A Comparison and Integration of Capture-Recapture Models and the Detection Profile Method

Lionel C. Briand, Khaled El Emam, Bernd G. Freimut
Fraunhofer Institute for Experimental Software Engineering
Sauerwiesen 6, D-67661 Kaiserslautern-Siegelbach, Germany
{briand, elemam, freimut}@iese.fhg.de

International Software Engineering Research Network Technical Report ISERN-98-11

A Comparison and Integration of Capture-Recapture Models and the Detection Profile Method

Lionel C. Briand, Khaled El Emam, Bernd G. Freimut
Fraunhofer Institute for Experimental Software Engineering
Sauerwiesen 6, D-67661 Kaiserslautern-Siegelbach, Germany
{briand, elemam, freimut}@iese.fhg.de

Abstract

In order to control inspections, the number of remaining defects in software artifacts after their inspection should be estimated. This would allow, for example, deciding whether a reinspection of supposedly faulty artifacts is necessary. Several studies in software engineering have considered capture-recapture models for performing such estimations. These models were initially developed for estimating animal abundance in wildlife research. In addition to these models, researchers in software engineering have recently proposed an alternative approach, namely the Detection Profile Method (DPM), that makes less restrictive assumptions than some capture-recapture models and that show promise in terms of estimation accuracy. In this study, we investigate how to select between these two approaches for defect content estimation. As a result of this investigation we present a selection procedure taking into account the strength and weaknesses of the two methods. A weakness known for capture-recapture models is that they tend to provide extreme under/over estimation. The existence of such extreme outliers can discourage their use because their consequences in terms of wasted effort or defect slippage can be substantial, and therefore it is not clear whether a particular estimate can be trusted. The evaluation of our selection procedure with actual inspection data indicates that this selection procedure provides the same accuracy as capture-recapture models alone and DPM alone, and most importantly does not exhibit extreme over/under estimation. Thus, this selection procedure can be used in practice with a high degree of confidence since its estimates are not likely to exhibit extreme estimation error.

Keywords: capture-recapture model, defect content estimation, software inspections.

1. Introduction

The construction of reliable software requires that the number of defects introduced and propagated during development is minimized. An important and widely-applied technique for achieving this objective is software inspections. The benefits of inspections stem from the fact that defects are detected early after their insertion, and that rework costs are reduced [10]. Also, inspections find defects that are less likely to be found using other defect detection techniques [4][14]. Thus it is possible to (1) increase reliability by removing defects and (2) reduce costs and cycle time by saving on rework.

Various inspection processes and reading techniques have been proposed to increase the effectiveness of inspections. However, regardless of the effectiveness of the reading technique employed or other changes to the inspection process, without any quality control on the inspection process, the use of inspections is destined to be suboptimal. One approach to optimize the effectiveness of inspections is to reinspect an artifact that is presumed to still have high defect content. The reinspection decision criterion could be based on the number of remaining defects after an inspection, which can be estimated with defect content models.

A promising approach for the estimation of defects in an inspected software artifact was proposed by Eick et. al. [7]. They applied Capture-Recapture (C/R) Models, which are used in biology to estimate animal abundance, to predict the number of defects in a software artifact after design inspections. Based on this idea, several researchers have explored this approach further ([17], [8], [19], [13], [9], [6], [1]).

A recent, comprehensive evaluation study of C/R Models [6] identified the type of model that seems to be the most accurate and usable in the context of software inspections. However, this 'best' model still exhibited some characteristics that make it difficult to use in practice: in some cases it exhibits extreme outliers in its estimates. This behavior does not give confidence to the user that the model is always providing reasonable estimates, and therefore could discourage its use.

A complementary approach for estimating defect content has been recently proposed by Wohlin and Runeson [18].

Their approach is based on plotting defect data according to some criterion. Based on this plot, a curve is fitted through the data points. Using the parameter of this fitted curve the number of defects in an artifact can be estimated.

In this paper we propose some strategies for enhancing the most promising approach proposed by Wohlin and Runeson, called the Detection Profile Method (DPM). We evaluate these proposed enhancements and identify one that improves over the DPM. We then compare the improved DPM to the best C/R Model found in [6], and propose a selection procedure that selects between the DPM and the best C/R Model. An evaluation of this selection procedure indicates that it provides the same accuracy as C/R Models but has an important characteristic that makes it more usable in practice: it does not exhibit extreme outliers in its estimates. This advantage should give confidence in the estimates selected by the procedure.

In the following section we provide an overview of the two defect content estimation methods that we consider in this paper, and then state our research objectives. In Section 3 we describe our empirical research method and the data sources used for evaluation. This is followed by our approach for enhancing the DPM method, and its evaluation, in Section 4. Section 5 consists of a comparison of the enhanced DPM method with C/R Models, and Section 6 presents a procedure selecting between C/R Models and the enhanced DPM, and its evaluation. We conclude the paper in Section 7 with a summary and directions for future research.

2. Basic Concepts of Defect Content Estimation Methods

In this section we present the two defect content estimation methods under study. Section 2.1 presents C/R Models and Section 2.2 presents the DPM. In Section 2.3 we present our research objectives.

2.1 Basic Concepts of Capture-Recapture Models

In order to describe the principles of C/R Models let us take a look at one of the most basic models stated in terms of inspections. Suppose a software artifact with a total of N defects is inspected by two inspectors. The first inspector detects n_1 of these defects while the second inspector detects n_2 of these defects. Usually, both inspectors do not detect exactly the same defects, thus let m_2 be the number of defects detected by both inspectors.

