H. Petersson and C. Wohlin, "Evaluation of using Capture-Recapture Methods on Software Review Data", Proceedings Empirical Assessment and Evaluation in Software Engineering, EASE99, Keele, UK, 1999.

Evaluation of using Capture-Recapture Methods on Software Review Data

Håkan Petersson and Claes Wohlin
Dept. of Communication Systems, Lund University
Box 118, SE-221 00 LUND, Sweden
{hakan.petersson, claes.wohlin}@tts.lth.se

Abstract

Software defect content estimation is important, to control and ensure software quality. One possible method to achieve this is by applying capture-recapture methods. This type of methods can be applied on data collected from reviews, and it can be used to estimate the remaining number of defects after a review. This paper focuses on replicating a previous study made by other researchers. The replication is concerned with looking at different estimation methods and the ability to select a good enough estimation method. The evaluation in this paper is based on three data sets from three different types of documents: a text document, code and a requirements document. It is concluded that no estimation method is generally superior, and the best method is dependent on several factors, including the actual data sets and the number of reviewers. Moreover, it is concluded that the selection method proposed in the previous study is not superior to one specific model. Thus, the selection procedure is unable to pinpoint the best estimation method in our data sets.

1. Introduction

Inspections [Fagan76] and reviews [Humphrey95] become more and more accepted in the software industry as a cost-effective way of improving software quality. However, normally the inspection methods do not support estimation of the remaining defect content. Capture-recapture methods, see Section 2, provide this opportunity. This paper addresses the challenge of estimating the number of defects after a review.

Capture-recapture methods aim at making estimations of the remaining defect content by using the information about what the individual reviewers have found. The knowledge about the remaining defect content allows us to control the software process through entry and exit criteria in the software process. Moreover, it allows us to direct our efforts at the most critical parts from a defect perspective.

Capture-recapture methods have traditionally been used in population estimations in biology [Otis78]. Researchers in software engineering have realised that this type of approach can be applied to estimate the “population” of defects. In other words, esti-

mating the total number of defects and hence allowing us to determine the remaining number of defects.

The work conducted with respect to capture-recapture and software engineering is so far rather limited, but there is a growing interest both within the research community and from industry. The work started with some initial studies in software engineering, which are presented in [Eick92] and [Vander Wiel93]. A follow-up study dividing defects into different classes as a means for improving the estimates is presented in [Wohlin95]. An evaluation of several capture-recapture methods applied to software engineering is presented in [Briand97]. Some new approaches for performing estimations based on curve fitting approaches are introduced in [Wohlin98]. A selection procedure choosing between two estimation methods is presented in [Briand98], and further discussed below. Finally, an experience-based approach to estimation is presented in [Runeson98].

One of the major challenges in the area is to identify which method to use for estimation of defect content. Several methods have been proposed in the literature, both in software engineering and in other areas such as population estimation in biology. The estimation methods include statistical methods based on, for example, Maximum-Likelihood and Jackknife estimations. The Jackknife estimator is briefly discussed in Section 2.1, since it is included in the selection procedure proposed by Briand et al. in [Briand98].

The traditional estimation models, for example the Maximum-Likelihood and the Jackknife, are evaluated in [Briand97]. In [Wohlin98], two new estimation approaches are added to the potential methods to use. Thus, several methods exist, and the challenge is to determine which method to use. One of them is the Detection Profile Method (DPM) which is briefly introduced in Section 2.2, because it, as the Jackknife estimator, is used in the replicated study [Briand98]. Unfortunately, no model is superior in general, hence it is important to be able to choose the best or a good enough method in a specific situation. A selection procedure is proposed and evaluated for one data set in [Briand98].

The objective of this paper is to evaluate and replicate the previous study and improve the approach suggested in [Briand98]. Replications are essential, as they will allow us to gain a better understanding and more general valid results, than stand-alone investigations.

The paper is organised as follows. A short description of the capture-recapture concept and the methods Jackknife and DPM is given in Section 2. Section 3 provides a brief summary of the, for our evaluation, relevant parts of the paper by Briand et al. [Briand98]. Some improvement proposals related to one step in the selection procedure are presented in Section 4, and evaluated in Section 5. The selection procedure is evaluated using three data sets in Section 6. Finally, a discussion and conclusions are presented in Section 7 and 8, respectively.

2. Capture-Recapture

Capture-recapture methods aim at making defect estimations by using information about the findings of the individual reviewers. One, for capture-recapture, important

characteristics of the review is the overlap among the defects found by the different reviewers. The size of the overlap may stretch from no overlap, where the reviewers have found totally disjunct sets of defects, to complete overlap, where the reviewers have found exactly the same defects. It is intuitively acceptable to assume that there would be more defects left in the reviewed document when there is no overlap than when there is complete overlap. This intuitive concept is used in capture-recapture methods to estimate the number of defects. However, the most probable number of remaining defects depends on the assumptions made when developing the statistical model. The different assumptions lead to different models and within the model, there can be different methods to deal with the statistics. Brief descriptions of two methods are given in the next two subsections.

