Data Sets and Data Quality in Software Engineering

Gernot A. Liebchen
Brunel University
United Kingdom
gernot.liebchen@brunel.ac.uk

Martin Shepperd
Brunel University
United Kingdom
martin.shepperd@brunel.ac.uk

ABSTRACT

OBJECTIVE - to assess the extent and types of techniques used to manage quality within software engineering 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 perform a systematic review of available empirical software engineering studies.
RESULTS - only 23 out of the many hundreds of studies assessed, explicitly considered data quality.
CONCLUSIONS - first, the community needs to consider the quality and appropriateness of the data set being utilised; not all data sets are equal. Second, we need more research into means of identifying, and ideally repairing, noisy cases. Third, it should become routine to use sensitivity analysis to assess conclusion stability with respect to the assumptions that must be made concerning noise levels.

Categories and Subject Descriptors

D.2.9 [**Software Engineering**]: Management—*Cost estimation, Software Quality Assurance, Time estimation*

General Terms

Experimentation, Management, Measurement

Keywords

Data sets, empirical research, prediction, data quality

1. INTRODUCTION

As the discipline of empirical or evidence-based software engineering matures there has been a growing move to publicly archive data sets with a view to encouraging secondary and meta style analyses. The PROMISE Group [11] are prominent in this regard. Clearly such initiatives are to be applauded.

However, from our own experience we have growing concerns about the quality of some of the data sets we have been using to learn, evaluate and compare competing prediction systems. Specifically we have questioned whether the noise levels prevent meaningful conclusions being drawn. This can be particularly important in circumstances where the researchers are somewhat remote from the actual data collection process or where secondary / meta-analyses are being performed. Thus we have started to explore techniques for the identification, elimination or repair of noisy cases within data sets. We are also interested in the approaches adopted by other researchers.

Note that there are many definitions of data quality and many inconsistent views. The most widely accepted data quality definition defines it in terms of "fitness for purpose" [36, 28, 33, 10], which is derived from a more general definition of quality due to Crosby [5]. We agree that it is only possible to meaningfully examine data quality in the presence of some purpose, i.e. consider what the data are to be used for. Typically, in the empirical software engineering community, this is to predict a dependent variable such as project effort or defect count. For example, quality problems may impact the goal of predicting X very differently from comparing development techniques A and B even though the *same* data set is involved in both cases.

Although data quality is frequently viewed in terms of accuracy or absence of noise there are many alternative and supplementary perspectives such as completeness and timeliness. Their importance and the number of these dimensions are again more problem dependent than absolute [28, 36]. See [3] for a more detailed discussion concerning the various data quality dimensions and [6] for a discussion of the extent and impact of data quality problems in general.

In this paper we adopt a specific view of data quality, namely its accuracy, i.e. noise. Whilst other aspects like completeness and timeliness are also important issues, they are beyond the scope of this investigation. There is a good deal of work elsewhere on completeness or its obverse missingness [24].

Noise can be defined as incorrect data. Some machine learning researchers also incorporate outliers in their definition of noise [9, 2]. Outliers can be unwanted for some analysis methods, but they may well be "true" or correct data points, and can be catered for with robust analysis techniques. Thus we do not, in general, see outliers as particularly problematic and indeed they can often form an important part of an empirical study, to investigate why some cases depart from a typical pattern.

The focus in this paper is on inaccurate or noisy data points, which occur not due to exceptional circumstances,

PROMISE'08, May 12–13, 2008, Leipzig, Germany.

but due to intentional or unintentional errors in capturing, transferring or editing of software metrics data. We explore how the community perceives and addresses this problem and propose a possible research agenda. The next section provides more details of our systematic review. This is followed by a description of the results. We conclude the paper with a discussion of the significance of these findings and a list of open research questions that we believe are of some importance for the empirical software engineering community.

