Information and Software Technology 115 (2019) 93–96

Contents lists available at ScienceDirect

Information and Software Technology

journal homepage: www.elsevier.com/locate/infsof

Evaluating probabilistic software development effort estimates: Maximizing informativeness subject to calibration

Magne Jørgensen

Simula Metropolitan Center for Digital Engineering, Norway & Oslo Metropolitan University, Norway

ARTICLE INFO

Keywords:
Probabilistic effort estimates
Estimation error measurement
Effort prediction intervals
Estimated effort distributions

ABSTRACT

Context: Probabilistic effort estimates inform about the uncertainty and may give useful input to plans, budgets and investment analyses.
Objective & method: This paper introduces, motivates and illustrates two principles on how to evaluate the accuracy and other performance criteria of probabilistic effort estimates in software development contexts.
Results: The first principle emphasizes a consistency between the estimation error measure and the loss function of the chosen type of probabilistic single point effort estimates. The second principle points at the importance of not just measuring calibration, but also informativeness of estimated prediction intervals and distributions. The relevance of the evaluation principles is illustrated by a performance evaluation of estimates from twenty-eight software professionals using two different uncertainty assessment methods to estimate the effort of the same thirty software maintenance tasks.

1. Introduction

The results in [1] document a large variation in what software professionals mean with an "effort estimate" and that, typically, the meaning is neither communicated nor requested. This makes the situation in software contexts similar to what has been found in other estimation (forecasting) contexts where "…the common practice of requesting "some" point forecast, and then evaluating the forecasters by using "some" (set of) scoring function(s), is not a meaningful endeavor" [2]. This paper argues that a probabilistic framework enables more precise communication of effort estimates and aims at giving support on how to evaluate probabilistic effort estimates.

2. Evaluating single point, probabilistic effort estimates

A single point probabilistic effort estimate (*est*) is an effort value given probabilistic interpretation through reference to an estimated distribution of effort (F). Examples of such estimates are the most likely use of effort (the mode of F), the P50-estimate (the median of F), the P85-estimate (the 85% percentile of F) and the expected value (the mean of F). The effort estimates are given for a purpose, e.g., input to planning or cost-benefit analyses, with a connected loss function (L), e.g., the deviation between the estimated and the actual effort, which we try to minimize. Formally, the optimal single point estimate ($\widehat{est}$) from a distribution F is the one that minimizes the expected ($\mathbb{E}$) loss, i.e., $\widehat{est} = \arg min_{est}\ \mathbb{E}_F L(est, G)$, where G is the actual distribution of effort usage (the inherent uncertainty in use of effort). We apply this connection to propose the first evaluation principle.

> **Principle 1**: There should be a match (consistency) between the selected estimation error measure and the loss function of the type of probabilistic, single point effort estimate to be evaluated.

This principle has been suggested in earlier papers on evaluation of single point probabilistic estimates, see for example [2], and implies that proper estimation performance evaluation should be based on that: i) We know or assume the loss function for the estimation, ii) The loss function is represented by our estimation error measure, and iii) The type of effort estimates we request and evaluate is one that minimizes the expected loss when $F = G$. What is emphasized in the above principle is consequently not that the estimate actually minimizes the expected value of the loss function, only that we should evaluate the estimation performance relative to a match between what the estimator *believes* is the effort distribution and the loss function. An evaluation in accordance with the proposed principle rewards honest effort estimates, i.e., estimates that minimizes the loss function relative to what the estimator believes is the actual effort distribution *G* and not something else, and enables a rational evaluation of the estimation performance. Sometimes there may be practical problems related to following the principle. It may, for example, be difficult to formulate the loss function or more

E-mail address: magnej@simula.no

https://doi.org/10.1016/j.infsof.2019.08.006
Received 27 May 2019; Received in revised form 7 August 2019; Accepted 8 August 2019
Available online 8 August 2019
0950-5849/

Table 1
Matching error measure, loss function and types of probabilistic single point estimate.

