How accurately do engineers predict software maintenance tasks ?

Les Hatton
CISM, University of Kingston[*]

December 19, 2005

Abstract

This paper is a contribution to the empirical literature on distributions of software maintenance amongst the primary activities of adaptive, corrective and perfective maintenance. Unusually, the software process under study here maintains detailed information on predicted versus actual effort and is therefore able to shed some light on how accurately engineers predict both the type of maintenance necessary and the duration. The quality of these predictions is subjected to formal statistical analysis, with some surprises, for example, prior estimates of maintenance category are either completely correct or completely incorrect with little in-between.

Keywords: Maintenance, Prediction, Case history

1 Overview

Historically, all activity carried out on software systems after they have been delivered for the first time has been called *maintenance.* The IEEE standard definition for this is *The modification of a software product after delivery to correct faults, improve performance or other attributes, or adapt the product to a modified environment*, [4]. Traditionally, [9], maintenance is itself split up into three activities:-

- *Adaptive maintenance.* This is nowadays usually taken as the effort devoted to adding new functionality, either because there was no time to complete it before first delivery or, because it has been suggested by use of the product since first delivery. (It should be noted however that functional enhancement was not originally included in this category, [11].)
- *Corrective maintenance.* This is simply that effort devoted to the removal of defects and is the category on which most sources agree the best.
- *Perfective maintenance.* This is effort for which no change in functionality is expected. For example, it might be devoted to cleaning up or performance improvements. Benefits deriving from perfective maintenance are

[*] L.Hatton@kingston.ac.uk, lesh@leshatton.org

> less tangible and this category of maintenance is consequently often neglected in commercial organisations although it appears to be a significant factor in open source developments such as the Linux kernel where there is continual mention of cleaning up in regular reports such as [10]. Like adaptive maintenance, this category has not been made particularly clear-cut in previous work with items like re-engineering sometimes factored out on their own, [1].

The distribution of maintenance effort between these three categories appears to be very variable but this may simply because maintenance categories are not really very well-defined with corrective maintenance probably the least affected. Some of the quoted distributions are shown as Table 1 in this case for effort by total time spent, (often number of requests for each category may be quoted). In some cases, additional categories such as documentation are reported. In this paper, the categories as defined above will be used exclusively and any documentation effort was folded into the respective category.

Source	Adaptive	Corrective	Perfective	Comments
[3]	29	18	28	Out of 76 as study contained 24% 'other'
[1]	46	17	25	Out of 89 as study contained 11% 'answering questions'
[2]	42	37	22	Generally small groups
[7]	52	8	33	Out of 96 as study contained 4% 'other'
[5]	83	12	5	Reported asymptotic defect density of 0.8 per KSLOC in Cobol
This study	54	6	40	Reported asymptotic defect density of 0.2 per KSLOC in C

Table 1: Quoted distributions of maintenance activity amongst the three primary areas

An additional reason for this considerable variation and one studied in more detail here is leakage between the categories due to an inaccurate initial appraisal of the type of maintenance necessary. It is not unusual whilst performing say an adaptive maintenance activity to find a defect, or perhaps to decide that some perfective re-writing is necessary to clarify the structure in order to be able to add a new feature. Normally, data is not reported on this so it clouds the distinctions between the three primary categories. Another confounding factor in the author's personal experience, even when these categories are reasonably well-defined within an organisation, is that adaptive and perfective work is often used to hide corrective work.

The case history studied here is a little unusual. All the usual information is kept but in addition, the following has been recorded for each maintenance request.

- A triplet of non-negative integer numbers (a_e, c_e, p_e) where $a_e+c_e+p_e = 10$ representing the *estimated* contribution of adaptive, corrective and perfective components respectively. For example, (4,5,1) would represent a maintenance contribution estimated to contain 40% adaptive work, 50%, corrective and 10% perfective.
- A triplet of non-negative integer numbers (a_a, c_a, p_a) where $a_a + c_a + p_a = 10$ representing the *actual* contribution of adaptive, corrective and perfective components respectively after the work has been completed. These are normally filled out by the same person as estimated the triplet, so they are probably consistent but their accuracy is not separately assessed.
- Estimated time for a particular maintenance request, (known here as *Change Requests or CRs*), and actual time recorded when completed.

The presence of these triplets for each CR allows analysis of the accuracy of estimation both for distribution of maintenance activity and also the duration over a period of over four years.