2. METHOD

In order to survey the empirical software engineering literature concerning the subject of data quality, a systematic literature review was carried out. In recent years there has been increasing interest in establishing software engineering as an evidence based discipline [19] and a crucial part of this process is the systematic review. Systematic reviews are widely adopted in many other disciplines such as medicine, social policy, educational psychology, etc. The aim is to make the process of identifying *all* relevant studies and synthesizing the results into some overall, coherent picture unbiased and repeatable. A review is the process which requires an exhaustive scanning of all available literature that satisfies some agreed protocol that, amongst other things, will contain an unambiguous description of the inclusion criteria that a study must satisfy in order to be entered into the review.

The main objective for our search was to discover which studies explicitly consider noise or data quality in empirical software engineering. We are aware, as we mentioned in the introduction, that data quality has other dimensions apart from accuracy, but we were principally interested in accuracy (noisiness) as measure for data quality.

For the search we looked for the terms "data quality" AND "software". We used the ScienceDirect, SCOPUS and IEEExplore bibliographic databases to make a general search for relevant articles. This resulted in 552 hits, omitting duplicates[1]. The articles were then scanned to determine if they concerned empirical software engineering applications. This was supplemented by an exhaustive, hand search of those sources that we considered particularly relevant, namely the journal Empirical Software Engineering and the conference series of ESEM, METRICS, ISESE, PROMISE and EASE[2]. We recognise that the lack of online availability of published studies may slightly restrict our analysis, however we are confident that our search has covered the major empirical software engineering publication venues. Therefore we consider that the results provide us with an adequate view of the state of affairs in this community.

The results of these two search strategies were combined and then the articles checked against the inclusion criteria. These criteria are that the article must:

- focus on an empirical investigation of some aspect of software engineering or address some methodological issue relevant to such empirical research
- address data quality explicitly
- be refereed
- be written in English

The located papers were then analysed according to the following further sub-objectives, namely:

- How significant do the community consider noise to be (in principle and in practice)?
- How do empirical analysts address this problem?
- Are there techniques that might be deployed to independently assess the quality of a given data set?

Next we go onto consider the results of this systematic review in more detail.

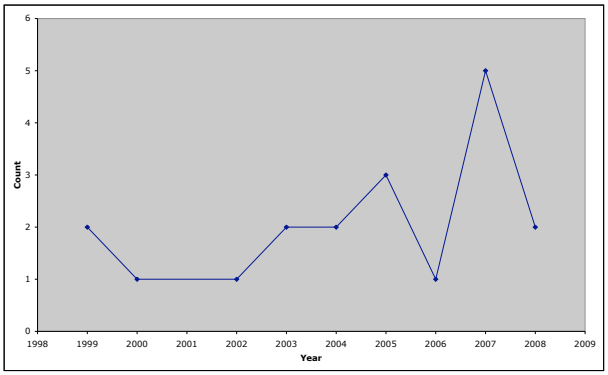

Figure 1: Data Quality Papers Retrieved by Year

3. FINDINGS

Although we scrutinised many hundreds of papers, our systematic review of the published literature identified only 23 refereed empirical software engineering papers that explicitly referenced the data quality and satisfied our inclusion criteria. There are other possible synonyms for "data quality" so our search may not have included every single relevant study, nevertheless we are somewhat surprised by the lack of explicit attention given to this topic by our community. This is particularly so when one considers the many hundreds of papers published that analyse empirical software engineering data.

Figure 1 gives a breakdown of the 23 papers by year. Note that 2008 is of course incomplete. Nonetheless there seems something of an increase over time, suggesting that the community is giving the topic of quality more explicit attention. We now look in more detail at the various messages from these papers.

[1]Papers that report the same empirical study are only counted once using the most recent publication, e.g. [7, 14, 15] describe the same study and so only [15] is included in our analysis.

[2]Results from EASE proceedings were limited to 2007 and 2006 due to the lack of online availability of the remaining proceedings.