Error measure	Loss function L	Type of single point effort estimates minimizing the loss function
Absolute Error (**AE**)	$L(\text{est,act}) = \lvert act - est \rvert$	* Judgment-based estimates reflecting "just as likely to spend more as to spend less" (median) effort than estimated. * Analogy-based estimates using the median value of a larger set of similar projects. * The output of linear regression-based models constructed from log-transformed values of *est* and *act* of historical projects (assuming log-normal distributions of *est* and *act*).
Inaccurate effort estimates (the inverse of **PRED**-types of measures)	$L_c(\text{est,act}) = \begin{cases} 1, & \lvert act - est \rvert > c, \\ 0, & otherwise \end{cases}$ where c is a constant	* Judgment-based estimates reflecting "most likely effort" (mode). * Analogy-based estimation models selecting the closest analogy of historical projects.
Squared Error (**SE**)	$L(\text{est,act}) = (act - est)^2$	* Analogy-based estimates using the mean value of a larger set of analogies. * The output of linear regression-based models constructed from historical values of *est* and *act*.
Mean Squared Error (**MSE**) of the sum of a set of estimates.	$L(\sum_1^n est_i, \sum_1^n act_i) = (\sum_1^n est_i - \sum_1^n act_i)^2$	* Task estimates added to find the total effort of a project (linearity of mean values). * Project estimates added to find the total effort of a portfolio.
Magnitude of relative error (**MRE**)	$\lvert \frac{act - est}{act} \rvert$	* Median of a random variable whose density if proportional to $f(act)/act$, where f is the density function of F (the median functional), see [2], i.e., a value *not* connected to any intuitive or commonly used interpretation of effort estimates.
Symmetric magnitude of relative error	$L(est, act) = \lvert \ln(\frac{est}{act}) \rvert = \lvert \ln(est) - \ln(act) \rvert = \lvert \ln(\frac{act}{est}) \rvert$	* Estimates aiming at minimizing the median of the expected percentage error, with symmetric penalty for effort over and under-runs. (The back-transformed measure $\zeta = e^{Md(\lvert \ln(\frac{est_i}{act_i}) \rvert)} - 1$ gives the median, symmetric magnitude of relative estimation error, see [3])

than one loss function that matters. The evaluation principle may nevertheless be useful to guide the estimation work and the performance evaluation. In particular it may be useful to avoid clear mismatch between what the estimates are meant reflect, e.g., the most likely use of effort, and how their performances are evaluated.

Table 1 displays a selection of effort estimation error measures with matching loss functions and single point estimates.

3. Evaluating effort prediction intervals and estimated effort distributions

The second evaluation principle, motivated below, is as follows:

> **Principle 2**: Effort prediction intervals and estimated effort distributions should be evaluated both regarding *calibration* with the actual effort distribution and the *informativeness* of the intervals and distributions.

This principle may be seen as an extension of the evaluation principle formulated in [4], i.e., that the goal to be evaluated against is that of: "Maximizing the sharpness of the predictive distributions subject to calibration". Our informativeness criterion is, however, wider (but also less precise) than the sharpness (i.e., narrowness of intervals or concentration of distribution) criterion. Our motivation for this extension is to include performance measures not clearly connected with sharpness, but still providing useful information. This is in particular the case for measures on how well an estimator separates high and low effort uncertainty situations. Frequently, both in experimental and in real-world contexts, see for example [5], the performance of probabilistic effort and cost estimates has been evaluated based on their calibration alone. Perfect calibration is, however, not sufficient for a good probabilistic effort estimate. It is for example possible to have properly calibrated P90-estimates of effort, without performing well on informativeness, e.g., when knowing that, historically, 90% of a certain types of software projects cost less than 1.000 work-hours and use 1.000 work-hours as the P90-estimate for all projects.