2.1 Jackknife

One of the capture-recapture models, assumes that the probability of a specific defect being found is independent of the ability of the different reviewers, but the detection probability differs between defects. One method, within this model, is the Jackknife method. [Otis78]

Jackknife uses the overlap information by calculating the number of total defects from the frequencies of how many defects that were found exactly by 'i' reviewers, denoted f_i. The Jackknife can be calculated with different orders. The most basic formula, order one, is:

$$\hat{N} = D + \frac{k-1}{k} f_1$$

where D denotes the number of unique defects, $\hat{N}$ denotes the estimated total number of defects and k denotes the number of reviewers.

2.2 DPM

Another method, a curve fitting method denoted the Defect Profile Method, is proposed in [Wohlin98]. The method is based on plotting the defects versus the number of reviewers that have found a specific defect, see Figure 1.

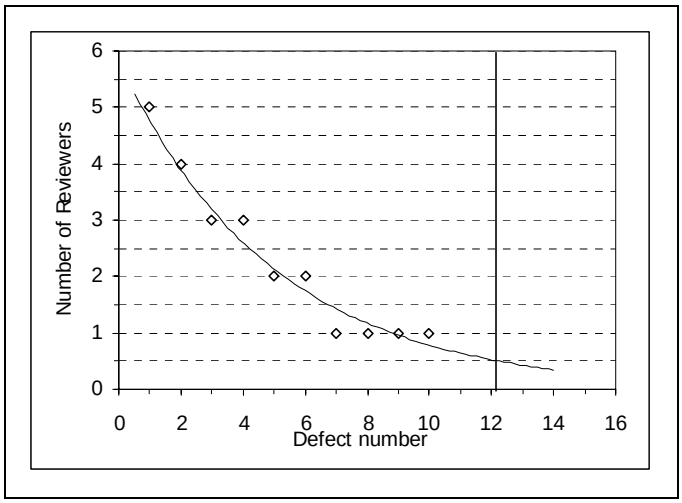

FIGURE 1. Example of the curve fitting method, DPM.

It is observed that the form of the curve resembled an exponentially decreasing function. The estimate of the remaining number of defects is obtained from where the exponential curve is equal to 0.5, see Figure 1. The proposed method is evaluated using two data sets, and it is concluded that the method behaved reasonably. The mean relative error is similar to other methods, but the variance is slightly smaller.

3. A Selection Method

3.1 Introduction

The focus in [Briand98] is on identifying a procedure for finding a good enough estimation method for the remaining number of defects. A problem identified is that most estimators produce extreme estimates or outliers for certain data sets. To overcome this problem, it is essential to choose a method, which for a specific data set produces a reasonable estimate. The observation made, by Briand et al. in [Briand98], is that by combining an enhanced version of one of the curve fitting methods [Wohlin98], with a Jackknife estimator, the outliers are avoided. This is the observation for their data set, and one of the important issues in this paper is to confirm or reject their results.

To summarize, the focus in [Briand98] is on the enhanced curve fitting method and on the proposed selection procedure. These two issues are further discussed in Section 3.2 and Section 3.3.

3.2 The Enhanced Defect Profile Method (EDPM)

As stated in Section 2.2, the mean relative error of the DPM is similar to other methods, but the variance is slightly smaller. These encouraging results were the starting point when Briand et al. developed their selection procedure. Briand et al. observed, however, that the DPM is unsuitable for certain data sets. In particular, the problem becomes apparent when none of the defects were found by only one reviewer, hence the plot (c.f. Figure 1) does not resemble an exponential function, or at least it does not decrease as fast as expected. Thus, Briand et al. suggest a criterion which selects

between an exponential function and a linear function. Letting f_i be the number of defects found by exactly i reviewers, they define the following criterion (denoted strict order criterion):

$$f_1 \geq f_2 \geq f_3 \geq \ldots \geq f_n;\ \text{n = the number of reviewers} \qquad \text{(Eq. 1)}$$

Hence, if the criterion is fulfilled, Briand et al. choose the originally proposed method, i.e. the DPM with an exponential function. If the criterion is not fulfilled, they use the linear function. They create a criterion, which determines whether the data actually look like an exponential function. Thus, they chose a linear function as a complement. Other functions are possible, but this remains to be investigated.

The method with the choice, based on the strict order criterion, between the exponential function and the linear function is referred to as the Enhanced Defect Profile Method (EDPM) in [Briand98]. This method forms one important input to their proposed selection procedure.

3.3 The Selection Procedure

The selection procedure proposed by Briand et al. is based on the EDPM. First, the EDPM is applied using the strict order criterion to decide which function to fit to the data. The function is fitted to the data and evaluated using an R^2 approach, i.e. the goodness of fit is evaluated by looking at the correlation between the data points and the fitted function. The EDPM estimate is trusted if R^2 is greater than 0.8, and is significant on the 0.01 level. This is made to ensure that the fit is good enough. If it turns out that the EDPM is not applicable according to the test, the Jackknife estimator is chosen. This decision is taken based on the evaluation presented in [Briand97], where the Jackknife estimator turned out to be the best one.

