Data Sets and Data Quality in Software Engineering: Eight Years On

Gernot Liebchen
Bournemouth University
Fern Barrow, Poole, BH12 5BB
UK
gliebchen@bournemouth.ac.uk

Martin Shepperd
Brunel University London
Uxbridge, UB8 3PH
UK
martin.shepperd@brunel.ac.uk

ABSTRACT

Context: We revisit our review of data quality within the context of empirical software engineering eight years on from our PROMISE 2008 article.
Objective: To assess the extent and types of techniques used to manage quality within data sets. We consider this a particularly interesting question in the context of initiatives to promote sharing and secondary analysis of data sets.
Method: We update the 2008 mapping study through four subsequently published reviews and a snowballing exercise.
Results: The original study located only 23 articles explicitly considering data quality. This picture has changed substantially as our updated review now finds 283 articles, however, our estimate is that this still represents perhaps 1% of the total empirical software engineering literature.
Conclusions: It appears the community is now taking the issue of data quality more seriously and there is more work exploring techniques to automatically detect (and sometimes repair) noise problems. However, there is still little systematic work to evaluate the various data sets that are widely used for secondary analysis; addressing this would be of considerable benefit. It should also be a priority to work collaboratively with practitioners to add new, higher quality data to the existing corpora.

Keywords

empirical software engineering; data quality; mapping study.

1. INTRODUCTION

It should go without saying that data quality is a central concept for any empirical discipline and this is certainly the case for empirical software engineering (ESE). In some senses there is a particularly strong reliance upon the underlying correctness of the data since we have little underlying theory to guide the researcher and the kinds of models that might be developed. The situation is most acute when using inductive methods such as machine learners coupled with a tendency to use secondary data, i.e., data not collected by the researchers themselves. Thus incorrect models and misleading conclusions can easily arise if the data are incorrect. In the worst case the researchers and practitioners could be unaware of such problems. Consequently it is no surprise that the community are starting to become concerned about such problems.

In 2008 we published a review article that considered the state of play regarding data quality in ESE [7]. One of the first challenges we identified was how to define data quality since there are a range of views. We adopted, and continue to hold, a narrow view of data quality, namely its accuracy, otherwise referred to as noise. It is important to point out that noisy instances are *not* synonymous with outliers. Outliers are simply extreme instances which stand out from the distribution of data observations, but they can be true exceptional instances. Noise are data items containing errors, i.e., the recorded value deviates from the true, but possibly unknown, value. Outliers may help detecting these errors, but are not necessarily errors in their own right.

We recognise that other aspects like "fitness for purpose" are also important issues, however, they remain beyond the scope of this investigation since we cannot know the purpose of a data set *a priori*. Moreover purposes can vary over time meaning that it would be extremely difficult to operationalise such a definition. Similarly, although it is a major research area in its own right [8], we exclude issues of incompleteness or missingness (and data imputation) since this can be clearly identified and therefore the potential for misleading analysis is reduced.

Likewise we also exclude redundancy and inconsistency as long as these reflect the 'true' data. Both of these 'problems' arise from the perspective of machine learning; these are circumstances that a prediction system must deal with, by whatever means, hence they are not strictly speaking data quality issues. So to summarise, we consider data quality to mean the absence of noise or incorrect data [7].

In 2008 we identified five main themes.

1. A very small proportion of studies directly consider quality. Using a slightly restricted search we located only 23 out of the many hundreds of studies.
2. The dominant approaches for handling quality were (i) manual inspection / triangulation and (ii) prevention through better data collection techniques.
3. We located little work to independently assess the quality of a given data set (typically such approaches made use of quality meta-data which were essentially surrogates for the level of incompleteness within data sets).
4. We commented on the need for more research into

Promise 2016 September 7, 2016, Ciudad Real, USA

DOI: 10.475/123_4

Table 1: Summary of Review Article Counts

Review	Acronym	Year	Articles
Liebchen [6]	L10	2010	161
Rosli et al. [12]	R13	2013	64
Bosu [1]	B16	2016	282
Bosu extra[1]		2016	57
snowball		2016	12
Total unique			460
Total relevant			283

automatically identifying, and ideally repairing, noisy cases.