2 Case history details

The data is taken from a small software development company. 957 CRs totalling some 5000 hours of work were completed between 14th March, 2001 and the end of the study period, 14th November, 2005 on a total of 13 packages of which about 60% were spent on two commercial products, about 10% on open source support and the remainder on web-site development including eCommerce activities, and utilities including graphics and licensing support. In other words, the spread of work is not untypical.

Figure 1 shows the distribution of activity by maintenance type in both total time and also in number of CRs. Compared with Table 1, this distribution represents a rather higher emphasis on perfective maintenance than is normal and a rather lower requirement for corrective maintenance. Both the main packages maintained by this company (each in the range 110,000 - 150,000 source lines of code, (SLOC)), have a comparatively low level of defect as shown by Table 2, but the relationship between this fact and the slightly unusual maintenance profile will not be explored here.

Component	Cum. defect density (per 1000 SLOC)
Package 1 GUI client	0.599
Package 1 Server	0.224
Package 2 GUI client	0.269
Package 2 Server	0.184

Table 2: Defect densities for the two main contributors to the maintenance records of Figure 2. GUI clients are written in Tcl/Tk and the servers in C.

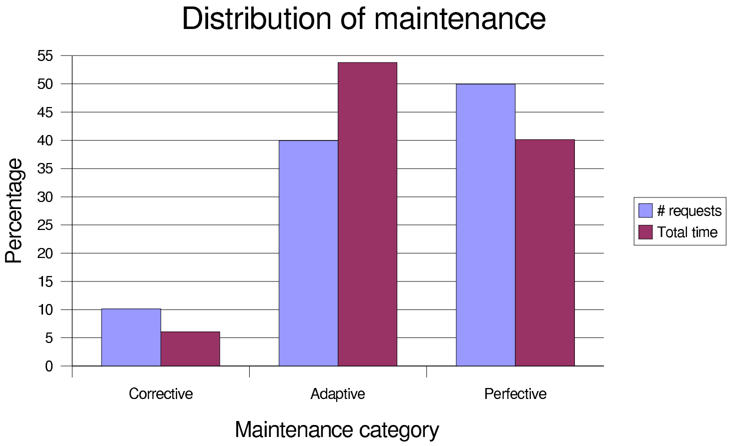

Figure 1: Distribution of activity by maintenance type

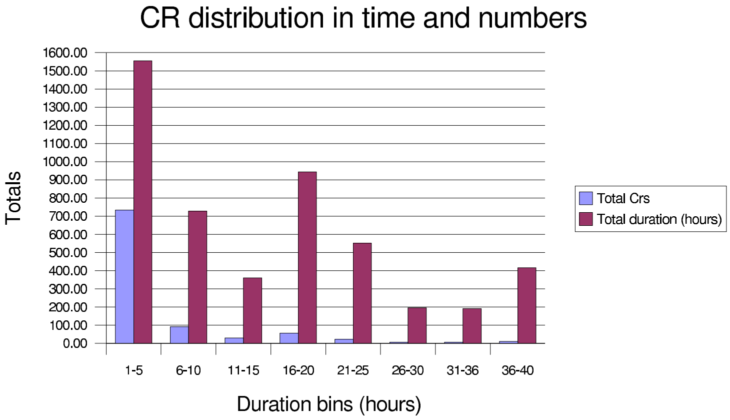

Figure 2: Distribution of Change Requests by both number and total duration.

In contrast, Figure 2 shows the CRs binned according to actual duration in both number and total duration for each 5 hour bin. Approximately 75% by number fall into the shortest 1-5 hour bin as can be seen, so maintenance activities are dominated by small changes measured either by number or total time.

3 Estimation accuracy of maintenance duration

It is interesting to ask the question whether engineers become more accurate in their estimates of maintenance tasks as time goes by. In this case history, the personnel did not change during the period covered by the maintenance records and so it might be expected that they would become progressively more accurate in assessing the time taken to perform a CR. This will be characterised and tested in two ways:-

- Any change in the bias.
- Any change in the spread.