The Jackknife estimator basically assumes that defects may have different probabilities of being detected, and that all reviewers have the same detection probability. This is probably not completely true in software development, where it is highly likely that different individuals based on, for example, background and experience have different detection probabilities. Although not being completely true, the results in [Briand97] indicate that the Jackknife estimator gives better results than most other estimators do. This issue has, however, to be further investigated. The actual estimation is carried out using a formula derived from the basic assumptions of the method.

4. EDPM Modifications

When studying the criterion, Eq. 1, used by the EDPM to choose between a linear curve and an exponential curve it sometimes appears to be too strict. As it is formulated, the criterion sometimes provides a bit too hard restraints on when to choose the exponential curve. The criterion can be rewritten as:

$$\begin{aligned} f_1 &\geq f_2 \\ f_2 &\geq f_3 \\ &\ldots \end{aligned} \qquad \text{(Eq.2)}$$

It is enough that one of all the parts in Eq. 2 fails, to make the EDPM use a linear curve instead of an exponential curve. One example of this, illustrated in Figure 2, is when one of the f_i values, with a high i index, is zero instead of one. The two data sets are almost identical. The only thing that differs is the f_4 value. The left diagram has $f_4 = 1$ and the right has $f_4 = 0$. Because of this, the right data set fails to pass the strict order criterion while the left one passes. The right one, however, still resembles an exponential curve more than a straight line.

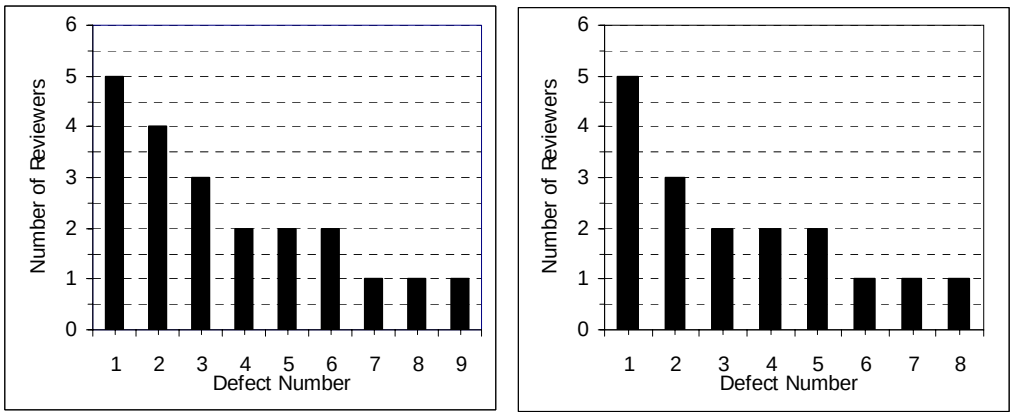

FIGURE 2. Example of criterion accept (left) and reject (right).

In [Briand98], the largest underestimates made by the EDPM are picked from the linear fit estimator. The linear fit has altogether a tendency of underestimating more than the DPM. Although the EDPM manages to handle some of the overestimates made by the DPM, it still provides too many underestimates. This leads to the conclusion that the EDPM might be helped by a slight change to the criterion.

To investigate this conclusion, a new version of the EDPM, denoted EDPM2, is constructed. In this we make a very small change in the criterion towards not choosing the linear fit as many times. The purpose of the change is to allow for one of the parts of the divided criterion, Eq. 2, to be less strict. This is made by subtracting 1 from the right side of one of the criterion parts, for example, changing the second line in Eq. 2 to $f_2 \geq f_3\text{-}1$. This would allow a data set to have a single flaw in its compliance to the strict order criterion, for example, the right one in Figure 2.

The modified variant of the EDPM is used as one of the candidates in the next section when investigating the parts, which the selection procedure proposed by Briand et al. consists of. The modification made to the criterion is investigated to evaluate whether a minor change to the criterion has any effect on the estimations of the EDPM.

5. An Evaluation of the EDPM

The main result in [Briand98] is the selection procedure. As described earlier, this selection procedure has two possible outcomes. One, choose a curve fitting approach, EDPM, or two, choose a statistical estimator, Jackknife. Before evaluating the selection procedure itself, the first part included in the procedure, is studied. As for the statistical part, we follow the recommendation made by Briand et al. and keep Jackknife as the statistical estimator based on the results presented in [Briand97]. Four curve-fit-

ting estimators are evaluated on three different data sets to assess which of the estimators performs best in general. The estimator candidates are:

- LF, curve fitting using a linear curve.
- DPM, curve fitting using an exponential curve.
- EDPM, the combination of the two estimators above, proposed by Briand et al., [Briand98].
- EDPM2, our modification of the EDPM using a slightly less strict criterion.

The two EDPM estimators are constructed using either a linear or an exponential function depending on the result of the selection criterion. By comparing all four, it can be evaluated whether the combined estimators perform better than its parts.