Table 2 suggests measures that cover both calibration and informativeness of effort prediction intervals and estimated distributions.

4. Illustration of the use of the estimation performance evaluation principles and measures

4.1. The data set[1]

Twenty-eight software professionals of a Norwegian software development organizations were invited to participate in a study on effort estimation. The participants were randomly divided into two groups: TRAD (traditional) and ALT (alternative). All of them were asked to estimate the most likely effort of the same set of thirty software maintenance tasks, previously completed in a Norwegian telecom organization. Those in the TRAD group were asked to estimate a 90% effort prediction interval (PI_{90}), while those in the ALT group were asked to estimate the probability (X-value of PI_X) to include the actual effort in the effort interval [0.5 × most likely effort; 2 × most likely effort]. After each task estimate and uncertainty assessment the participants received information about the actual effort used by the company, how much their estimate of most likely effort deviated and whether their effort prediction interval included the actual effort or not.

4.2. Evaluating single point estimates

In accordance with the first evaluation principle, the estimation performance measures should match a reasonable loss function of the type of estimate requested. Asking for the most likely effort matches a loss function minimizing the expected value of the proportion of inaccurate estimates, which may be represented by a PRED-type of error measure. We use PRED(25), defined as the proportion of projects with actual effort within ±25% of the estimated effort, for this purpose. Possibly, the participants were unaware of the loss function implications of requesting the most likely effort (or did not reflect on loss functions at all), and actually tried to minimize the distance between the estimated and the actual effort or the percentage error with equal emphasis on under- and

[1] The data we use for illustrative purpose are the same used for a different purpose in [7].

Table 2
Measures for the evaluation of effort prediction intervals and distributions.

$HitRate_X$	$\frac{1}{n}\sum_i h_i,\ h_i = \begin{cases} 1, & minimum_i \le act_i \le maximum_i \\ 0, & otherwise \end{cases}$, i.e., the proportion of observations where the actual effort is included in the PI_X (the X% effort prediction interval). It may be used on both centralized effort prediction intervals (minimum-maximum intervals) and PX-estimates of effort (i.e., *X*% confidence intervals with minimum 0 and maximum PX).
$PI_X Calib$	$HitRate_X - X$, i.e., the difference between the actual frequency ($HitRate_X$) and the estimated probability (confidence level *X*) of including the actual effort in the effort prediction interval.
$RWidth_X$	$\frac{PI_X}{\frac{1}{2}(maximum+minimum)}$, i.e., the prediction interval divided by its mid-point. A lower value means higher informativeness, of the effort prediction interval or PX-estimate.
$r_{RWidthX,err}$	Correlation between the relative width of effort prediction intervals and the estimation error. Very low, zero or negative values indicate low or no ability to separate high and low uncertainty situations. Notice that, given the random nature of use of effort, we cannot expect very high correlations even when $F = G$ (i.e., when we have perfect estimates of the outcome distribution).
CRPS	Continuous Ranked Probability Score. $CRPS(F, \text{act}) = \int_{\mathbb{R}} (F(z) - \mathbb{I}(\text{act} \le z))^2 dz$, where $\mathbb{I}(\text{act} \le z)$ is the identification function with value 1 when $act \le z$, otherwise 0. The value of CRPS (the integral) is minimized when F equals G. CRPS combines calibration and sharpness. A lower expected CRPS, for the same estimation tasks, indicates better estimation performance.
PIT	Probability Integral Transform: F(*act*), i.e., the probability value we get when giving the actual effort as input to our estimated (cumulative) effort distribution. Given perfectly calibrated estimated effort distributions, the distribution of PIT-values will be uniform, a ∩-shape indicates too wide (dispersed) and a ∪-shape a too narrow estimated effort distribution.
PIT-uniformity	Uniformity of a PIT-distribution. We suggest the use of *PIT-histograms* for this purpose, i.e., we count and display the number of observed PIT-values for categories of probabilities, as recommended in for example [6].