5. There was little guidance for researchers to locate and use the higher quality data sets.

Since our 2008 article there have been a number of systematic reviews / mapping studies[1]. In 2010 Liebchen published a significantly extended data quality mapping study as part of his doctoral thesis [6]. In 2013 Rosli et al. [12] conducted another mapping study and in parallel Bosu and MacDonell published a further mapping study [2]. Most recently Bosu updated this as part of his doctoral thesis [1]. These studies are integrated with a further snowball search on our part to include studies up until May 2016.

The remainder of this paper describes how we collated and updated these various reviews of data quality. We provide some bibliometric data on the current state of play and then consider each of the five findings from 2008 and the extent to which there have been changes. The paper concludes with a series of recommendations to the ESE community: researchers and practitioners.

2. METHOD

In order to identify relevant post-2008 literature the findings of all five literature reviews were collated [7, 6, 12, 2, 1]. The process was simplified since in practice [6] subsumes all but one paper in [7] and [1] subsumes [2] leaving us with three reviews to integrate. Although these studies focus on how data quality is dealt with in the software engineering community, they vary in their inclusion criteria. Liebchen [6] and Rosli et al. [12] focussed on the accuracy of data, whereas Bosu [1] has a more general approach that included data quality dimensions such as missingness and timeliness.

Our review was then brought up to date by a snowball search based on papers that cite[2] our original 2008 review [7] which yielded a further 12 articles. The studies and paper counts are summarised in Table 1.

The articles were then combined and re-checked using the same inclusion criteria used in L10. Specifically a paper must be (i) in the domain of ESE, (ii) explicitly address some aspect of data accuracy or noise, (iii) be refereed, (iv) written in English and (v) be available. No time limit was imposed. The raw data comprising the detailed searches, all papers and their categorisations may be found online[3].

3. FINDINGS AND DISCUSSION

[1]Technically the reviews are mapping studies since they do not have precise questions but rather seek to understand the research work conducted in a particular area [5].

[2]62 citing papers were identified from Google Scholar in June 2016 resulting in 12 additional articles.

[3]Please see https://github.com/gliebchen/Data-Sets-and-Data-Quality-in-Software-Engineering-Eight-Years-On

The original study [7], which this paper revisits, identified 23 articles. In contrast, two years later, the review L10 identified 161, R13 identified 64 and B16 identified 282. Bosu [1] also identified an additional 57 papers. This results in 460 unique articles. Note that the union of the reviews yields 460 articles and is less than the sum since duplicates have been removed (see Table 1). Relevant articles are further reduced to 283 due to our more inclusion criteria being more stringent than for B16.

A question arises from the substantial difference between the 23 articles identified in 2008 [7] and the considerably greater numbers found by the subsequent studies. This is due to the restrictive search strategy of the original study which searched for "data quality" only. Clearly the concepts around data quality can be expressed in multiple ways.

Subsequent to the original 2008 review, L10 extended the search strategy, and included variations of search terms such as "noise", "inconsistent pieces" and "erroneous data". R13 used a similar search strategy, but restricted the search to articles published 2008-2012. B16 is less specific concerning the search string used, but the focus was on articles published between 2007 and 2014. Also recall B16 included other aspects of data quality than just accuracy like timeliness. This contributes to the reduction in the number of articles considered as relevant from 460 to 283 (see Table 1).

In order to see the level of agreement between the three reviews the years 2008 and 2009 were inspected more closely since these are the only years completely covered by all three reviews (see Table 2). In this period there are a total of 69 known relevant papers, i.e., the union of L10, R13 and B16 for this time period. The largest agreement is between the reviews by L10 and R13. B16 and L10 identified most papers separately from the other reviews. It can also be seen that a number of papers was missed by each review respectively. The lowest overlap is between B16 and R13 where only 9 articles were in common out of a theoretical possibility of 69. The inconsistencies between the studies show that there can be a degree of imprecision for complex searches and more general mapping studies, unlike the consistency that has been observed for precisely specified systematic reviews e.g., [9]. Consequently, we focus on the general patterns rather than the niceties of exact counts.