If we now assume, that each inspector has a probability p_i ($i=1,2$) of detecting defects, we have $E(n_i)=Np_i$ and $E(m_2)=Np_1p_2$, where E(x) denotes the expected value of x. Thus, we can denote N as

$$N = \frac{E(n_1) \cdot E(n_2)}{E(m_2)} \qquad \text{(Eq. 1)}$$

and derive an estimator for the number of defects as

$$\hat{N} = \frac{n_1 \cdot n_2}{m_2} \qquad \text{(Eq. 2)}$$

This estimator is known in biology and wildlife research as Lincoln-Peterson Estimator ([12],[16]).

One of the major differences between the various C/R Models are the assumptions about the detection probabilities. In Table 1 those models suitable for inspections, their assumptions, as well as their corresponding estimators are listed [6].

Model	Assumptions on detectability	Estimators
M0	Defects are equal with respect to their probability of being detected, the probability of detecting defects among inspectors is the same.	M0(MLE)
Mt	Defects are equal with respect to their probability of being detected, the probability of detecting defects among inspectors varies.	Mt(MLE) Mt(Ch)
Mh	Defects have different probabilities of being detected, the probability of detecting defects among inspectors is the same.	Mh(JE) Mh(Ch)
Mth	Defects have different probabilities of being detected, the probability of detecting defects among inspectors varies.	Mth(Ch)

Table 1: Summary of Capture-Recapture Models for Inspections

Model M0 assumes that the inspection can be described by one parameter, namely the probability p that any defect is detected by any inspector. Thus, it assumes that all defects have the same probability of being detected and all inspectors have the same probability to detect defects. Model Mt relaxes this strict assumption by assuming that each inspector i has associated a specific probability p_i of detecting any defect. Thus, this model still assumes that all defects have the same probability of being detected.The example presented above belongs to this model and the Lincoln-Peterson Estimator is the Maximum-Likelihood Estimator Mt(MLE) for two inspectors. Model Mh relaxes the strict assumption of Model M0 by assuming that each defect j can have a different probability p_j of being detected, which is the same for all inspectors. Finally, Model Mth combines both alleviations by assuming that each inspector i has a probability p_i to detect defects and that each defect j has a probability p_j of being detected. The probability that inspector i detects defect j becomes then $p_i \cdot p_j$.

In a previous study we have shown [6], that Model Mh and Model Mth are the most appropriate ones for inspections based on the dataset we evaluated. We recommended the Jackknife Estimator Mh(JE) and Chao Estimator Mh(Ch) for Model Mh when used in the context of software inspections.

2.2 Detection Profile Method

According to Wohlin and Runeson [18], real inspections usually violate the assumptions made by C/R Models, and thus it is necessary to find estimation methods that do not rely on these assumptions. Therefore, they proposed alternative methods for defects content estimation.

These methods are based on sorting and plotting defect data. Assuming that a curve can be fitted through the actual plot, the parameters of this curve can be determined and used to estimate the total number of defects. Based on this principle, Wohlin and Runeson present two graphical methods [18]. These two methods differ in the criterion used to plot the defect data and the selection of the curve to be fitted. In this paper we consider only one of these methods, namely the one that turned out to provide the more accurate results in Wohlin and Runeson's evaluation.

This method is called Detection Profile Method (DPM), since a profile is created based on how many inspectors detected each defect. For each defect, one calculates how many inspectors detect that defect. The defects are then sorted in descending order according to the number of inspectors detecting them, and are plotted in a graph as shown in Figure 1.

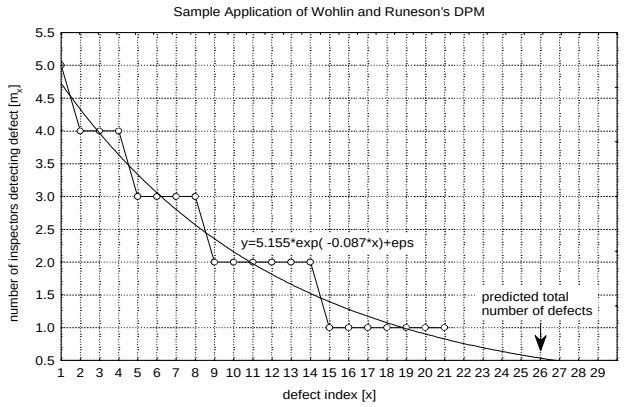

Figure 1: Example application of the Detection Profile Method.

Wohlin and Runeson assume that:

(a) adding more inspectors will lead to more detected defects; thus, after adding an appropriate number of inspectors all defects will be detected, and

(b) the data can be approximated by an exponential function, then

we can now obtain an estimate by fitting a decreasing exponential curve to the data:

$$m_x = A \times e^{-bx} \qquad \text{(Eq. 3)}$$

where m_x indicates the total number of inspectors to find a defect x, b is a factor describing the decrease of the exponential function, and A is a constant. Taking the log on both sides yields:

$$\ln(m_x) = \ln(A) - bx\,. \qquad \text{(Eq. 4)}$$

Linear regression can be used to estimate the parameters A and b. Once these parameters are determined, the total number of defects is determined by the largest x-value for which Equation 3 provides a result larger than or equal to 0.5.

In [18] Wohlin and Runeson conclude that the performance of the Detection Profile Method is not statistically significantly better than the Maximum Likelihood Estimator for the C/R Model Mt (in Table 1 denoted as Mt(MLE)). Nevertheless, the DPM showed better results in terms of mean relative error and standard deviation and should thus be considered for further research. In addition, DPM has some appealing properties: it is easy to operationalize using commonly available tools, and its ease of visualization makes it intuitively appealing to non-specialists who would have to apply the method in practice.

2.3 Research Objective

In the preceding sections two different methods for defect content estimation were presented. Each of these approaches makes different assumptions.

The fundamental assumption underlying Wohlin and Runeson's DPM is that the defect data fits an exponential curve. However, they also point out that this assumption might not be true in all circumstances. Therefore, we enhance the DPM approach (1) by taking into account different shapes a fitted curve might have and (2) by providing a way to choose between these different curves.