	No	n.a.	Yes
Paper count	2	4	17

Table 1: Studies Commenting Upon the Potential Significance of Data Quality

The papers cover a range of topics within empirical software engineering. These include: meta-analysis, defect prediction and reliability modelling, effort prediction, PSP, reuse and studies evaluating different noise reduction strategies. Only two studies addressed the quality of qualitative data. Whilst Li et al. [20] focused on data derived from interviews, Johnson [13] highlighted the importance of quality review documents as a basis for software development reviews, but Johnson also pointed out the importance of accuracy of the metrics used in the review process.

Although a substantial majority of writers considered data quality to be a threat to empirical data analysis (see Table 1), not all papers agreed with this proposition. Indeed one author suggested that the random, i.e. unbiased nature of noise meant that it could be ignored since presumably it would average out in the long run [37]. Whilst this may mean that measures of centre are largely unaffected there may be considerable impact upon the variance and our ability to fit and differentiate between predictors. Thus we cannot remain so sanguine. This author also notes the possibility of intentional errors and states that these should be dealt with by validating the data, a sentiment hard to disagree with!

Four papers merely flagged up the potential problem of data quality and how it might impact analysis, but made no suggestions as to how this might be combatted. Otherwise we classify the contribution of each paper in Table 3. Note that a paper may make multiple contributions so the Paper Count does not sum to 23.

Data Quality Topic	Paper Count / 23
Data collection	7
Manual quality checking	12
Empirical quality analysis	9
Automated quality checking	6
Usage of quality meta data	2
Special analysis techniques	1

Table 2: Data Quality Topics Addressed by Papers in the Systematic Review

Approximately 30% of the papers emphasize the role of data collection, particularly the usage of software tools and the need to make the task less onerous for those involved in reporting the data. One might characterise this as noise prevention activity.

The area of most interest (more than 50% of the papers retrieved) is that of manual data quality checking. Typically this involves increasing one's confidence in a data set by some manual intervention such as independent scrutiny or the use of triangulation, for example measuring the same attribute in different ways or through inter-rater reliability analysis. This cleaning or scrubbing precedes the main analysis.

In the earliest paper retrieved by our search, Gulezian [12] recognises that in empirical software engineering the analyst is often working with secondary data, which requires serious attention to the nature and quality of the data. He then lists data quality considerations to be taken into account when working with that sort of data. Though not directly addressing accuracy, these indicators of data quality can be used to scrutinise the given data.

Another important topic is empirical assessment of the data quality problem. It is one thing to speculate about the extent and another to collect hard data. Of course this in itself can be challenging since in many cases the "true" value of a data item may be unknown. Thus unless the value is implausible in some way, the level of noise in a data set may be difficult to measure. However, an important example of data quality assessment, and an early paper on the topic, comes from Johnson and Disney [15]. They report that as part of the data recording process of the Personal Software Process (PSP) for 89 projects completed by ten participants they discovered 1539 primary errors. However, it must be stated that almost half (46%) of the errors were incorrect calculations and so can be addressed by the provision of better tool support. Another significant problem they encountered were missing data. There is a good deal of research on data imputation [24], however, we consider this to beyond the main thrust of our investigation. Overall they concluded that to improve data quality manual data collection should be avoided and "external measures" should be used which we interpret to mean for triangulation purposes.

Particularly relevant for data archives such as PROMISE is the ability to evaluate data quality independently. Ideally it is then possible to identify and quarantine noisy items or cases. This is a challenging area since we are not simply looking for outliers. The main work in this area has been by Khoshgoftaar and colleagues [17, 18, 16, 34, 35] and ourselves [21, 22, 23]. Such work is still at a relatively early stage with a major challenge being how to evaluate such techniques since the "true" value of each data item may be unknowable. Typically the approach is to learn some classifier over the data and then treat misclassified cases as suspect.