5.1 Data Sets

Three different review data sets are used to carry out the evaluation and the data sets are picked from three different studies [Wohlin95], [Runeson98] and a recent replication of [Basili96] performed at Lund University. The replication is based on the lab package provided by University of Maryland [UMD98]. The three data sets represent data gathered from reviews in different contexts. The data sets are:

- Text document: A text document written in English was reviewed by 22 reviewers to find grammatical errors and spelling errors[1] [Wohlin95]. Context: Documents written in English.
- Code: Five programs written in C, where the original defects had been reinjected, were reviewed by a total of eight reviewers [Runeson98]. Each person reviewed three programs that were randomly assigned to them. Thus, all programs were not reviewed by five reviewers, therefore this group size is not applied on the this data set. Context: Code written in C.
- Requirements document: This data set is produced by a replication in [Regnell99] of the Perspective Based Reading (PBR) experiment first presented in [Basili96]. The reviewers in the replication were 30 final year Master students at Lund University. The students were divided randomly between the two requirements documents. There are, however, some questions raised whether capture-recapture methods can or may be used on PBR data. The objective of PBR is to make sure that the reviewers focus on different aspects. Thus, the objective of PBR is contradictory to the assumptions of the capture-recapture methods. In this data set, however, it could not be significantly shown that there were any differences in what defects were found by reviewers using different perspectives. Therefore, the capture-recapture methods can be applied to this data set. Context: Requirements document written in English.

To create a larger basis for the evaluation, each data set, as in [Briand97] and [Briand98], is combined into a large number of so called virtual reviews. For each group size of reviewers, all possible combinations of the reviewers are run as a single

1. Later in the text, both errors and defect are called defects.

review creating one sample of the data set. The number of virtual reviews for each data set are listed in Table 1.

TABLE 1. Number of virtual reviews for the different contexts.

	Group size			
Data Set	**2**	**3**	**4**	**5**
Text	231	1540	7315	26334
Code	46	44	21	N/A
Req. Doc.	210	910	2730	6006

5.2 Results

The results are evaluated using the same approach as in [Briand98]. The objective is to make the results comparable.

In Figure 3 to Figure 5, boxplots of the results from the three data sets are presented.

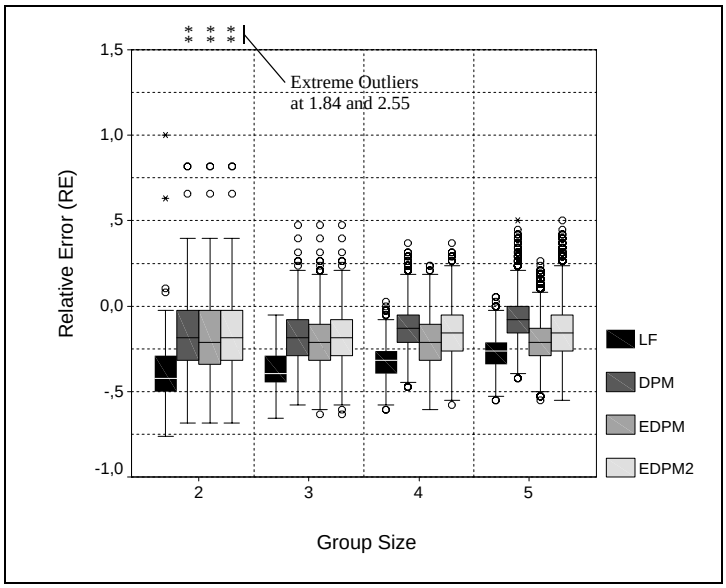

FIGURE 3. Text Document.

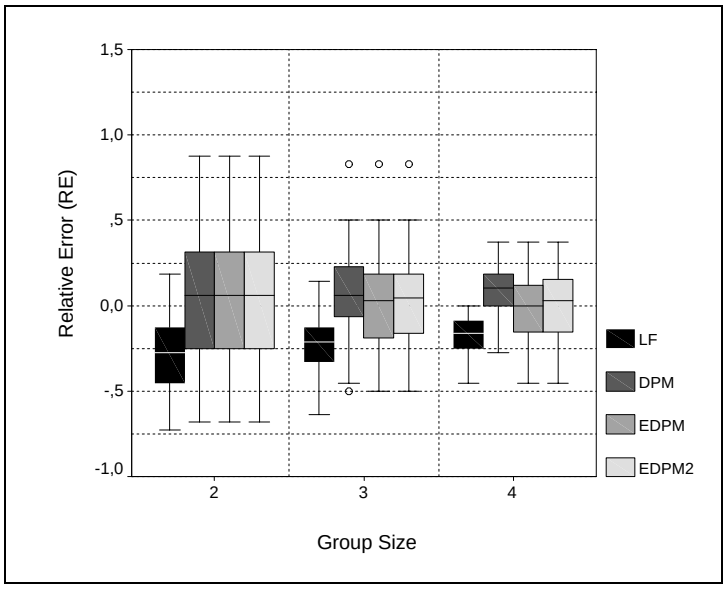

FIGURE 4. Code.