The maintenance records contain the required information on both the estimated and actual duration of each CR. Figure 3 shows the difference between estimated and actual CR durations for the full set of records in increasing chronological order. Points above the zero line correspond to CRs which were performed more

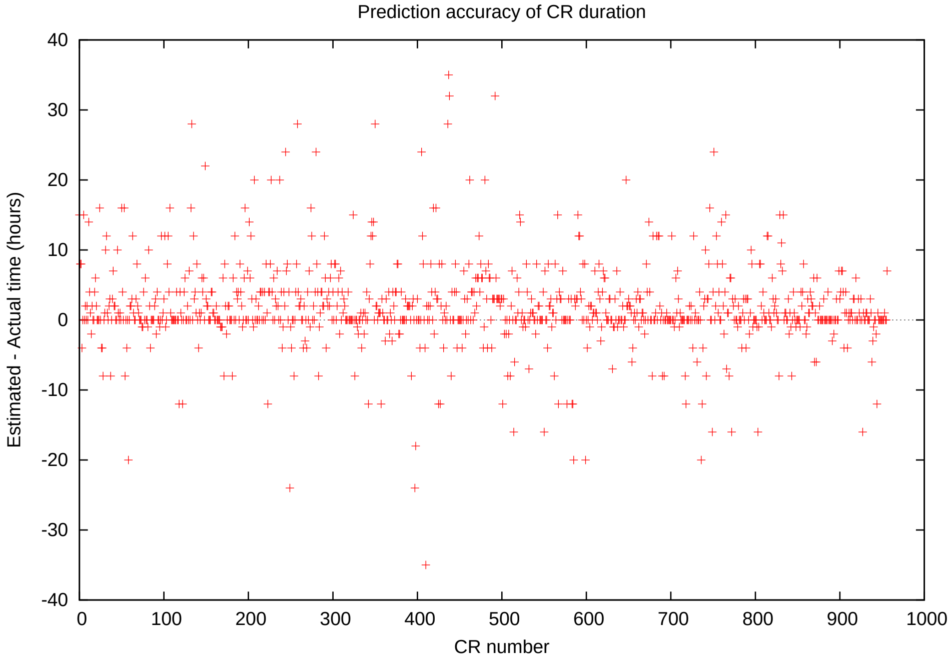

Figure 3: Estimated minus actual CR duration in chronological order

quickly than expected. Note that the frequency at which CRs were resolved did not change significantly throughout the measurement period.

3.1 Changes in bias

It is clear from Figure 3 that there is a systematic optimistic bias. Using the data of Figure 2, the average duration of a CR was 5.17 hours and the average bias is 1.83 hours so, on average, engineers over-estimated across all CRs by around 35% for any kind of maintenance change. The significance of these will now be established.

Splitting the data into two halves reveals that the average bias in the first half is 2.45 hours and the average bias in the second half of the data is 1.2 hours. The following hypotheses will therefore be made:-

- H_0 The null hypothesis, the average bias in each half came from the same population.
- H_1 The alternative hypothesis, the average bias in each half year did not come from the same population.

The data will be analysed using the z-test for the difference of means in a population, [8]. This states that the following statistic is approximately distributed as N(0,1).

$$z = \frac{\overline{X_1} - \overline{X_2}}{(\frac{(s_1)^2}{N_1} + \frac{(s_2)^2}{N_2})^{\frac{1}{2}}} \qquad (1)$$

where $\overline{X_i}$, s_i and N_i are respectively the sample means, standard deviations and number of samples for each of the half samples. Substituting the appropriate values yields,

$$z = \frac{2.45 - 1.20}{(\frac{44.95}{477} + \frac{29.78}{478})^{\frac{1}{2}}} \simeq 3.16 \qquad (2)$$

There is a probability of less than 0.001 that this could have occurred by chance so the null hypothesis H_0 is rejected and it must be concluded that there is a highly significant drop in the average bias in the two half samples. From now on tests for significance, will be abbreviated somewhat.

3.2 Spread of estimates

Here, the variances in the two half samples will be compared using the F-test, [8]. In this case computing the F-statistic gives a value of:-

$$F = \frac{{S_1}^2}{{S_2}^2} \simeq 1.51 \qquad (3)$$

The number of degrees of freedom for the numerator and denominator respectively is 476 and 477 respectively. At the 5% significance level, the corresponding value of the F statistic is less than 1.25. It can therefore be deduced that the drop in the variance is significant at the 5% level.

3.3 Relationship with CR duration

The distribution of CR durations over the maintenance period appeared random so it is useful to see if there is any relationship between estimation accuracy and the actual duration of the CR, for example, it might reasonably be expected that shorter duration CRs were predicted more accurately. The relationship is shown in Figure 4 and a clear anti-correlation can be seen. Shorter duration CRs are generally optimistically predicted whereas longer duration CRs are generally pessimistically predicted.

The product moment correlation coefficient for this sample is -0.30 with 957 samples. A standard significance test on this (one-tailed or two-tailed) yields a probability of this happening by chance as < 0.0001 in both cases so a null hypothesis that there was no such trend is comprehensively rejected.