Some data sets, most notably the ISBSG project effort and productivity benchmark data set, contain meta data that describe the perceived quality of each case or project. For ISBSG quality is graded between A (highest quality) and D (lowest quality). This is quite important for situations where organisations elect to contribute project data to a central repository and thus there is a reduction in control over collection procedures. In the situation of ISBSG the classification is principally guided by the completeness of a case, in other words high quality data are interpreted as possessing low levels of missingness. In our review the two studies that utilised the ISBSG both adopted the strategy of only using data graded as A or B. Note though that this doesn't accord with our view that data quality concerns the difference between the "true" and recorded values for a data item. Indeed a complete case may contain many inaccuracies.

The final area of activity that we identified was the explicit adoption of specialised analysis techniques that are robust to the presence of noise. However such an approach does not feature very strongly amongst the papers we identified. Only one paper made any explicit mention of using a particular

analysis method due to data quality considerations and even this was limited to merely aggregating low level data on the grounds that if the noise were unbiased this would reduce the variance. Surprisingly sensitivity analysis does not appear to have been widely considered as means of assessing the potential impact of noise upon conclusion stability.

	Not sig.	n.a.	Sig.	Total
Effort estimation	0	2	5	7
Other	2	2	12	16
Total	2	4	17	23

Table 3: The Significance of Data Quality by Problem Domain

As indicated earlier, these papers encompass a breadth of topic under the general banner of empirical software engineering. The contingency table (Table 3) gives some support to the notion that the problem of data quality is perceived as particularly acute in the domain of project effort. This is unsurprising given that this type of data must of necessity be collected in the field rather than the lab and since project completion is a relatively infrequent event the data will tend to be historical. This underlines the fact that different problem domains will tend to have their own particular data quality issues.

4. DISCUSSION

So what is the significance of this investigation? We believe that it is rather important to explicitly consider the quality — meaning the accuracy — of the data sets that form the basis of our research. Clearly, poor data quality can threaten the meaningfulness of our conclusions. Worse still, we don't wish to perpetuate such problems by reusing suspect data sets.

So how problematic do the community consider noise to be? We are surprised by the very small proportion of papers that consider this issue directly. We found a total of 23 articles and of these approximately a quarter merely stated that quality might be or was an issue. Compared with many other empirically based disciplines this is a low level of interest. Of the papers that mention data quality more than 73% claimed that it is a significant problem and only about 9% that it was not, so it would seem that when we consider the problem of data quality it is generally viewed as being important.

So how do empirical analysts address this problem? The dominant approach is that manual inspection so as to provide some opportunity for triangulation. Other recommendations are preventitive techniques such as appropriate tool support for the data collection.

Presently there is little work to independently assess the quality of a given data set and where approaches such as the use of quality meta-data are deployed these are essentially surrogates for the level of missingness within the data set.

In the past, our community has been primarily concerned with the challenges of promoting the need for empirical evidence and simply obtaining data. From here we have moved onto issues such as replication and independent scrutiny of analyses through the public provision of data. And in this regard the PROMISE Group have been very active. At the time of writing (January, 2008) the repository contains 44 data sets. Consequently, researchers are beginning to have choices particularly if their concerns are to evaluate or compare different modelling or learning techniques. In which case does it matter which data sets we use? There are many issues here. One is the sampling bias of how data sets are "selected" to go into a repository and the attendant danger that our analysis techniques are skewed to those that perform well on some unrepresentative data sets[3]. Another is we may be wasting research effort on data sets that contain such levels of noise as to prevent meaningful conclusions. Where the noise leads to bias this will be particularly acute.

Thus we consider the problem of data quality to be pressing. Therefore we would make the following suggestions.

First, it may be that there is more data quality activity such as scrubbing or cleaning than is immediately visible. We would urge for more transparency in dealing with noise and that researchers should explicitly describe what procedures they have carried out prior to their analysis and often prior to reporting / archiving their data.