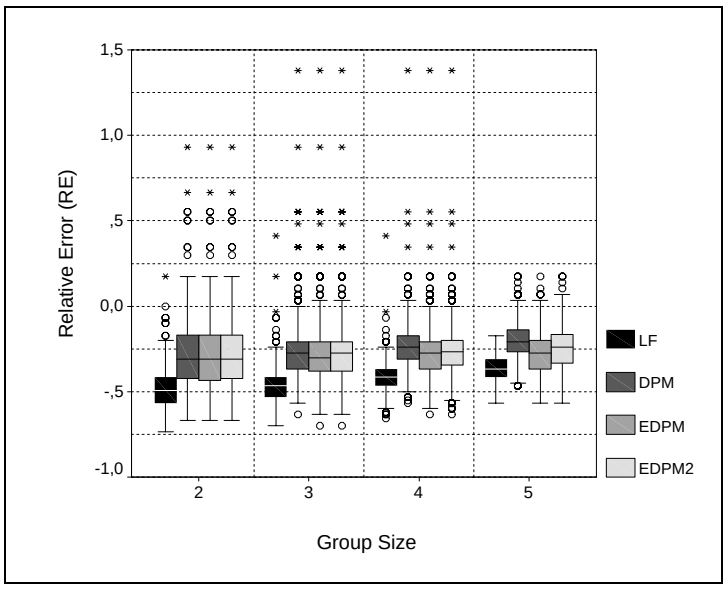

FIGURE 5. Requirements Document.

The boxplots show the results of all estimations in terms of the relative error (RE):

$$RE = \frac{\text{Estimated number of defects} - \text{Actual number of defects}}{\text{Actual number of defects}}$$

The box in the boxplots extends from the 25th percentile, lower quartile, to the 75th percentile, upper quartile, of the estimates and the whiskers (lines extending from the

boxes) show the limit for non-outlier values. Outlier values have the following characteristics:

$$Outlier > UQ + 1.5(UQ - LQ)$$
$$or$$
$$Outlier < LQ - 1.5(UQ - LQ)$$

$$\begin{cases} UQ - \text{Upper quartile} \\ LQ - \text{Lower quartile} \end{cases}$$

The outliers are marked with circles. Extreme outliers follow the same characteristics but with 1.5 changed to 3.0 instead. They are marked with stars.

When evaluating estimators there is often a trade-off between finding an estimate with a good variance and a good bias. Briand et al. use three different measures to illustrate variance and bias:

- absolute median relative error (Bias)
- maximum outlier value (Variance)
- quartile range, range between 25th and 75th percentile (Variance).

All of these measures can be found in a boxplot, see Figure 6

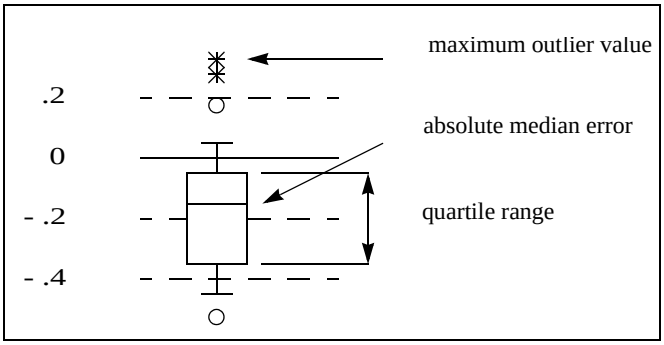

FIGURE 6. Bias and variance measures

When studying the boxplots of the text document and the requirements document, Figure 3 and Figure 5, it can be seen that the EDPM2 has slightly lower bias than EDPM. The modification of the criterion has resulted in avoiding some of the linear fit's lower values, unfortunately not as many as would have been preferred. At the same time, when looking at larger reviewer group sizes, the EDPM2 tends to pick out too many of the DPM's high values resulting in that the maximum outlier variance increases. If we study the LF's and the DPM's performance we find, in common for all the documents, that LF almost all the times underestimate and therefore has a large bias. The quartile variance, however, is very low but having such a large bias, we are not helped by this. The DPM in its turn has smaller bias than the EDPM in both the text document and the requirements document though it has slightly larger maximum outlier variance in the text document. In the code data, the DPM has larger bias than both the EDPM and the EDPM2.

The decision of which estimator is best depends on both the data set and the group size hence it is difficult to find one method which is superior. It also depends on what we want to achieve, low variance and few outliers or a low bias. A mix of these characteristics would have been preferable but not shown in any case. In the light of this, there is no strong enough reason to replace the EDPM as the curve fitting part of the selection procedure.

A best estimator can be picked, based on the outlier and variance measures. However, it might not be the best when it comes to the selection procedure. The best estimator for the selection procedure is the estimator that works best together with the Jackknife estimator. An estimator that performs well when the Jackknife estimator does not and vice versa.

6. The Replication

In the previous section it is shown that although the EDPM does not show any major advances over the DPM and the EDPM2 it is not noticeably worse either. This leads to the conclusion that we can proceed to make a replication and evaluation of the selection procedure without any modifications to its design. It contains the same parts as in [Briand98], namely the EDPM and the Jackknife estimator.