Table 2: Overlap between the Reviews

	B16 (including 57 extra)	R13
L10	17	40
B16	–	9

3.1 RQ1: Proportion of Studies to Consider Quality

In our 2008 review we commented on the low number of studies—23 to be precise—to explicitly address data quality within ESE. This has now dramatically increased to 283. As discussed in the previous section, whilst the number risen it should also be noted that subsequent reviews have been more wide ranging, in particular in their exploration of alternative formulations to "data quality". Thus an additional 76 articles have been found in the time period covered by the original 2008 review meaning we under-estimated the literature by about 75%. There are many ways to describe data quality and it is possible that even now some relevant articles have been missed. The problem could be lessened if researchers used standard reporting protocols as this would (i) facilitate searching and (ii) encourage more widespread

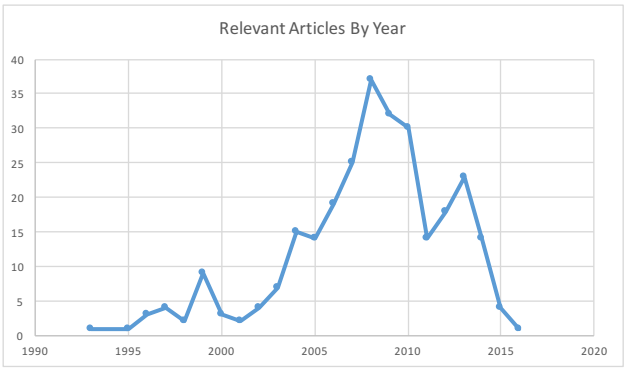

Figure 1: Publication Trends Over Time

consideration of data quality amongst researchers.

The next aspect to consider is any trend information concerning attention to data quality (see Fig. 1). Here it is clear that there has been growth since 2005, however, this seems to have tailed away in recent years. Obviously 2016 is incomplete, however, it does appear that data quality is still not widely regarded as a very important topic or at least it is something that can be taken for granted.

It is not easy to estimate the total population of ESE studies, but order of magnitude we are dealing with thousands[4], thus, the underlying conclusion that the majority of studies continue not to explicitly consider data quality remains inescapable.

3.2 RQ2: Approaches to Data Quality

As was the case in 2008, most articles note that noise is a potential difficulty ($\sim$ 90%) and remarkably few (1 out of 283) explicitly exhibit confidence that this is not the case. Within the overall total of papers software quality (defect prediction) and cost estimation remain the dominant application domains and there seem little difference in the mix of approaches other than a greater exploration of automated detection algorithms in software quality (see Table 3).

Table 3: Paper Categorisation vs Domains

	Data Collection	Manual Noise Checking	Automated Noise Checking	Empirical Analysis of Noise	Data Quality Meta Data	Is noise a problem?	Total
	Y	Y	Y	Y	Y	Y	
Software Quality	19	14	19	25	4	106	110
Cost Prediction	13	14	9	21	36	100	122
Other Domains	26	20	4	8	2	64(No:1)	72
Total unique	57	47	30	46	40	253(No:1)	283

In terms of techniques, although improving data collection seems to be the most popular, by a rather small margin, approach. However, manual checking for example through triangulation is increasing in popularity as has use of meta-data principally with the ISBSG data set and the quality indicator fields [4].

Valverde et al. [15] suggested the application of a data quality model as a form of data protocol for new data sets as initially suggested by Liebchen [6]. In a similar approach Rosli et al. [13] suggested the application of their meta-model to evaluate new data sets.

3.3 RQ3: Dealing with Data Quality for Secondary Data

The pattern of using secondary data, typically data sets that have been made publicly available through various repositories, remains the norm. Thus some data sets are highly reused. There are few articles that systematically evaluated data quality and made recommendations as to data set usefulness or otherwise. Probably the most work in this regard concerns the NASA defect data sets and some studies have considered the impact of problematic instances [14, 10]. Data cleaning appears to impact at least some experimental findings.