Second, we need to further investigate independent means (manual and automated) for assessing quality. We believe this has to be with respect to some purpose so it may be naïve to expect a single measure or indicator. Of course this is an extremely challenging problem that is compounded by the difficulty of knowing, in general, the "true" value of a data item. For this reason empirical techniques may have to be supplemented by simulation studies. Nevertheless it would be extremely helpful when using data archives to have some sense of the data quality, not least because of the increasing levels of separation of researcher and data collection environment.

Third, the use of sensitivity analysis should be commonplace. So even if we are unable to make very definitive statements concerning data quality we can at least comment on the minimal assumptions necessary concerning noise levels for conclusion stability.

Acknowledgements

We would like to thank Dr Tim Menzies for some fruitful discussions and valuable insights.

5. REFERENCES

[1] S. Biffl and W. J. Gutjahr. Using a reliability growth model to control software inspection. *Empirical Software Engineering*, 7(3):257–284, 2002.

[2] C. E. Brodley and M. A. Friedl. Identifying and eliminating mislabeled training instances. In *AAAI/IAAI, Vol. 1*, pages 799–805, 1996.

[3] C. Cappiello. *Data Quality and Multichannel Services.* PhD thesis, Politecnico di Milano, 2005.

[4] S. Counsell, G. Loizou, and R. Najjar. Quality of manual data collection in java software: an empirical investigation. *Empirical Software Engineering*, 12(3):275–293, 2007.

[5] P. B. Crosby. *Quality without tears: The art of hassle free management.* McGraw-Hill, New York, USA, 1984.

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

[3]Tim Menzies in a private communication observed that there is a similar problem amongst the machine learning community who use the UCI repository of data sets.

Paper Ref.	Year	Quantitative	Qualitative	Data collection	Manual noise checking	Empirical analysis of quality	Automatic noise checking	Data quality meta-data	Special analysis techniques	Is noise a problem?	Cost prediction?
[29]	2008	Y	N	N	N	N	N	Y	N	n.a.	Y
[25]	2008	Y	N	N	N	N	N	Y	N	n.a.	Y
[23]	2007	Y	N	N	Y	Y	Y	N	N	Y	Y
[27]	2007	Y	N	N	N	N	N	N	N	Y	N
[35]	2007	Y	N	N	Y	Y	Y	N	N	Y	N
[34]	2007	Y	N	N	Y	Y	N	N	N	Y	N
[16]	2007	Y	N	N	Y	Y	Y	N	N	Y	N
[4]	2007	Y	N	Y	Y	Y	N	N	N	Y	N
[20]	2006	Y	Y	N	Y	N	N	N	N	n.a.	N
[22]	2006	Y	N	N	N	Y	Y	N	N	Y	Y
[18]	2005	Y	N	N	Y	Y	Y	N	N	Y	N
[21]	2005	Y	N	N	Y	N	N	N	N	Y	Y
[31]	2005	Y	N	Y	N	N	N	N	N	Y	N
[30]	2004	Y	N	N	N	N	N	N	N	n.a.	N
[17]	2004	Y	N	N	Y	Y	Y	N	N	Y	N
[26]	2003	Y	N	N	N	N	N	N	Y	Y	Y
[32]	2003	Y	N	N	N	N	N	N	N	Y	Y
[1]	2002	Y	N	Y	Y	N	N	N	N	N	N
[37]	2000	Y	N	Y	N	N	N	N	N	N	N
[15]	1999	Y	N	Y	Y	Y	N	N	N	Y	N
[8]	1999	Y	N	N	N	N	N	N	N	Y	N
[13]	1998	Y	Y	Y	N	N	N	N	N	Y	N
[12]	1995	Y	N	Y	Y	N	N	N	N	Y	N

Table 4: Classification of Retrieved Papers

[7] A. M. Disney and P. M. Johnson. Investigating data quality problems in the psp. *Proceedings of the ACM SIGSOFT Symposium on the Foundations of Software Engineering*, pages 143–152, 1998. Cited By (since 1996): 2.