4 Estimation accuracy of maintenance type

In this section, the two triplets (a_e, c_e, p_e) and (a_a, c_a, p_a) for each maintenance record will be compared by calculating their normalised cross-correlation giving a range of comparison between 0.0 for no similarity and 1.0 for exact similarity.

An example of the kinds of insight this data presents is shown in Figure 5. Here the amount of time spent as a function of match between estimated and actual maintenance categories is shown. The graph shows a relatively unusual distribution in that around 77% of all time was spent on CRs whose spread of activity between adaptive, corrective and perfective was estimated correctly.

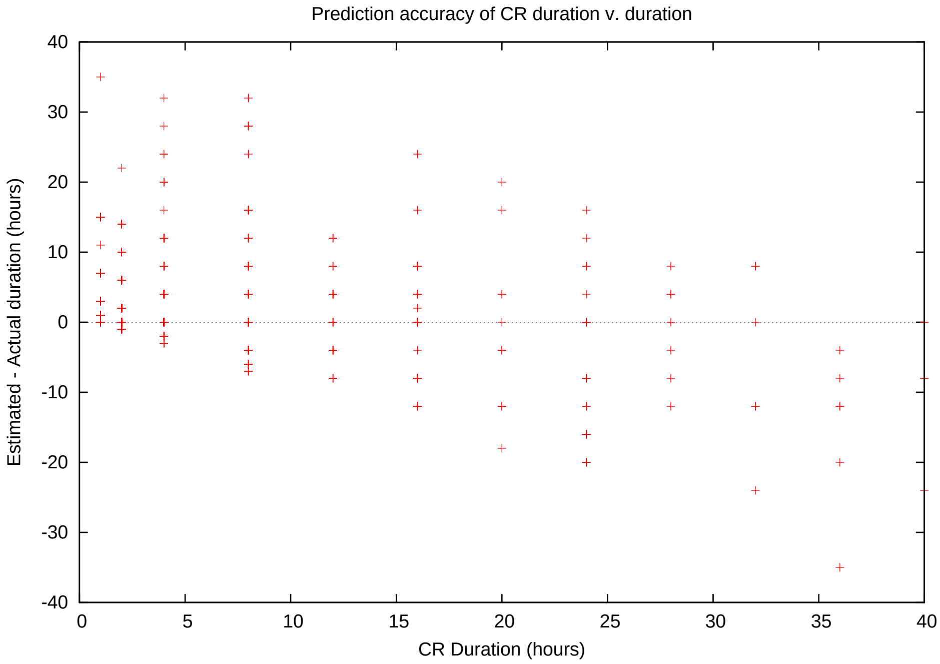

Figure 4: Estimated - actual CR duration v. actual CR duration

At the other end of the scale about 16% of the total time was spent on CRs where the estimate was estimated completely incorrectly, representing a major surprise when work was actually carried out. The remaining 7% was spread more or less evenly in between these two extremes. Overall, 1 in 4 of the initial estimates for maintenance distribution turned out to be incorrect when the work was eventually carried out and most of these were completely incorrect, (i.e. no correlation between the estimated and actual maintenance triplets).

An obvious relationship to investigate is whether the degree of correlation between actual and estimated categories of maintenance was related perhaps in an inverse way to the duration of the CR when it was actually carried out. The raw data is shown in Figure 6 which plots the value of the normalised cross-correlation between the estimated and actual maintenance triplets against the actual duration of the CR in hours. There is a clear and unexpected pattern that shorter CRs are much less predictable in terms of maintenance type than the longer CRs. Formal statistical analysis gives a value for the product moment cross-correlation of 0.53 with 957 samples which is very highly significant. It is not clear what perceptual mechanism would cause this and more work will be necessary.

5 Unexpected transitions

Analysis of the estimated and actual maintenance triplets allows unexpected transitions to be quantified, for example, how often adaptive maintenance led to

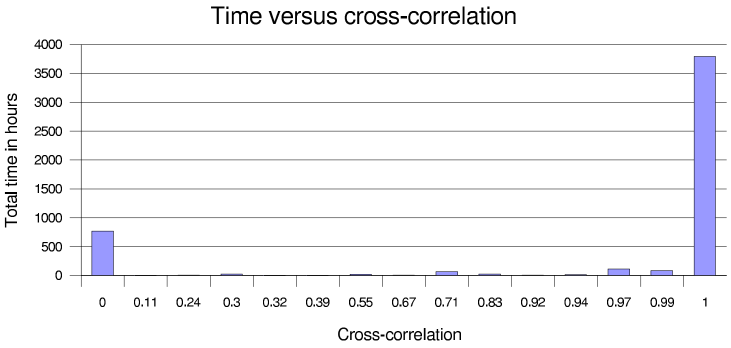

Figure 5: Time spent on maintenance as a function of maintenance category estimation accuracy.