The same approach as in section Section 5 is used for evaluation. We compare the performance of the combined estimator, the selection procedure in this case, with its parts. The estimations for every virtual review are made and the relative errors are presented in boxplots.

6.1 Results

The results are presented in Figure 7 to Figure 9, which show boxplots from the data sets. The selection procedure are denoted SELPROC, in the boxplots.

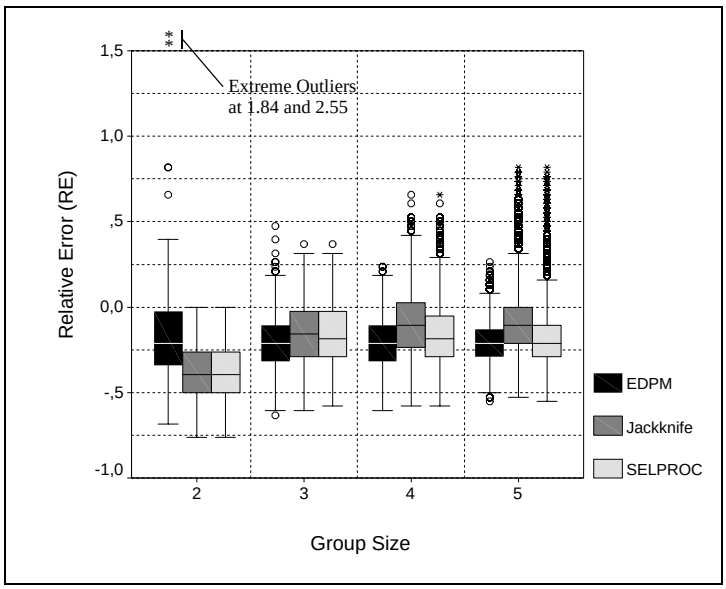

FIGURE 7. Text Document.

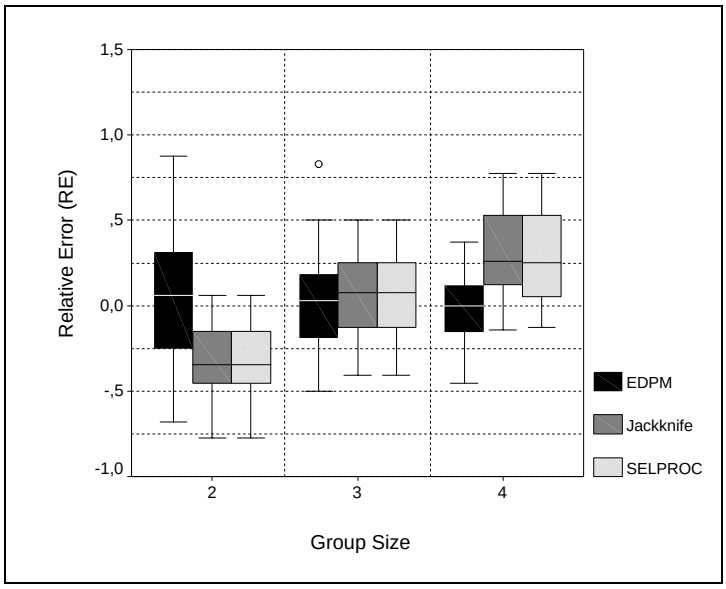

FIGURE 8. Code.

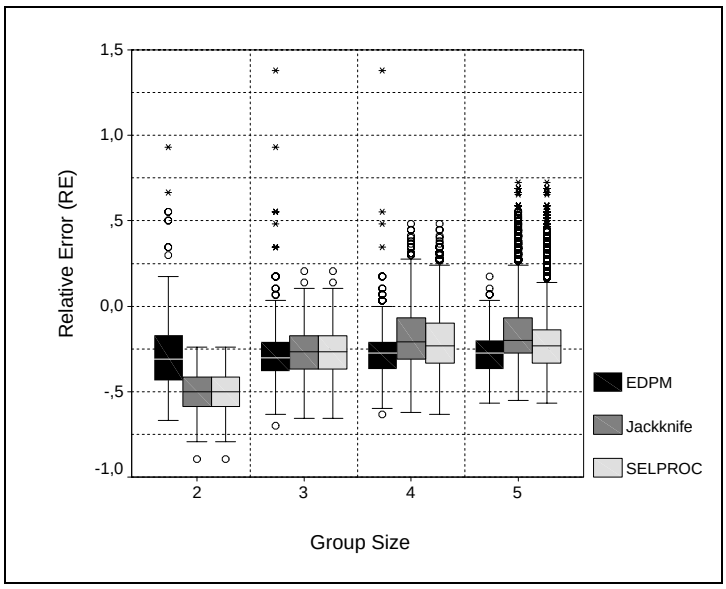

FIGURE 9. Requirements Document.