[8] N. E. Fenton and M. Neil. A critique of software defect prediction models. *IEEE Transactions on Software Engineering*, 25(5):675–689, 1999.

[9] D. Gamberger, N. Lavrač, and C. Grošelj. Experiments with noise detection algorithms in the diagnosis of coronary artery disease. In *IDAMAP-98, Third Workshop on Intelligent Data Analysis in Medicine and Pharmacology*, pages 29–33, Brighton, UK, 1998. University of Brighton.

[10] M. Gertz, M. T. Özsu, G. Saake, and K.-U. Sattler. Report on the dagstuhl seminar: "data quality on the web". *SIGMOD Record*, 33(1):127–132, 2004.

[11] T. P. Group. Promise data sets. Available: http://promisedata.org/repository/, Last accessed 10 January, 2008.

[12] R. Gulezian. Software quality measurement and modeling, maturity, control and improvement. *Proceedings of the IEEE International Software Engineering Standards Symposium*, pages 52–59, 1995.

[13] P. M. Johnson. Reengineering inspection. *Communications of the ACM*, 41(2):49–52, 1998.

[14] P. M. Johnson and A. M. Disney. Personal software process: A cautionary case study. *IEEE Software*, 15(6):85–88, 1998. Cited By (since 1996): 9.

[15] P. M. Johnson and A. M. Disney. A critical analysis of psp data quality: Results from a case study. *Empirical Software Engineering*, 4(4):317–349, 1999. Cited By (since 1996): 4.

[16] T. M. Khoshgoftaar and P. Rebours. Improving software quality prediction by noise filtering techniques. *Journal of Computer Science and Technology*, 22(3):387–396, 2007.

[17] T. M. Khoshgoftaar, N. Seliya, and K. Gao. Rule-based noise detection for software measurement data. In *IRI*, pages 302–307, 2004.

[18] T. M. Khoshgoftaar and J. D. Van Hulse. Identifying noise in an attribute of interest. In *ICMLA '05: Proceedings of the Fourth International Conference on Machine Learning and Applications (ICMLA'05)*, pages 55–62, Washington, DC, USA, 2005. IEEE Computer Society.

[19] B. Kitchenham. Procedures for performing systematic reviews (technical report tr/se-0401). Technical Report Technical Report TR/SE-0401, Keele University, Keele, UK, July 2004.

[20] J. Li, F. O. Bjørnson, R. Conradi, and V. B. Kampenes. An empirical study of variations in cots-based software development processes in the norwegian it industry. *Empirical Software Engineering*, 11(3):433–461, 2006.

[21] G. A. Liebchen and M. Shepperd. Software productivity analysis of a large data set and issues of confidentiality and data quality. *Proceedings of the*

11th IEEE International Software Metrics Symposium (METRICS'05), 00:46, 2005.

[22] G. A. Liebchen, B. Twala, M. Shepperd, and M. Cartwright. Assessing the quality and cleaning of a software project data set: An experience report. In *Proceedings of 10th International Conference on Evaluation and Assessment in Software Engineering (EASE)*. British Computer Society, 2006.

[23] G. A. Liebchen, B. Twala, M. Shepperd, M. Cartwright, and M. Stephens. Filtering, robust filtering, polishing: Techniques for addressing quality in software data. *First International Symposium on Empirical Software Engineering and Measurement (ESEM 2007)*, 0:99–106, 2007.

[24] R. J. A. Little and D. B. Rubin. *Statistical analysis with missing data*. John Wiley & Sons, Inc., New York, NY, USA, 1986.

[25] E. Mendes and C. Lokan. Replicating studies on cross- vs single-company effort models using the isbsg database. *Empirical Software Engineering*, 13(1), 2008.

[26] E. Mendes, I. Watson, C. Triggs, N. Mosley, and S. Counsell. A comparative study of cost estimation models for web hypermedia applications. *Empirical Software Engineering*, 8(2):163–196, 2003.