After the DPM has been improved, we have to explore the strengths and weaknesses of both the C/R Models and the improved DPM. This enables us to provide a procedure which selects the most appropriate defect content estimation method in a given situation.

This selection is of value since C/R Models are not appropriate in all conditions. The study in [6] found that many C/R Models exhibit rare but extreme outliers in their estimates when applied with real inspection data, even though their aggregate behavior may be desirable (e.g., high accuracy). An extreme outlier is a high over/under estimation and points out conditions under which the estimator should not have been applied.

From an applied point of view, extreme outliers pose a dilemma. If an estimator provides accurate estimates most of the time, but occasionally gives extreme values, can you trust it most of the time? Most likely the answer is no, hence making the, rather accurate, estimators unusable in practice.

A selection procedure between C/R Models and the DPM should try to overcome this drawback of C/R Models

and result in estimates that do not produce outliers.

The remaining parts of the paper follows the objectives stated above. The enhancement of the DPM is described in Section 4, the comparison between C/R Models and the DPM follows in Section 5, and Section 6 provides a selection procedure between the two approaches. But first we explain in the following section how we evaluate the different estimation methods.

3. Approach for Evaluating Estimators

To evaluate different defect content estimation methods we must estimate the defect content for several inspections based on actual inspection data and judge their accuracy. In this section we describe the criteria according to which we evaluate, the data source for the inspection data, and how we obtain defect content estimates.

3.1 Evaluation Criteria

In order to evaluate the different estimation methods, it has to be determined how accurate the estimates are. For this evaluation, we must be able to measure how close the estimated value is to the actual value. As a measure of this accuracy, we use the relative error (RE) defined as:

$$RE = \frac{\text{estimated \# of defects} - \text{actual \# of defects}}{\text{actual \# of defects}} \quad \text{(Eq. 5)}$$

The RE value allows us to distinguish between overestimation (too many defects were estimated, thus, a positive RE is obtained) and underestimation (too few defects were estimated, thus, a negative RE is obtained).

When dealing with estimators, two properties of these estimators should be investigated: bias and variability.

The bias of an estimator tells us how accurate the estimates are on average. It can be expressed as the central tendency across a number of estimates. This can be, for example, the mean or the median. A disadvantage of the mean is that it is sensitive to extreme values or outliers. A problem we are addressing in this paper is extreme over/ under estimation. Large outliers would distort the mean. Therefore, bias is defined here as the median RE.

Besides the average RE of the various estimators, it is also important to look at their RE variability. Variability tells us whether a large variation around the central tendency can be expected, e.g., whether extreme outliers can be produced by the model. The inter-quartile ranges and the possible presence of outlier values are used as measures of variability.

In addition to these two properties, it has to be investigated how often an estimate can be computed. Almost for all estimators conditions exist where they fail to provide an estimate. In order to assess, whether too large a number of failures occurs, the failure rate of the estimators has to be determined. We define the failure rate as the percentage of estimates that fail (i.e., cannot be computed).

3.2 Data Source

The data that we use for our evaluation comes from experiments to assess different reading techniques for inspections. These experiments were performed between 1994 and 1995 at the NASA/Goddard Space Flight Centre (NASA/GSFC) [2].

Since the goal of the experiments was to compare reading techniques, they focused exclusively on the defect detection step of an inspection process. All the data that is considered here follow an “Ad-hoc” preparation process, i.e., no specific reading technique was used. Neither inspection meetings nor corrections to the inspected documents were performed.

The documents being inspected during the experiments were two actual requirements documents from NASA/ GSFC describing functional specifications for satellite ground support software. They were structured according to the IEEE standard [11] and the different requirements were stated in natural language.

All defects in these documents were detected during subsequent development phases.

The two different documents were inspected in two experimental runs each. Since the documents were modified for the second run, we treat both runs independently (i.e., we treat them as four different documents).

The inspectors reading the documents were software professionals at NASA/GSFC with various levels of experience in the application domain and the development techniques used. This can therefore be considered representative of the circumstances in actual projects.

3.3 Evaluation Strategy

Since the experiment described above focused on the detection step of an inspection, no meetings were performed during the experiment. However, any number or combination of inspectors could be grouped and called a “virtual inspection meeting”.

The number of inspectors was systematically varied between two and six. This can be considered as representative for real-world inspections [3], [5], [15]. Additionally, it has been shown [6] that the number of inspectors has an influence on the accuracy of C/R Models and, thus, should be taken into account here.

For our evaluation, we compiled all possible “virtual inspection meetings” for a given number k ($2 \le k \le 6$) of inspectors and for all documents. For example, if for a document a total of six inspectors were available and

inspections with three inspectors were investigated, 20 "virtual inspection meetings" are then formed (there are 20 ways in which three inspectors can be selected from six). For each of these "virtual inspection meetings" an estimate was obtained with each estimator that is under evaluation. This is repeated for each of the four documents.

Bias and variability for a given estimator and number of inspectors was then determined by computing the median RE and the inter-quartile range, as well as the absolute maximum values for all combinations from that estimator and k inspectors.

4. Evaluation and Enhancements of DPM

In this section we propose a number of enhancements to the DPM. We then evaluate these alternative enhancements and select the one that is most promising.

4.1 Strategies for Detection Profile Method

The basic concepts behind the DPM have been outlined earlier. The basic assumption behind Wohlin and Runesons DPM proposal is that an exponential curve can be fitted through the plot. But they also indicate that other curves might be more appropriate as well.