When studying the boxplots, we find that for group sizes of two and three reviewers the selection procedure almost exclusively picks the Jackknife estimates. The Jackknife's and selection procedure's boxes are identical with only one exception. The EDPM shows better bias, nevertheless, it has much higher maximum outlier variance than the other two.

It is not until group sizes of four reviewers that we see clear tendencies that the EDPM has any effect on the selection procedure. However, even then, and for five reviewers, the selection procedure seems to select Jackknife quite often. The EDPM lowers the bias of the selection procedure but in the cases of the text document and the requirements document this is an impairment. The big problem, however, for four and five reviewers is maximum outlier variance. The selection procedure seems to pick most of

the outliers produced by the Jackknife estimator. This is unfortunate because the effect that the selection procedure aimed for was to reduce the number of outliers.

The results found here are also discernible in [Briand98]. The selection procedure follows the Jackknife estimator for small group sizes and are not influenced by the EDPM until group sizes of four or five. This is actually according to expectations since the few number of data points, due to that we have few reviewers, means that it is difficult to obtain a significant fit with any function. The major difference compared to [Briand98] is that with our data sets the selection procedure is unable to avoid the outliers. In [Briand98] the selection procedure managed to remove the outliers for five reviewers. We are, however, unable to replicate this result with our data sets.

It has to be taken under consideration that in both the text and requirements documents there are many estimations made. The number of estimates that are summarized in the boxplots differ quite a lot depending on the group size, see Table 1. This fact makes a comparison within the same document type difficult. However, some observations can still be made. The bias of the EDPM changes very little with increasing group size compared to the Jackknife estimator and the selection procedure. This tendency can also be seen in [Briand98].

In the code document, the EDPM shows extremely good bias while the bias of the Jackknife estimator and the selection procedure varies with the group size.

An interesting detail is found when studying the bias levels within one group size, for example, when comparing EDPM, the Jackknife estimator and the selection procedure for groups of size three. The relative position of the bias levels can be viewed as a pattern, and it is interesting to note that the pattern seems to change with the group size. Moreover, the same pattern is visible for all three document types, and it is also discernible in [Briand98]. This may indicate that the group size is actually one important aspects to include in the selection of the estimator to use.

As a conclusion, the selection procedure shows no major improvements in comparison with its parts in our investigation.

7. Discussion

In Section 4, we started with making modifications to the criterion used by the EDPM because we saw examples where the criterion was too strict when determining which curve fits the specific data best. Even if there were small improvements in some cases, the modification did not succeed to become an improvement in general. It is worth noticing though, that even such a small change managed to produce noticeable effects in the boxplot. Especially in the text document reviewed by four reviewers, when EDPM2 picked all instead of some of the DPM's overestimating outliers. This leads us to that the choice of criterion is important.

The criterion used by the selection procedure appears not to be enough. The criterion and use of a linear curve was introduced in [Briand98] and they pointed out a group of cases where the DPM tends to make highly inflated estimates. This group is when f_1 is equal to zero. The linear curve then managed to make a better estimate. However, lets turn this case around and think of what would happen if f_1 instead is very large, i.e. we

have very little overlap among the defects found. This would result in that both the DPM and the LF estimate a value very close to the total number of found defects.[1] This goes against the basic idea of capture-recapture, i.e. if there are little overlap among the defects found, we should estimate a large number of remaining faults. So the criterion and the use of only a linear fit in collaboration with the DPM does not cover all the cases. Still, the selection procedure might take care of this if EDPM fails on achieving significance and thereby the choice will fall on the Jackknife estimator instead.

If we look at our results concerning the selection procedure that was evaluated in Section 6, we find that the selection procedure does not manage to remove all the outliers as in [Briand98]. Mostly because it follows the Jackknife estimator most of the times and fail to avoid the cases when the Jackknife estimator produces overestimating outliers.

The selection procedure is, built from different criteria picking estimators that already exist. It is a way of trying to find the estimator that fits the specific data. The concept of choosing, automatically limits our opportunity to produce a good estimate. The best we can hope for is that we all of the time choose the best method. This is, however, a very difficult task. The problem is particularly obvious when studying outliers. The selection procedure is based on two estimators that perform well in general. Nevertheless, if we come across the case where both estimators produce outliers we can pick nothing but an outlier. There are a couple of different ways to work around this:

- Use more than two estimators to choose from. This would increase the chance of having one estimator that does not produce an outlier in that specific case. However, this would lead to increased problems in which to choose and especially find the one that produces the result that we want.
- Find or develop estimators that complement each other in a way that they do not produce outliers at the same time.
- Diverge from the pure idea of selection by creating new interpolated estimates from two estimators where one always estimates high and the other always estimates low. A variant of this would be to identify one estimator, which is likely to make an overestimate, and another, which is likely to make an underestimate, and then use this knowledge to modify or combine the estimates.

All of these approaches have their difficulties but should be investigated in the future research.

Another interesting approach is if we can identify the special cases that produces outliers, as for the example Briand et al. did with $f_1 = 0$. If we could do this then it is a question of either saying, the estimator can not make an estimate, or choose another special way of handling the case. There seems to be a lot to gain to study the reasons of why outliers appear. Do they appear at random or are there patterns to be recognised?