[27] P. Mohagheghi and R. Conradi. Quality, productivity and economic benefits of software reuse: a review of industrial studies. *Empirical Software Engineering*, 12(5):471–516, 2007.

[28] T. C. Redman. *Data Quality for the Information Age*. Artech House, Inc., Norwood, MA, USA, 1996. Foreword By-A. Blanton Godfrey.

[29] P. Sentas, L. Angelis, and I. Stamelos. A statistical framework for analyzing the duration of software. *Empirical Software Engineering*, 2008 (accepted), Available online: [http://www.springerlink.com/content/g82h031117l38336/].

[30] F. Shull, M. G. Mendoncça, V. Basili, J. Carver, J. C. Maldonado, S. Fabbri, G. H. Travassos, and M. C. Ferreira. Knowledge-sharing issues in experimental software engineering. *Empirical Software Engineering*, 9(1-2):111–137, 2004.

[31] R. Sison, D. Diaz, E. Lam, D. Navarro, and J. Navarro. Personal software process (psp) assistant. In *APSEC '05: Proceedings of the 12th Asia-Pacific Software Engineering Conference (APSEC'05)*, pages 687–696, Washington, DC, USA, 2005. IEEE Computer Society.

[32] E. Stensrud, T. Foss, B. Kitchenham, and I. Myrtveit. A further empirical investigation of the relationship between mre and project size. *Empirical Software Engineering*, 8(2):139–161, 2003.

[33] D. M. Strong, Y. W. Lee, and R. Y. Wang. Data quality in context. *Communications of the ACM*, 40(5):103–110, 1997.

[34] J. D. Van Hulse and T. M. Khoshgoftaar. A comprehensive empirical evaluation of missing value imputation in noisy software measurement data. *Journal of Systems and Software*, 2007.

[35] J. D. Van Hulse, T. M. Khoshgoftaar, and H. Huang. The pairwise attribute noise detection algorithm. *Knowledge and Information Systems*, 11(2):171–190, 2007.

[36] R. Y. Wang, H. B. Kon, and S. E. Madnick. Data quality requirements analysis and modeling. In *Proceedings of the Ninth International Conference on Data Engineering*, pages 670–677, Washington, DC, USA, 1993. IEEE Computer Society.

[37] A. Wesslén. A replicated empirical study of the impact of the methods in the psp on individual engineers. *Empirical Software Engineering*, 5(2):93–123, 2000.

APPENDIX

Table 4 describes the papers we identified in the systematic literature review that satisfied our inclusion criteria. The meaning of each column is as follows:

Paper ref. is the citation marker in the References.

Year refers to when the paper was formally published.

Quantitative is if the paper addresses data quality specifically in terms of quantitative data.

Qualitative is if the paper addresses data quality specifically in terms of qualitative data.

Data Collection indicates if the paper addresses or makes suggestions concerning data quality at the collection stage, i.e. a preventative strategy.

Manual Noise Checking refers to if the paper suggests how to, or actually, addresses data quality problems by some manual noise checking procedure.

Empirical Analysis of Quality indicates if the paper contains an empirical analysis of the extent of any noise problems.

Automatic Noise Checking indicates if the paper proposes or evaluates some automated procedure for noise detection.

Data Quality Meta-Data indicates if the paper proposes or utilises meta-data, e.g. some data sets like ISBSG contain case quality variables that can be used to filter out 'low' quality cases from the analysis.

Special Analysis Techniques refers to explicitly using or recommending particular analysis techniques that are robust to data quality problems.

Is Noise a Problem? indicates how significant the paper considers noise to be for their particular analysis. The value "n.a." indicates a mere flagging of data quality as a potential problem for the analysis.

Cost Prediction indicates if the paper is concerned with data in the cost prediction domain. This is included because we believe the problems of noise to be particularly acute in this problem domain.