Actually, there are situations where fitting a decreasing exponential curve to the inspection data may not be appropriate as this leads to highly inflated estimates. For instance, this will occur when there are no defects found by exactly one inspector. This could happen, for example, when a large number of inspectors participate. When fitting an exponential curve, the last number of defects that provides a value greater than 0.5 can be quite large. For example, consider the situation in Figure 2. Here we have an inspection where 13 defects were found and there are actually 15 defects in the document. No defects were found by exactly one inspector. Application of DPM would provide us with a highly inflated estimate of defect content.

In addition, visual inspection of real inspection data indicated to us that an exponential fit may not always be the most appropriate. For example, consider the data in Figure 3 which comes from one inspection in our data set. Here a linear fit gives an R^2 value of 0.92, whereas the exponential curve has an R^2 value of 0.84. In addition, the estimate from the linear fit is closer to the actual than that from the exponential fit. This indicates that, in some cases, a linear fit may perform better than an exponential fit. This represents an alternative strategy to fitting a curve to the data.

To apply this strategy, there must be objective criteria for selecting the type of curve to fit (i.e., exponential vs. linear). We propose two alternative criteria:

- **R^2 Criterion:** Select the fit that has the largest R^2 value (i.e., the best goodness of fit). Intuitively, this would seem to be a good enough criterion. However, in some cases when the number of defects found by one inspector is zero, the exponential curve may still provide a very good fit on the range where data are available, but also an inflated estimate due to an inaccurate extrapolation of the model. Therefore, we consider an alternative criterion.

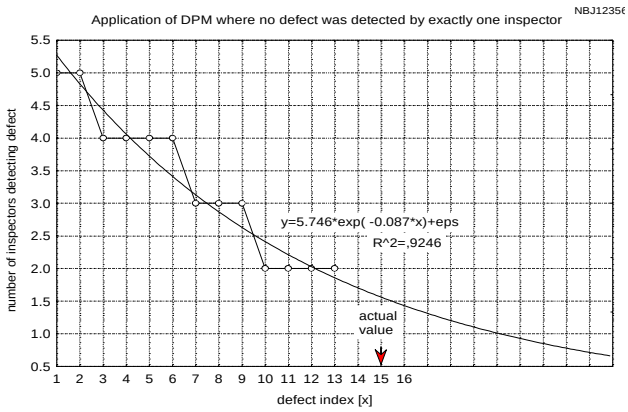

Figure 2: Example of overestimation when DPM is applied and no defects exist that were found by only one inspector.

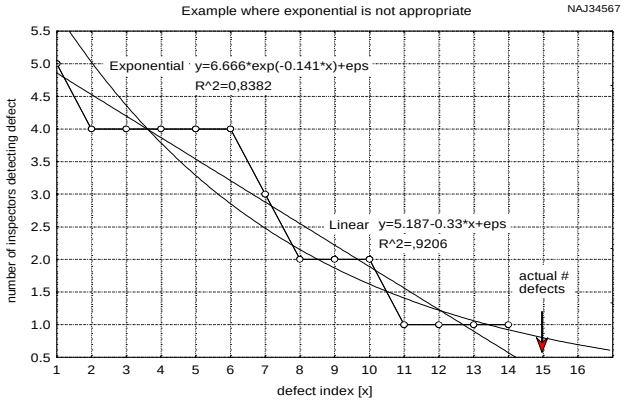

Figure 3: Example of a situation where application of an exponential fit is not appropriate. The data used here comes from our data set that is presented later in the paper.

- **Strict Order Criterion:** The second criterion that we propose is a rather strict one for selecting an exponential fit:

$$f_1 \geq f_2 \geq f_3 \geq \ldots \geq f_k \qquad \text{(Eq. 6)}$$

where k is the number of inspectors and f_i is the number of defects found by exactly i inspectors. This ordering would be expected if the underlying true relationship is exponential, and ensures that an exponential fit would serve better than a linear fit.

Using this criterion, the situation where f_1=0 will lead to the selection of linear fit.

4.2 Comparison of DPM Strategies

We now evaluate each of the four strategies outlined above: always fit exponential (original Wohlin and Runeson strategy), always fit linear, select between exponential and linear based on R^2, and select between exponential and linear based on the strict order criterion.

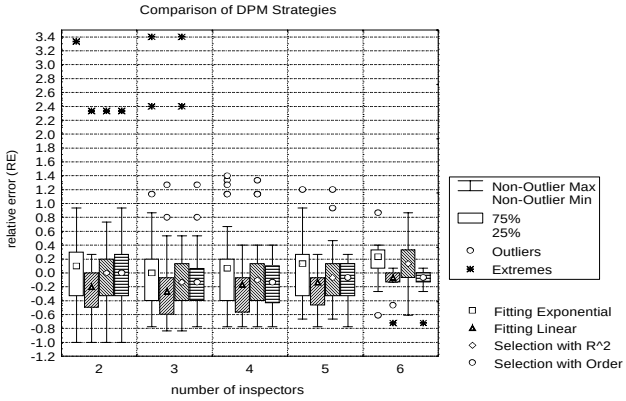

Figure 4: Comparing different Detection Profile Strategies

The results of comparing these four strategies are given in Figure 4 (In this and the following boxplots we use the following notation: the interquartile range is denoted as a box and a data point is deemed to be an outlier if the following conditions hold: datapoint>UBV+1.5*(UBV-LBV) or datapoint<LBV-1.5*(UBV-LBV), where UBV is the 75th percentile and LBV is the 25th percentile. For extreme values the factor 1.5 is changed to 3.0).

Additionally, numeric comparison of bias and variability are provided in Table 2 to Table 4, where the different approaches are ranked. In general, while the exponential fit strategy in some cases has less bias (e.g., see three inspectors) than the other strategies, it consistently has extreme outliers and also high variability. Since one of the problems witnessed with other approaches for estimating defect content has been extreme outliers, then this strategy is clearly less preferred.