In Section 6, it is noted that there seems to be a pattern of how the bias levels of the estimators are placed relative each other. In our three data sets as well as in the one used in [Briand98], general observations such as, the DPM has a larger mean than the

1. In some cases even below this number.

EDPM and that the linear function has always the lowest mean, can be seen. This gives some knowledge of how the different estimators lie relative each other though we do not know where they lie on an absolute scale.

8. Conclusions

The importance of knowing the number of remaining defects after a review is undisputed. It is, however, a difficult task to estimate the number of defects. The research in the field has so far been unable to come up with a method, which is superior to the others. The methods seem to be very dependent on the actual data set being studied and the number of reviewers involved. The results in this paper provide some new insights, including:

- The selection procedure proposed in [Briand98] does not perform better than its parts for our data sets,
- no estimation model can be depicted as being generally better than the others,
- the EDPM may be suitable for few reviewers although never selected by the selection procedure.

The results show one important aspect of performing replications. In this particularly case, the results were not the same hence indicating that the results obtained by Briand et al. may be a result of the actual data set. Moreover, it is clear that more work is needed in this area in order to improve the existing approaches. Several challenges have been identified in this paper and they should be addressed in the future research.

Acknowledgement

We would like to thank Thomas Thelin at the Department of Communication Systems, for his valuable comments on this paper and help during the implementation of the algorithms. We would also like to thank Bernd Freimut for taking his time to help us make this replication possible by helping us with some details in the paper by himself and his colleagues at Fraunhofer Institute of Experimental Software Engineering, [Briand98]. This work was partly funded by The Swedish National Board for Industrial and Technical Development (NUTEK), grant 1K1P-97-09673.

References

[Basili96] V. R. Basili, S. Green, O. Laitenberger, F. Lanubile, F. Shull, S. Sørumgård and M. V. Zelkowitz: “The Empirical Investigation of Perspective-Based Reading”, *Empirical Software Engineering: An International Journal*, Vol. 1, No. 2, pp. 133-164, 1996.

[Briand97] L. Briand, K. El Emam, B. Freimut and O. Laitenberger: ”Quantitative Evaluation of Capture-Recapture Models to Control Software Inspections”, In *Proc. of the 8:th International Symposium on Software Reliability Engineering*, pp. 234-244, 1997.

[Briand98] L. Briand, K. El Emam and B. Freimut: “A Comparison and Integration of Capture-Recapture Models and the Detection Profile Method”. In Proc. of the 9:th International Symposium on Software Reliability Engineering, 1998.

[Eick92] S. Eick, C. Loader, D. Long, L. Votta and S. Vander Wiel: “Estimating Software Fault Content Before Coding”. In *Proc. of the 14th International Conference on Software Engineering*, pp. 59-65, 1992.

[Fagan76] M. E. Fagan: “Design and Code Inspections to Reduce Errors in Program Development”, *IBM System Journal*, Vol. 15, No. 3, pp 182-211, 1976.

[Humphrey95] W. S. Humphrey: *A Discipline for Software Engineering*. Addison-Wesley, USA 1995.

[Otis78] D. Otis, K. Burnham, G. White and D. Anderson: “Statistical Inference from Capture Data on Closed Animal Populations”. *Wildlife Monographs*, No. 62, October 1978.

[Regnell99] B. Regnell, P. Runeson, and T. Thelin “Are the Perspectives Really Different? - Further Experimentation on Scenario-Based Reading of Requirements”, Technical Report CODEN: LUTEDX(TETS-7172) / 1-38 / 1999 & local 4, Dept. of Communication Systems, Lund University, 1999.

[Runeson98] P. Runeson and C. Wohlin: “An Experimental Evaluation of an Experience-Based Capture-Recapture Method in Software Code Inspections”, *Empirical Software Engineering: An International Journal*, Vol. 3, No. 4, pp. 381-406, 1998.

[UMD98] V. R. Basili, S. Green, O. Laitenberger, F. Lanubile, F. Shull, S. Sørumgård and M. V. Zelkowitz, Lab package for the Empirical Investigation of Perspective-Based Reading, available on the WWW: http://www.cs.umd.edu/projects/SoftEng/ESEG/manual/pbr_package/manual.html.

[Vander Wiel93] S. Vander Wiel and L. Votta: “Assessing Software Designs Using Capture-Recapture Methods”, In *IEEE Transactions on Software Engineering*, Vol. 19, No. 11, pp 1045-1054, November 1993.

[Wohlin95] C. Wohlin, P. Runeson and J. Brantestam: “An Experimental Evaluation of Capture-Recapture in Software Inspections”, In *Journal of Software Testing, Verification and Reliability*, Vol. 5, No. 4, pp. 213-232, 1995.

[Wohlin98] C. Wohlin and P. Runeson: “Defect Content Estimations from Review Data”, In *Proc. of the 20th International Conference on Software Engineering*. pp. 400-409, 1998.