It is also encouraging to see that two articles [11, 13] looked at the quality meta-data provided for publicly available data sets. In fact Rosli et al. [13] suggested that 61 out of 70 investigated data sets did not contain enough information in the form of meta-data to allow correct interpretation of these data sets' data.

3.4 RQ4: Automated Detection and Repair

There is a small but growing body of work exploring automated detection and in some cases repair algorithms. We found 46 articles empirically assessing the quality of software engineering data sets, of these 30 articles employed some form of automated process and 25 articles combined empirical analysis with automated noise detection algorithms.. These algorithms are largely based on outlier detection techniques, however, care needs to be taken that outliers are not automatically treated as noise.

The researcher most active in the areas of automated noise detection in software engineering is Khoshgoftaar who has contributed to 12 papers concerned with these issues, with the latest published paper in 2009 [16]. The next most frequent author is Van Hulse who has five papers; all of them co-authored with Khoshgoftaar e.g., [16].

Khoshgoftaar and his team's contribution is valuable, but we believe that the community needs more research groups to be engaged on automated noise detection and empirical analyses of noise. This would enrich the community since results could be verified and analysed by independent groups of researchers.

Yoon and Bae [17], for instance, evaluated six different data cleaning techniques from different research groups. They compared these techniques' ability to detect noise in three real world and 48 artificial software engineering data sets. Evaluating data cleaning techniques on artificial data sets is interesting, since the true level of noise can be known with absolute certainty. We believe that evaluations of data cleaning techniques ought to be compared against different evaluation methods as different evaluation methods may result in outcomes which may contradict each other [6].

3.5 RQ5: Guidance to the Community

Unfortunately we still lack systematic guidance for researchers

[4]To get a feel for the number of papers written in the ESE domain, the search phrase "empirical AND "software engineering"" was entered in basic searches of the ACM, IEEE Xplore, ScienceDirect, Scopus and Springer databases (data: 17/06/2016), resulting in the following counts of retrieved papers; ACM-6385, IEEE Xplore-3077, ScienceDirect-7928, Scopus-5396, Springer-6087 (articles only). The total is in excess of 28873. As a very crude measure 283 out of 28873 (about 1%) papers written in the ESE literature explicitly discussed data quality.

to use the better quality data sets that are available via various repositories such as PROMISE.

Data quality meta-data and protocols as proposed by Phannachitta et al. and Rosli et al. [11, 13] would be useful to understand, not just the data quality issues, but also the actual data itself. This would help researchers to draw better conclusions, and it would hopefully help to build better models. Protocols would also be helpful to understand the impact of any preprocessing. Of interest would clearly be the original state of a data sets, reasons for any preprocessing, the instances in a data sets which were excluded (or preprocessed) and the reasons for excluding any instances. Data sets have been referred to as clean after a preprocessing process, but this is difficult to understand as the absence of data quality problems is somewhat difficult to prove [3].

The encouraging increase of papers suggesting or evaluating the use of automated data quality algorithms is very promising. However, it would be good to see more experimental evaluations such as Yoon and Bae's [17] who evaluated a number of automated data cleaning techniques used to deal with noise in software engineering data sets. It would also be interesting to compare results of different data cleaning techniques against different cleaning performance measures since use of the latter can impact on the conclusion about their effectiveness [6].

4. CONCLUSIONS

The work presented in this paper departs from a full systematic literature review or mapping study in that (i) we did not develop a formal protocol and (ii) there has been no independent validation of the application of the inclusion criteria or the coding by different researchers. This paper's main interest is revisiting the findings of our 2008 study to get a feel of the state of play in the ESE community and in particular identify trends and changes since 2008. It also highlights issues with our original study and subsequent studies and we believe maps out the "big picture".

There are two sets of implications: first for the research community and second, for practitioners.