	2 insp.	3 insp.	4 insp.	5 insp.	6 insp.
Rank1	Order R (0.0000)	Exp (0.0000)	Exp (0.0666)	Order R (0.0666)	Order Lin (0.0666)
Rank2		Order R (0.1333)	R (0.1000)		
Rank3	Exp (0.1000)		Order (0.1333)	Exp Lin (0.1333)	R2 (0.1333)
Rank4	Lin (0.2000)	Lin (0.2666)	Lin (0.1666)		Exp (0.2333)

Table 2: Bias in terms of absolute median relative error for different DPM Strategies

	2 insp.	3 insp.	4 insp.	5 insp.	6 insp.
Rank1	Order Lin R (2.3333)	Order Lin (1.2666)	Order Lin(0.4000)	Order Lin (0.2666)	Order Lin (0.0666)
Rank2					
Rank3		Exp R (3.4000)	Exp R (1.4000)	Exp R (1.2222)	Exp R (0.8666)
Rank4	Exp (3.3333)				

Table 3: Variability in terms of outliers (maximum values) for different DPM strategies

	2 insp.	3 insp.	4 insp.	5 insp.	6 insp.
Rank1	Lin (0.5000)	Order (0.4666)	Lin (0.5000)	Lin (0.4000)	Order Lin (0.1333)
Rank2	R (0.5333)	R Lin (0.5333)	Order) R (0.5333)	Order R (0.4666)	
Rank3	Order (0.6000)				Exp (0.2666)
Rank4	Exp (0.6333)	Exp (0.6000)	Exp (0.6000)	Exp (0.6000)	R (0.4000)

Table 4: Variability in terms of quartile-range for different DPM strategies

Out of the three other strategies, the strict order and linear fit strategies consistently have the least amount of variability. However, except for six inspectors, the strict order strategy has less bias. The R^2 selection strategy exhibits extreme outliers for 3, 4, and 5 inspectors, and has a large variation compared to the strict order strategy for 6 inspectors. Therefore, we conclude that the strict order strategy is the best one out of the four.

For the remainder of this paper we refer to the strict order selection strategy between the exponential and linear fits as the Enhanced DPM approach (EDPM).

5. Comparison of the EDPM with Capture-Recapture Models

The original motivation for the DPM approach was to introduce alternative defect content estimation methods making less restrictive assumptions than C/R Models.

Therefore, it would be prudent to compare the EDPM approach with the C/R Models to see if this relaxation of assumptions has any impact.

In a previous study [6] using the same data set that we are using in our current study we found that among all C/R Models the Model Mh was preferred. Therefore, it would be appropriate to compare Model Mh with the EDPM. Two estimators for Model Mh are considered, the Jackknife Estimator Mh(JE), and the Chao Estimator Mh(Ch) [6]. We also compare the EDPM approach with the Model Mt (Mt(MLE)) since this is the one that was used for comparison purposes in [18].

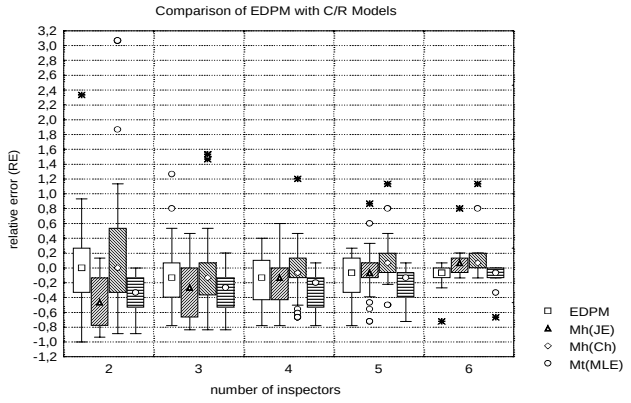

Figure 5: Comparison between best C/R Models and best DPM Strategy.

The results of this comparison are shown graphically in Figure 5. The different approaches are ranked in terms of their bias and variability in Table 5 to Table 7. In general, it can be seen that the EDPM approach is not a major improvement over C/R Models. In terms of bias, EDPM is comparable with Model Mh although it does not provide better results. In terms of variability, C/R Models outperform the EDPM approach for all numbers of inspectors except six, in which case it is tied with Model Mt. Therefore, despite its promise, the EDPM approach cannot be considered as a major improvement over the best existing C/R Models.

	2 insp.	3 insp.	4 insp.	5 insp.	6 insp.
Rank1	EDPM Mh(Ch) (0.0)	EDPM Mh(Ch) (0.1333)	Mh(Ch) (0.0666)	Mh(JE) Mh(Ch) EDPM (0.0666)	Mh(JE) Mh(Ch) Mt(MLE) EDPM (0.0666)
Rank2			EDPM Mh(JE) (0.1333)		
Rank3	Mt(MLE) (0.3333)	Mt(MLE) Mh(JE) (0.2666)			
Rank4	Mh(JE) (0.4666)		Mt(MLE) (0.2000)	Mt(MLE) (0.1333)	

Table 5: Bias in terms of absolute median relative error for C/R and EDPM.

	2 insp.	3 insp.	4 insp.	5 insp.	6 insp.
Rank1	Mt(MLE) (0.0)	Mt(MLE) (0.2)	Mt(MLE) (0.0666)	Mt(MLE) (0.0666)	Mt(MLE) (0.0)
Rank2	Mh(JE) (0.1333)	Mh(JE) (0.4666)	EDPM (0.4)	EDPM (0.26667)	EDPM (0.0667)
Rank3	EDPM (2.3333)	EDPM (1.2666)	Mh(JE) (0.6)	Mh(JE) (0.8666)	Mh(JE) (0.8)
Rank4	Mh(Ch) (3.0666)	Mh(Ch) (1.5333)	Mh(Ch) (1.2)	Mh(Ch) (1.1333)	Mh(Ch) (1.1333)

Table 6: Variability in terms of outliers (maximum values) for C/R and EDPM.