Table 3
Performance measurement of probabilistic effort estimates.

Person Id	Single point estimate accuracy			Prediction interval and distribution calibration and informativeness			
	Pred(25)	Mean MAE	ζ	P_{90}Calib	Mean $RWidth_{90}$	$Corr_{RWidth90,\zeta}$	Mean CRPS
id01 (T)	0.20	6.20	0.30	−0.07	1.36	−0.16	4.75
id02 (T)	0.33*	4.75	0.16*	−0.07	1.31	−0.09	3.96
id03 (T)	0.27	4.66	0.22	−0.20	1.14	−0.09	3.87
id04 (A)	0.27	4.58	0.30	0.01*	1.42	−0.19	3.55
id05 (A)	0.33	4.83	0.28	0.00*	1.34	0.05	4.19
id06 (T)	0.23	4.98	0.27	−0.37	0.95*	−0.14	4.33
id07 (T)	0.27	4.38	0.29	−0.17	1.14	−0.06	3.74
id08 (T)	0.27	4.76	0.26	−0.17	1.26	−0.27	4.15
id09 (A)	0.23	5.50	0.30	0.11	1.61	0.03	4.12
id10 (A)	0.27	5.40	0.30	−0.27	1.09*	−0.31	4.87
id11 (T)	0.27	6.17	0.17*	−0.33	1.17	0.20*	6.02
id12 (T)	0.33*	7.02	0.30	−0.23	1.24	−0.08	6.54
id13 (T)	0.27	4.11	0.18*	−0.80	0.20*	0.03	3.90
id14 (T)	0.23	5.03	0.30	−0.03*	1.62	−0.12	4.25
id15 (T)	0.27	4.32	0.28	−0.23	1.18	−0.12	3.56
id16 (T)	0.33*	3.73*	0.18*	−0.13	0.96	−0.30	3.05*
id17 (A)	0.30	5.38	0.23	−0.27	1.00*	−0.24	4.94
id18 (T)	0.27	5.36	0.30	−0.33	0.98*	−0.01	4.64
id19 (T)	0.17	4.00	0.23	−0.20	1.21	0.11	3.13*
id20 (A)	0.17	6.52	0.30	−0.09	1.42	−0.33	5.21
id21 (A)	0.37*	3.57*	0.21	−0.05	1.27	−0.31	3.01*
id22 (A)	0.33*	4.67	0.20	0.02*	1.32	−0.13	3.96
id23 (A)	0.27	4.07*	0.28	0.03*	1.44	0.28*	3.18*
id24 (A)	0.37*	4.07*	0.21	−0.04	1.35	0.18*	3.24
id25 (A)	0.27	5.18	0.30	−0.04	1.49	−0.25	3.94
id26 (A)	0.33*	3.63*	0.17*	0.25	1.62	0.30*	2.70*
id27 (A)	0.23	5.58	0.30	−0.03*	1.43	0.41*	4.22
id28 (A)	0.27	4.93	0.22	0.10	1.51	−0.21	3.71
Group							
TRAD	0.26	4.96	0.22	−0.24	1.12	−0.08	4.28
ALT	0.29	4.85	0.26	−0.02	1.38	−0.05	3.92

over-run. In this case, error measures with a fairer evaluation and better match with the loss function would be the mean MAE and the median relative error (ζ). We include these three error measures in Table 3, where the participants with (*) are those with the five most accurate (with ties included) single point estimates for each measure. As can be seen, it would be misleading to make claims about who gave more accurate estimates without stating which loss function the effort estimates were meant to minimize.