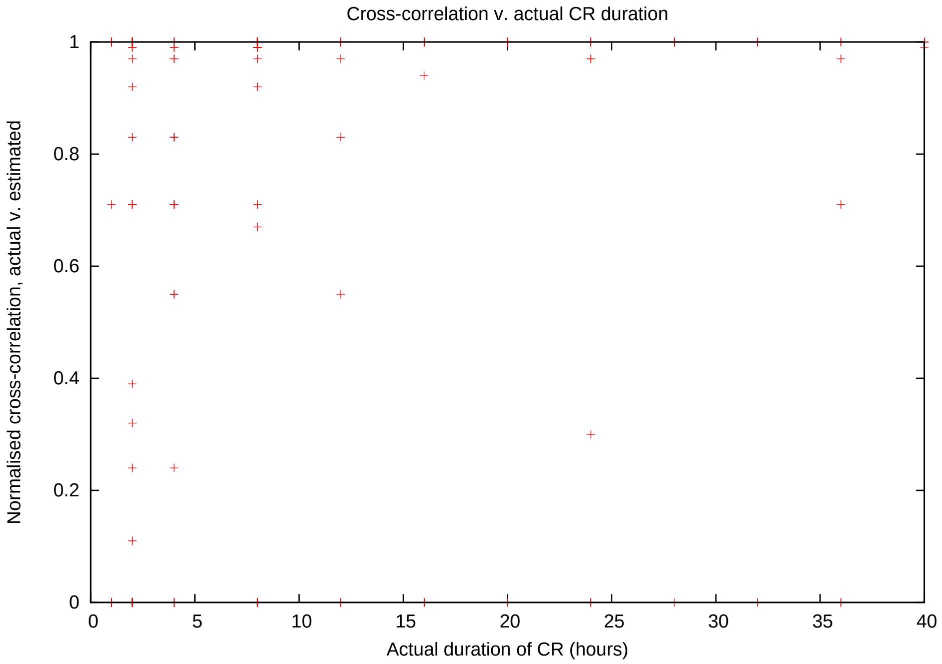

Figure 6: Estimated - actual CR duration in chronological order

some form of corrective maintenance or how often corrective maintenance led to some form of perfective maintenance. To clarify this, the following nomenclature will be introduced:-

Define the following:-

- A *pure* adaptive triplet will be defined to be (10,0,0).
- A *pure* corrective triplet will be defined to be (0,10,0).
- A *pure* perfective triplet will be defined to be (0,0,10).
- *Impure* triplets will be defined as having a value v, such that $0 < v < 10$ in the appropriate position. Triplets were constrained by the change management mechanism so that the sum of their components was always 10.

With these definitions, Table 3 shows transitions from pure triplets to impure triplets in decreasing order of frequency. These comprise around 19% of all the CRs.

Transition	Number of occurrences
Pure perfective to impure adaptive	66
Pure adaptive to impure perfective	50
Pure perfective to impure corrective	27
Pure corrective to impure perfective	22
Pure adaptive to impure corrective	9
Pure corrective to impure adaptive	6
Pure perfective to impure adaptive/corrective	3
Pure corrective to impure adaptive/perfective	1
Pure adaptive to impure corrective/perfective	0

Table 3: Transitions between pure and impure maintenance triplets in decreasing frequency.

As can be seen, the most frequent transitions occurred when perfective maintenance was scheduled and opportunities for adaptive work naturally arose and vice versa. It is also interesting to note that transitions between adaptive and corrective maintenance occurred rather less frequently than corresponding transitions between perfective and corrective maintenance. This may suggest that developers are more alert to defects during perfective maintenance than during adaptive maintenance. This observation will be analysed using the z-test for proportions, [8]. Let p_{ac} be the proportion of all changes with transitions between adaptive and corrective and let p_{pc} be the proportion of all changes between perfective and corrective. Then,

$$p_{ac} = \frac{15}{957} = 0.0157; q_{ac} = 1 - 0.0157 = 0.984 \tag{4}$$

and

$$p_{pc} = \frac{49}{957} = 0.0512; q_{pc} = 1 - 0.0512 = 0.949 \tag{5}$$

It will be assumed as a null hypothesis that the two binomial populations for these proportions are the same.