	2 insp.	3 insp.	4 insp.	5 insp.	6 insp.
Rank1	Mt(MLE) (0.4)	Mt(MLE) (0.4)	Mh(Ch) (0.2666)	Mh(JE) (0.1999)	EDPM Mt(MLE) (0.133)
Rank2	EDPM (0.6)	Mh(Ch) (0.4333)	Mt(MLE) (0.4)	Mh(Ch) (0.2666)	
Rank3	Mh(JE) (0.6444)	EDPM (0.4666)	Mh(JE) (0.4333)	Mt(MLE) (0.3334)	Mh(JE) (0.1999)
Rank4	Mh(Ch) (0.8666)	Mh(JE) (0.6666)	EDPM (0.5333)	EDPM (0.4666)	Mh(Ch) (0.2)

Table 7: Variability in terms of quartile ranges for C/R and EDPM.

6. Selecting between EDPM and C/R Models

A potential strategy for selecting between the EDPM estimate and the C/R estimate is to select the estimate that is better or most trust-worthy. In this section we present the rationale for a selection procedure, and then present the procedure itself, followed by its evaluation. During this evaluation we would like to answer the question: "does the selection procedure provide better results than the EDPM approach and the best C/R Model used in isolation?".

6.1 Strategy for a Selection Procedure

To motivate our selection strategy we first look at the relationship between the R^2 value from using EDPM and the relative error. For various numbers of inspectors, this is presented in Figure 6. We also make a distinction between the R^2 values that are statistically significant at an alpha level of 0.01, and those that are not. When there are a few inspectors (i.e., small inspection teams) or few defects discovered, the fit in the EDPM approach is made with very few observations. This can provide high R^2 values, but is

misleading because the fit would be quite unstable. Hence the importance of taking into account the statistical significance of the EDPM fit in addition to the R^2 value. Based on the number of virtual inspections analyzed, a p-value of 0.01 is selected to ensure that only a small number of R^2 values is wrongly accepted as statistically significant.

Visual inspection of the scatterplots indicates that large statistically significant R^2 values tend to cluster around zero RE. This behavior suggests that the EDPM method would likely provide reasonable results whenever the R^2 value is above a certain threshold and is statistically significant. We choose an R^2 threshold of 0.8. However, we found that the results of our evaluations are quite insensitive to variations of threshold value (up to 0.9 and down to 0.7).

Figure 7 provides an illustration where we compare the EDPM results with those of the three C/R Models for the case where the R^2 value is above the threshold and significant (upper panel), and when it is not (lower panel). As can be seen in the upper panel, the EDPM approach has generally consistently low bias for different numbers of assessors and does not exhibit the extreme outliers that the C/R models exhibit. If we look at the lower panel, we find that the EDPM bias is affected quite a bit by the number of inspectors, and for a low number of inspectors has some outliers.

The selection strategy then involves selecting either an estimate of a C/R Model or the EDPM estimate. If the R^2 value for the EDPM approach is greater than 0.8 and statistically significant, then select the EDPM estimate, otherwise select the C/R estimate. Since in a previous study it was found that, in general, the model Mh with the Jackknife Estimator is the most appropriate C/R Model [6], we use that estimator in our selection procedure.

6.2 Evaluation of the Selection Procedure

The results of the comparison of this selection procedure with Model Mh and the Jackknife Estimator and the EDPM approach are provided in Figure 8. We can see from this figure that the selection approach consistently provides good bias for more than three inspectors, and it does not suffer from extreme outliers as the other two approaches do. For less than four inspectors, the selection approach has a larger bias, but this is comparable to the C/R Model for 2 inspectors and comparable to both the C/R Model and EDPM approach for three inspectors. However, it does not yield extreme estimates as the EDPM approach does.

Based on this result we can conclude that:

- For two or three inspectors, there is not much difference between using the Jackknife Estimator for Model Mh, Mh(JE), and our selection procedure.
- For more than three inspectors, the selection procedure will provide consistently more reliable estimates (i.e., no extreme over/under estimation) and has similar bias to the C/R Model and the EDPM

