependent variable is not significantly explained by the independent variables for ES.

The estimated effect of EP on the EB and ET parameters are 0.18 and 0.14, with a t-statistic equal to 4.85 and 3.75. The effect of EI on EB is (0.08) with a t-statistic equal to (2.60), and a P-value of (<0.0001 and 0.0002) – see Table 20.

Since the t-statistic is greater than 2 and the P-value less than 0.1, we can conclude that the effect of EB and ET on the EP parameter and EB on the EI parameter is statistically significant.

Table 22 presents a summary of these results of the average estimate model of Effort Plan after they have been combined, without outliers.

Parameter	N=103 Projects, after removal of 3 outliers								
	Estimate	Std Error	95% Confidence Interval		t-Statistic	Variance			P-value
						Between BM	Within $\overline{U}m$	Total	
intercept	172	60	53	291	2.86	483	3028	3608	0.01
ES	0.01	0.06	-0.10	0.13	0.22	0.001	0.002	0.003	0.82
EB	0.08	0.03	0.02	0.15	2.60	0.0003	0.001	0.001	0.01
ET	0.07	0.09	-0.16	0.30	0.77	0.006	0.001	0.007	0.48

Table 19: Variance information from MI for Effort Implement (N=103 projects, after removal of outliers).

Parameter	N=103 Projects, after removal of 3 outliers								
	Estimate	Std Error	95% Confidence Interval		t-Statistic	Variance			P-value
						Between BM	Within $\overline{U}m$	Total	
intercept	80	75	-66	226	1.07	41	5512	5561	0.28
ES	-0.09	0.06	-0.21	0.04	-1.34	0.00001	0.004	0.004	0.18
EB	0.18	0.04	0.10	0.25	4.85	0.00002	0.001	0.001	<.0001
ET	0.14	0.04	0.07	0.22	3.75	0.00002	0.002	0.002	0.0002

Table 20: Variance information from MI for Effort Plan (N=103 projects, after removal of outliers).

Parameter	After outlier removal N= 103 projects	
	Significant t- test	Significant P-values
Intercept	Yes	Yes
ES	No	No
EB	Yes	Yes
ET	No	No

Table 21: Statistical significance of the parameter estimates of Effort Implement (N=103 projects).

Parameter	After outlier removal N= 103 projects	
	Significant t- test	Significant P-values
Intercept	No	No
ES	No	No
EB	Yes	Yes
ET	Yes	Yes

Table 22: Statistical significance of the parameter estimates of Effort Plan (N=103 projects).

The test of the null hypothesis P-value in Table 22 shows that, of the three variables (ES, EB, and ET), ES has a less significant impact on the Effort Plan estimate, while the P-value of EB and ET are much more statistically significant.

Summary of Observations

In summary, this paper identified a number of data quality issues associated with the ISBSG repository, and proposes a number of empirical techniques for preprocessing the data in order to improve the quality of the samples. It then focused on the issues of outliers and missing values: the presence of outliers in the ISBSG repository, and the use of MI to deal with missing values in the ISBSG repository, as well as considering the implications of the presence of outliers in numerical data fields.

The fact that a large number of data are missing from this repository, which comprises project data from a number of different companies, can considerably reduce the number of data points available for building productivity and estimation models. A few techniques have been developed for handling missing values, but it is essential to apply them appropriately, otherwise biased or misleading inferences may be made.

This paper worked on Release 9 (R9) of the ISBSG data repository, which contains information on 3,024 software projects developed worldwide.

We re-examined a statistical model that explains the variability in the total project effort field (Summary Work Effort), which was conditioned on a sample from the repository of 179 observational projects, and contains covariate effort by activity (Plan, Specification, Build, Test, and Implement).

Our investigation included an analysis of outlier behavior in the ISBSG repository, and outlier tests were performed on the effort estimation model built based on functional size and on the ISBSG's total work effort variables. This model was conditioned on an initial sample of 106 observational projects from the repository. When effort estimation models are built using data samples with outliers, these models degrade the effort estimation models available for future projects. Therefore, we examined the effort estimation model when the outlier test method is applied on functional size and the total work effort variables for the ISBSG repository datasets. The results of the model changed substantially, depending on whether they were computed with or without outliers. We show that applying the outlier test method avoids some biases in the results of the effort estimation model.

This paper investigated the use of the multiple imputation (MI) technique with the ISBSG repository for dealing with missing values, and reported on its use. Five imputation rounds were undertaken to produce parameter estimates which reflect the uncertainty associated with estimating missing data.

This paper also investigated the impact of MI in the estimation of the missing values of the effort variable by project activity using the ISBSG repository, and applied regression models, both withand withoutoutliers, and examined their specific influence on the results.