With this assumption, the following holds, [8]:-

$$z = \frac{p_{ac} - p_{pc} - 0}{\sqrt{\widehat{p}\widehat{q}\{\frac{1}{n_1} + \frac{1}{n_2}\}}} \sim N(0,1) \tag{6}$$

where $n_1 = 15$ and $n_2 = 49$ are the number of transitions found in adaptive to corrective and in perfective to corrective transitions respectively. An estimate for $\widehat{p} = 0.5 * (0.0157 + 0.0512) = .0335$ and for $\widehat{q} = 0.5 * (0.984 + 0.949) = 0.9665$. This gives

$$z = \frac{0.0355}{0.0547} = 0.649 \tag{7}$$

which is *not* significant. So although there appears to be a pattern, it is not a significant one and the null hypothesis cannot be rejected implying that transitions between perfective and corrective maintenance were not significantly different in frequency to transitions between adaptive and corrective frequency.

Repeating this test comparing the most numerous transitions (adaptive ⇔ perfective) and the least numerous (adaptive ⇔ corrective) confirmed significance but only at the 10% level so there were no compelling patterns favouring one kind of transition over another.

6 Conclusions

The literature still shows very considerable overlap between activities traditionally called maintenance. This study adds further empirical evidence on the distribution of such activities for two relatively recent products and additionally throws light on how well programmers estimate the maintenance category initially.

There is good statistical support for the following observations:-

- Predictions of the amount of time necessary for a change of any kind across a range of products became considerably more accurate as time went by. The reduction in over-optimism is significant at the 0.1% level and the reduction in the variance is significant at the 5% level. This is not a surprise but given the complexity of applications, it is reassuring to be able to quantify it although the variance of the estimates remains high.

- Short CRs were consistently over-estimated in the time necessary to complete them but long CRs were consistently underestimated. This relationship was highly significant. In other words, programmers appear to perform a form of averaging when predicting the duration of a maintenance task.

- Unusually, shorter CRs were much more likely to be predicted incorrectly in terms of maintenance category than longer CRs. This result was not expected and was confirmed at high significance. No potential explanation is offered here.

- The data showed that most maintenance requests were predicted correctly, ($\sim$ 77%) but of the remainder, the most likely eventuality was to predict the maintenance category completely *incorrectly* showing that even in a stable environment (no staff turnover and stable products), the frequency of total surprise remains significant.

- Transitions between adaptive and corrective maintenance, (i.e. corrective maintenance unexpectedly forming part of an adaptive task and vice versa) and transitions between perfective and corrective maintenance were not significantly different in frequency and no compelling patterns were found favouring unexpected transitions between one type of maintenance and another.

The overall picture presented here is of a gradually improving but rather coarse-grained ability to predict duration of a maintenance task but with significant surprises continuing even in a personnel- and product-stable environment.

References

[1] Dekleva, S. (1992) *Software maintenance: 1990 status*, J. Software Maintenance: Research and Practice 4(4), p. 233-247

[2] Glass R.L. (1996) *Results of the first IS State-of-the-Practice survey*, The Software Practitioner, 6(3-4), May-August

[3] Helms G.L. and Weiss I.L. (1985) *Applications software maintenance: can it be controlled*, DataBase, 16(2),p.16-18

[4] IEEE (1993) *IEEE Standard for Software Maintenance*, IEEE, New York, 39 pp.

[5] Kemerer C.F. and Slaughter S.A. (1997) *Determinants of software maintenance profiles: An empirical investigation*, Journal of Software Maintenance, 9(4), p. 235-251

[6] Lientz, B.P. and Swanson E.B. (1980) *Software Maintenance Management*, Addison-Wesley, 1980

[7] Sneed H.M. (1996) *Modelling the maintenance process at Zurich Life Insurance*, International conference on Software Maintenance Proceedings, 1996, p. 217-226

[8] Spiegel M.R. and Stephens L.J. (1999) *Statistics, 3rd edition, (Schaum series)*, McGraw-Hill, New York.

[9] Swanson, E.B. (1976) *The Dimensions of Maintenance*, Proceedings on the second international conference on Software Engineering, pp 492-497, 1976

[10] Brown, Z. (2006) *Zack's Kernel News*, Linux Magazine, Issue 62, January 2006, www.linux-magazine.com

[11] Zvegintsov, N. (1998) *Personal communication*