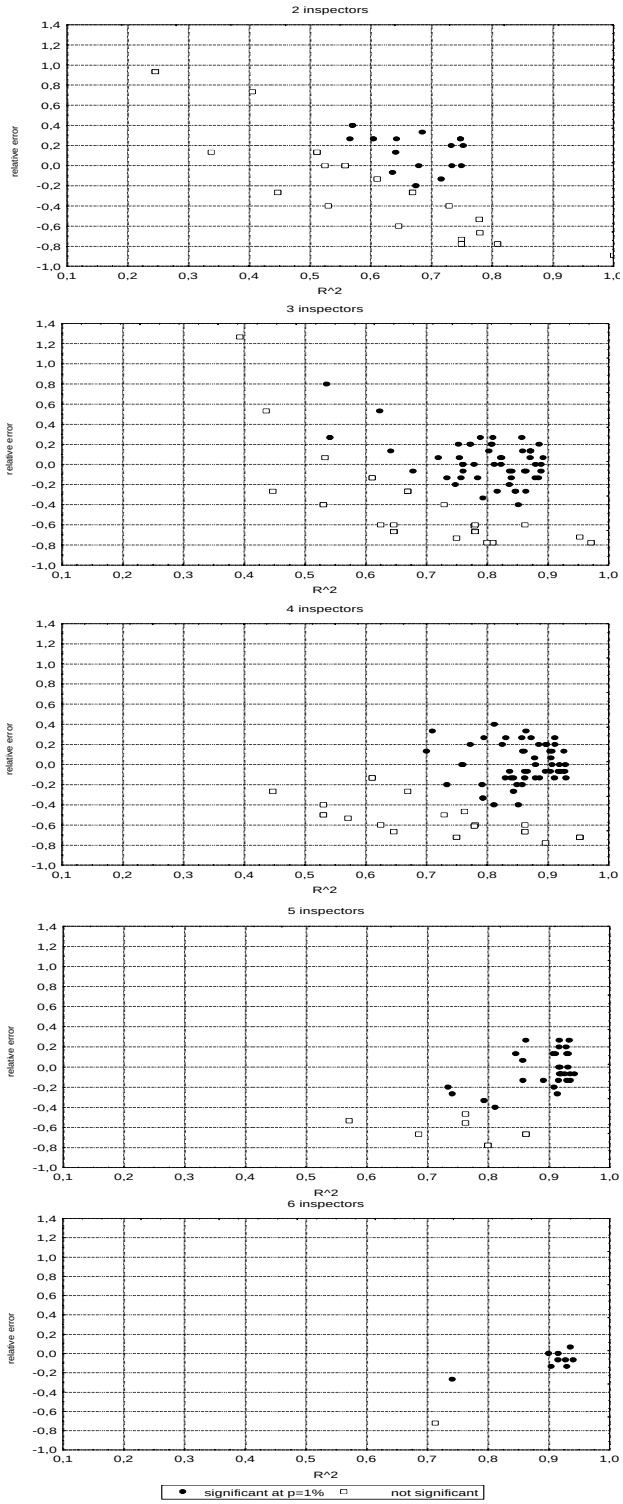

Figure 6: Relationship between R^2 and relative error

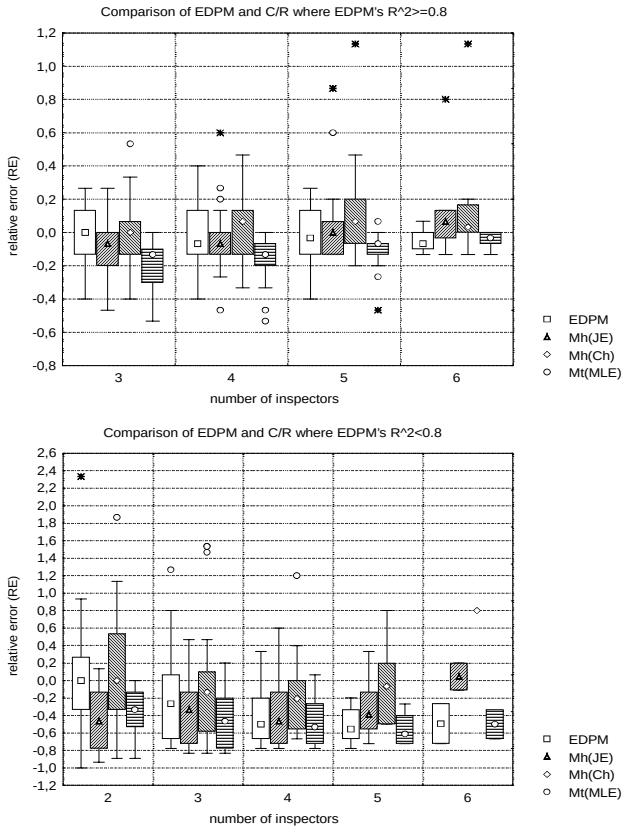

Figure 7: Selecting R^2 as threshold.

approach.

These two findings make it reasonable to recommend using the selection procedure since users can have a high degree of confidence in the estimates i.e., they are not likely to yield extreme errors which would lead to either a substantial waste of inspection effort or a larger amount of defect slippage to the next development phase. This conclusion is strengthened by the fact that the bias obtained by the selection procedure is comparable to other approaches, and therefore shows gains without any loss.

An additional advantage of the selection procedure is that the selection of EDPM indicates a reliable estimate. The upper part of Figure 7 shows that in this case a low bias and small inter-quartile range can be expected.

It is also informative to look at the failure rates of all the methods that we have considered. This is summarized in Table 8. The methods that are based on the DPM, including EDPM, will fail if all defects were found by the same number of inspectors. In such a case the fitted curve may not approach the x-axis (for example, a linear fit would be a horizontal line). It is also worthy of notice that the selection procedure gives consistently better or equally good results as all the other alternatives, whereby it fails very infrequently. This characteristic also makes it more attractive for use in practice since it is more likely to give an estimate, even for a small number of inspectors.

	2 insp. (66 comb.)	3 insp. (95 comb.)	4 insp. (80 comb.)	5 insp. (39 comb.)	6 insp. (10 comb.)
Fitting Exponential	27%	8%	0%	0%	0%
Fitting Linear	27%	8%	0%	0%	0%
Selection with R^2	27%	8%	0%	0%	0%
EDPM	27%	8%	0%	0%	0%
Selection Proc.	0%	2%	0%	0%	0%
Mh(JE)	0%	8%	0%	0%	0%
Mh(Ch)	29%	12%	9%	13%	10%
Mt(MLE)	29%	2%	0%	0%	0%