4.3. Evaluating effort prediction intervals and estimated distributions

In accordance with our second principle, we aimed at evaluating both calibration (P_{90}Calib) and informativeness (mean $RWidth_{90}$ and $Corr_{RWidth90,\zeta}$) of the effort prediction intervals. We also included the Continuous Ranking Probability Score (CRPS), which combines calibration and informativeness in one measure. All values are displayed in Table 3. To illustrate the use of PIT-histograms as a measure of

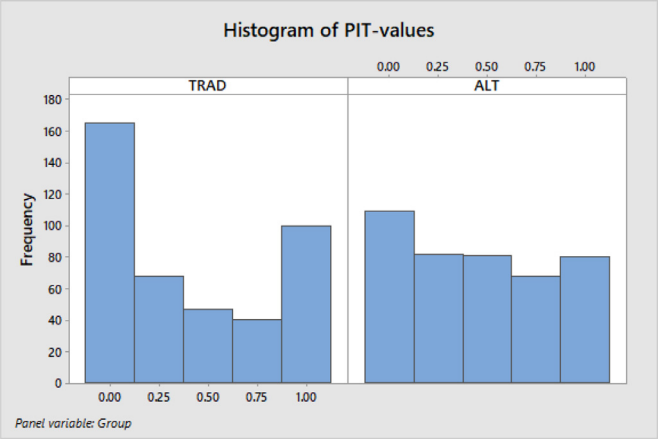

Fig. 1. Uniformity of the PIT-value distributions.

calibration, we display in Fig. 1 the histogram of the PIT-values of the two uncertainty elicitation groups. To identify the 90% effort prediction intervals of those in the ALT-group, and the estimated functions for the CRPS-values and the PIT-values we assumed a log-normal distribution and fitted this to the participants effort prediction intervals.

Table 3 and Fig. 1 show that the ALT group's effort prediction intervals were better *calibrated*, i.e., had P_{90}Calib closer to zero and more uniform PIT-values. When including measures of informativeness, however, the benefits of the alternative uncertainty elicitation method become less obvious. Those using the alternative method gave less informative effort intervals (wider mean $RWidth_{90}$) and were similarly poor at separating low and high uncertainty situations (low $Corr_{RWidth,\zeta}$). Statements about which was the best uncertainty assessment method should consequently also include how much we emphasize calibration compared to different types of informativeness.

Who of the participants were in total the best estimator? Not surprisingly, this depends on our loss function of the single-point estimates and how much we value calibration compared to informativeness for the prediction intervals. If we look at the number of times a participant had among the five best scores on a performance measures, it may be participant with id26. This participant scored well on all three single point estimation error measures, had uncertainty analysis that separated tasks with low and high uncertainty and had the best score on the combined informativeness and calibration measure CRPS. Even this participant, however, did not score well on all performance measures. An evaluation emphasizing the calibration only, which is what we did in [7], would rank him as just average.

Declaration of Competing Interest

None.

References

[1] M. Jørgensen, Communication of software cost estimates, 18th International Conference on Evaluation and Assessment in Software Engineering (EASE), ACM, 2014 28.
[2] T. Gneiting, Making and evaluating point forecasts, J. Am. Stat. Assoc. 106 (494) (2011) 746–762.
[3] S.K. Morley, T.V. Brito, D.T. Welling, Measures of model performance based on the log accuracy ratio, Space Weather 16 (1) (2018) 69–88.
[4] T. Gneiting, F. Balabdaoui, A.E. Raftery, Probabilistic forecasts, calibration and sharpness, J. R. Stat. Soc. 69 (2) (2007) 243–268.
[5] G.H. Volden, K. Samset, in: Quality Assurance in Megaproject Management, The Oxford Handbook of Megaproject Management, 2017, p. 406.
[6] T. Gneiting, M. Katzfuss, Probabilistic forecasting, Annu. Rev. Stat. Appl. 1 (2014) 125–151.
[7] M. Jørgensen, K.H. Teigen, *Uncertainty intervals* versus *interval uncertainty: an alternative method for eliciting effort prediction intervals in software development projects*, in: International Conference on Project Management (ProMAC), Singapore, 2002, pp. 343–352.