In addition, the averages of the effort distribution by activity were determined for three profiles (PSBTI, PSBT, and SBTI) and for each of the five imputation rounds. The PSBT profile presents a missing activity (Effort Implementation), and the SBTI profile presents a missing activity (Effort Plan). As a result, the average of the effort distributions of the other **activities** (Effort Specification, Effort Build, and Effort Test), as well as the combined average of the effort distribution of all the projects, varied accordingly in each imputation.

The regression analysis was trained with the five imputed datasets from 63 projects (without outliers). It was observed that the adjusted R^2 is lower for the dataset without outliers, indicating that the outliers

unduly influenced the estimation models, leading to statistical over confidence in the results.

This paper then showed:

A) the results of multiple imputation variance information, and

B) imputed values for the Effort Implement and Effort Plan variables over the five imputed datasets.

A. The results of this investigation revealed that the variance results of the standard error of the imputed

values decreased from 105 hours to 73 hours for Effort Implement, and from 106 hours to 60 hours for

Effort Plan for a multiple regression analysis with and without outliers respectively – see Tables 7 and 8.

B. Furthermore, the multiple regression analysis results were statistically significant for the Effort Plan and

Effort Implement parameters, as illustrated by the t-test and P-values without outliers.

The paper also presented the results of five effort estimation models that were combined with the five imputed dataset estimates, and obtained the averages of the parameter estimates. The results of this investigation show the results of three variables (ES, EB, and ET):

A. The P-value of the EB and ET variables presented a statistically much higher significant impact on the effort estimate than the ES variable.

B. The estimated effect of ES and ET on the EI parameter was 0.02 and 0.07 respectively, with a t-statistic equal to 0.22 and 0.77, and P-values of 0.82 and 0.48 respectively. Note that the values of the t-statistic were also less than 2 – see Table 19.

C. The estimated effect of EP on the ES parameter was -0.09, with a t-statistic equal to -1.34 and P-values of 0.18. Note that the values of the t-statistic were less than 2 – see Table 20.

D. The intercept coefficient is not statistically significant – see Table 20.

This means that the multiple regression analysis results did not find evidence that ES and ET have any impact on the EI (Effort Implement) or EP (Effort Plan) parameters, but they do have an impact on the EB (Effort Build) parameter.

References

1. ISBSG (Release12) (2013) International Software Benchmarking Standards Group.
2. Myrtveit I, Stensrud E, Olsson UH (2001) Analyzing data sets with missing data: an empirical evaluation of imputation methods and likelihood-based methods. IEEE Transactions on Software 27: 999-1013.
3. Rubin DB (1987) Multiple imputation for non response in surveys. John Wiley & Sons.
4. Abran A, Ndiaye I, Bourque P (2007) Evaluation of a black-box estimation tool: a case study. Software Process Improvement and Practice 12: 199-218.
5. Pendharkar PC, Rodger JA, Subramanian GH (2008) An empirical study of the Cobb–Douglas production function properties of software development effort. Info and Soft Tech 50: 1181-1188.
6. Xia V, Ho D, Capretz LF (2006) Calibrating Function Points Using Neuro-Fuzzy Technique. 21st International Forum on COCOMO and Software Cost Modeling, Herndon, Virginia.
7. Deng K, MacDonell SG (2008) Maximising data retention from the ISBSG repository. 12th International Conference on Evaluation and Assessment in Software Engineering (EASE).
8. Déry D, Abran A (2005) Investigation of the Effort Data Consistency in the ISBSG Repository. 15th International Workshop on Software Measurement -- IWSM'2005, Montreal, Canada, Shaker-Verlag.
9. Jiang Z, Naudé P, Jiang B (2007) The effects of software size on development effort and software quality. Intr J comp inform Sci Eng 1: 230-4.
10. Graham JW, Schafer JL (1999) On the performance of multiple imputation for multivariate data with small sample size. In Hoyle R (ed.). Statistical strategies for small sample research 1999: 1-29.
11. Schafer JL, Graham JW (2002) Missing data: Our view of the state of the art. Psychological Methods 7: 147-77.
12. John WG, Scott MH, Stewart ID, MAcKINNON DP, Joseph LS (1997) Missing Data Analysis in Multivariate Prevention Research. In: K. Bryant, M. Windle, & S. West (eds.), Washington D.C. pp: 325-66.
13. SAS Institute Inc (1999) SAS Procedures Guide. Version 8, Cary, NC: SAS Institute Inc.
14. Kuhnt S, Pawlitschko Jr (2003) Outlier identification rules for generalized linear models. Innovations in Classification, Data Science, and Information Systems.
15. Davies L, Gather U (1993) The Identification of Multiple Outliers. Journal of the American Statistical Association 88: 782-92.
16. Abran A (2015) Software Project Estimation – The Fundamentals for Providing High Quality Information to Decision Makers. Wiley & IEEE-CS Press – Hoboken, New Jersey.