Table 8: Failure Rate (percentage of estimates that fail) for all methods

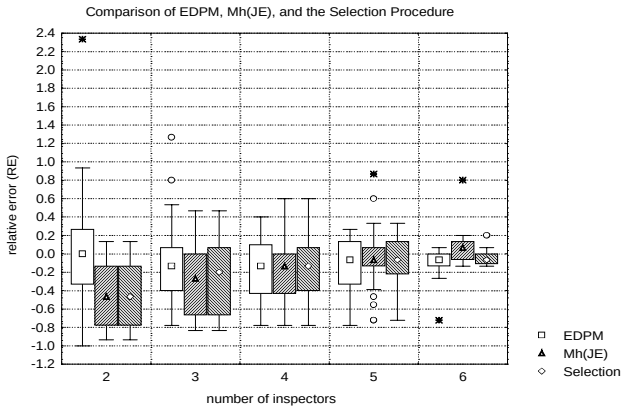

Figure 8: Comparing the selection procedure with EDPM and C/R Models

7. Summary and Conclusions

The objective of the study reported in this paper was threefold: (1) assess the Detection Profile Method (DPM) proposed by Wohlin and Runeson, (2) try to refine the DPM based on the results obtained in (1), (3) design a selection procedure to choose between C/R Models and the DPM method. Our goal was to improve the applicability of defect content estimation of inspected documents in realistic conditions. A previous study had identified that extreme outliers in C/R estimates are a potential problem that may limit the application of C/R Models: if a model works well most of the time, but occasionally provides extreme over/ under estimates, it would be difficult to trust its estimates. The new selection procedure proposed in this paper alleviates some of the weaknesses of the DPM approach, and takes advantage of the strengths of C/R Models. It was evaluated using actual inspection data.

Our evaluation results indicate that the DPM approach does not improve over the best C/R Model. In addition, although the improvements to DPM we proposed here provide better estimates than the original DPM approach, they still do not significantly improve over C/R Models. However, most importantly, the selection procedure shows as little bias in the estimates as the best C/R Models alone, but that it does not exhibit the extreme outliers that are often a characteristic of C/R estimates in the context of inspections. This result is encouraging in that it provides a solution that is more likely to be trusted, and hence adopted, in practice.

Further research in this area ought to consider improving the accuracy of the selection procedure of defect content estimation models. In particular, none of the approaches we have looked at provides completely satisfactory accuracies for inspections with a low number of inspectors.

Additionally, the results obtained in this study should be validated using alternative data sets from additional environments.

Acknowledgment

We would like to thank Claes Wohlin and Per Runeson, University of Lund, our colleagues from the QPE/QM group, and the reviewers of this paper for their valuable and helpful comments.

References

[1] Maria Paola Ardissone, Massimiliano Spolverini, and Mario Valentini. Statistical Decision Support Method for In-process Inspections. In *Proceedings of the 4th International Conference on Achieving Quality in Software*, pages 135–143, 1998.

[2] V. R. Basili, S. Green, O. Laitenberger, F. Lanubile, F. Shull, S. Sorumgard, and M. V. Zelkowitz. The Empirical Investigation of Perspective-Based Reading. *Empirical Software Engineering - An International Journal*, 1(2):133–164, 1996.

[3] D. Bisant and J.Lyle. A Two-Person Inspection Method to Improve Programming Productivity. *IEEE Transactions on Software Engineering*, 15(10):1294–1304, October 1989.

[4] Frank W. Blakely and Mark E. Boles. A Case Study of Code Inspections. *Hewlett-Packard Journal*, 42:58–63, October 1991.

[5] K. V. Bourgeois. Process Insights from a Large-Scale Software Inspection Data Analysis. *Crosstalk*, October 1996.

[6] Lionel Briand, Khaled El Emam, Bernd Freimut, and Oliver Laitenberger. Quantitative Evaluation of Capture Recapture Models to Control Software Inspections. In *Proceedings of the 8th International Symposium on Software Reliability Engineering*, pages 234–244, 1997.

[7] Stephen G. Eick, Clive R. Loader, M. David Long, Lawrence G. Votta, and Scott Vander Wiel. Estimating Software Fault Content before Coding. In *Proceedings of the 14th International Conference on Software Engineering*, pages 59–65, 1992.

[8] Stephen G. Eick, Clive R. Loader, Scott A. Vander Wiel, and Larry G. Votta. How many errors remain in a software design after inspection? In *Proceedings of the 25th Symposion on the Interface*. Interface Foundation of North America, 1993.

[9] Bernd G. Freimut. Capture-Recapture Models to Estimate Software Fault Content. Master's thesis, University of Kaiserslautern, Germany, June 1997.

[10] Tom Gilb and Dorothy Graham. *Software Inspection*. Addison-Wesley Publishing Company, 1993.

[11] Institute of Electrical and Electronics Engineers. *IEEE Standards Collection, IEEE Std. 830-1993*, 1994.

[12] David L. Otis, Kenneth P. Burnham, Gary C. White, and David R. Anderson. Statistical Inference from Capture Data on Closed Animal Populations. *Wildlife Monographs*, 62:1–135, October 1978.

[13] Per Runeson and Claes Wohlin. An Experimental Evaluation of an Experience-Based Capture-Recapture Method in Software Code Inspections. Accepted for publication in *Empirical Software Engineering - An International Journal*.

[14] Glen W. Russell. Experience with Inspection in Ultralarge-Scale Developments. *IEEE Software*, 8:25–31, January 1991.

[15] Edward F. Weller. Lessons from Three Years of Inspection Data. *IEEE Software*, 10(5):38–45, September 1993.

[16] G.C. White, D.R. Anderson, K.P. Burnham, and D.L. Otis. Capture-Recapture and Removal Methods for Sampling Closed Populations. Technical report, Los Alamos National Laboratory, 1982.

[17] Scott A. Vander Wiel and Lawrence G. Votta. Assessing Software Designs Using Capture-Recapture Methods. *IEEE Transactions on Software Engineering*, 19(11):1045–1054, November 1993.

[18] Claes Wohlin and Per Runeson. Defect Content Estimations from Review Data. In *Proceedings of the 20th International Conference on Software Engineering*, 1998.

[19] Claes Wohlin, Per Runeson, and Johan Brantestam. An Experimental Evaluation of Capture-Recapture in Software Inspections. *Software Testing, Verification and Reliability*, 5:213–232, 1995.