Researchers need to have concern for data quality. Some details may seem small but can nevertheless have significant impact upon results. We need to develop mechanisms for creating quality labels and deprecating data sets where there are serious concerns. It does not seem helpful for either researchers or practitioners to invest resources in continuing to work with suspect data.

The theme of protocols both to describe data and quality issues and for reporting of individual studies is gaining momentum. We believe this is something that the community could very usefully undertake and would facilitate the wiser and more effective use of secondary data.

In order to have trustworthy data-driven research then we need to have greater confidence in its quality and the lack of noise. Collaboration with *practitioners* in the collection and interpretation of data would be invaluable. Ultimately this will lead to more reliable and more actionable research which will be for the benefit of all.

5. REFERENCES

[1] M. F. Bosu. *Data Quality in Empirical Software Engineering: An Investigation of Time-Aware Models in Software Effort Estimation*. PhD thesis, University of Otago, Dept. of Information Science, NZ, 2016.

[2] M. F. Bosu and S. G. MacDonell. Data quality in empirical software engineering: a targeted review. In *17th Intl. Conf. on Evaluation and Assessment in Software Engineering*, pages 171–176. ACM, 2013.

[3] R. D. De Veaux and D. J. Hand. How to lie with bad data. *Statistical Science*, 20(3):231–238, 2005.

[4] F. González-Ladrón-de Guevara, M. Fernández-Diego, and C. Lokan. The usage of ISBSG data fields in software effort estimation: A systematic mapping study. *J. of Systems & Software*, 113:188–215, 2016.

[5] B. A. Kitchenham, D. Budgen, and O. P. Brereton. Using mapping studies as the basis for further research–a participant-observer case study. *Inf. Softw. Technol.*, 53(6):638–651, 2011.

[6] G. Liebchen. *Data Cleaning Techniques for Software Engineering Data Sets*. PhD thesis, Brunel University, London, 2010.

[7] G. Liebchen and M. Shepperd. Data sets and data quality in software engineering. In *4th Intl. Workshop on Predictor Models in Software Engineering*, pages 39–44. ACM, 2008.

[8] R. Little and D. Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, New York, 2nd edition, 2002.

[9] S. MacDonell, M. Shepperd, B. Kitchenham, and E. Mendes. How reliable are systematic reviews in empirical software engineering? *IEEE Trans. on Softw. Eng.*, 36(5):676–687, 2010.

[10] J. Petric, D. Bowes, T. Hall, B. Christianson, and N. Baddoo. The jinx on the NASA software defect data sets. In *20th Intl. Conf. on Evaluation and Assessment in Software Engineering*. ACM, 2016.

[11] P. Phannachitta, A. Monden, J. Keung, and K. Matsumoto. Case consistency: A necessary data quality property for software engineering data sets. In *19th Intl. Conf. on Evaluation and Assessment in Software Engineering*.

[12] M. M. Rosli, E. Tempero, and A. Luxton-Reilly. Can we trust our results? a mapping study on data quality. In *Software Engineering Conference (APSEC), 2013 20th Asia-Pacific*.

[13] M. M. Rosli, E. Tempero, and A. Luxton-Reilly. What is in our datasets?: Describing a structure of datasets. In *Proceedings of the Australasian Computer Science Week*, ACSW '16, pages 28:1–28:10, New York, 2016. ACM.

[14] M. Shepperd, Q. Song, Y. Sun, and C. Mair. Data quality: Some comments on the NASA software defect data sets. *IEEE Trans. on Softw. Eng.*, 39(9):1208–1215, 2013.

[15] M. C. Valverde, D. Vallespir, A. Marotta, and J. I. Panach. Applying a data quality model to experiments in software engineering. In *Advances in Conceptual Modeling - ER 2014 Workshops*, pages 168–177, 2014.

[16] J. Van Hulse and T. M. Khoshgoftaar. Knowledge discovery from imbalanced and noisy data. *Data Knowl. Eng.*, 68(12):1513–1542, 2009.

[17] K.-A. Yoon and D.-H. Bae. A pattern-based outlier detection method identifying abnormal attributes in software project data. *Inf. Softw. Technol.*, 52(2):137–151, 2010.