**When should we (not) use the mean magnitude of relative error (MMRE) as an error measure in software development effort estimation?**

Magne Jørgensen, Torleif Halkjelsvik, Knut Liestøl

**Abstract**

**Context**: The mean magnitude of relative error (MMRE) is an error measure frequently used to evaluate and compare the estimation performance of prediction models and software professionals.

**Objective**: This paper examines conditions for proper use of MMRE in effort estimation contexts.

**Method**: We apply research on scoring functions to identify the type of estimates that minimizes the expected value of the MMRE.

**Results**: We show that the MMRE is a proper error measure for estimates of the most likely (mode) effort, but not for estimates of the median or mean effort, provided that the effort usage is approximately log-normally distributed, which we argue is a reasonable assumption in many software development contexts. The relevance of the findings is demonstrated on real-world software development data.

**Conclusion**: MMRE is not a proper measure of the accuracy of estimates of the median or mean effort, but may be used for the accuracy evaluation of estimates of most likely effort.

# Introduction

When comparing the error of effort estimation models and when evaluating the estimates of software professionals, it is common to use the measure MMRE (Mean Magnitude of Relative Error):

, where *acti* and *esti* are the actual and the estimated effort for task *i*.

The popularity of the use of MMRE in software effort estimation contexts may be due to its interpretation as the mean absolute percentage error. MMRE is scale-free, which means that it can be used for the aggregation of estimation errors from software development projects of very different sizes. While MMRE has no upper score limits for over-estimation, under-estimation of effort will never give MMRE scores higher than one. In spite of this scoring asymmetry limitation, MMRE (MAPE) may be the most widely used estimation error measure in research and industry [1, 2].

This paper aims at identifying when it is meaningful to use the MMRE as an effort estimation error measure. The main premises of the analyses in the paper are the following:

* The effort required to complete a software task is uncertain and can be described by a probability (density) distribution. The actual use of effort can be viewed as a value randomly sampled from this effort probability distribution.
* A single-point effort estimate should aim at estimating a specific point of the outcome distribution, i.e., it should have a probabilistic interpretation. Common points of the outcome distribution subject to software development effort estimation are the mode (the most likely), the median and the mean effort. For example, effort estimates by software developers of smaller tasks may frequently give what they believe is the mode (most likely) effort [3], the median effort may be prescribed as estimates for the planned or budgeted efforts [4], and the estimates given as output from linear regression-based (Optimized Linear Square) estimation models or the PERT (Program Evaluation and Review Technique) approach aim at estimating the mean effort [5].
* A proper error measure, see for example [2], should give its lowest expected error when the estimated and actual point of the outcome distribution have the same probabilistic interpretation, e.g., when the estimated mean equals the actual mean of the effort usage distribution. In the (realistic) situation where the estimator does not know the probability distribution of the actual effort usage, the above criterion for a proper error measure, without changes in what is a proper measure, transforms into that of a match between what the estimator *believes* is the specified point of the effort distribution and the selected error measure [2].
* Not meeting the above criterion, either due to lack of match between type of estimate and error measure or due to evaluating estimates with unknown probabilistic interpretation, means that we will not be able to ensure a fair evaluation of the estimation error.

# The MMRE minimizing type of estimate

The MMRE is the mean of a measure we can refer to as the magnitude of relative error (MRE), where . Any measure that minimizes the expectation of MRE will, due to linearity of the mean, also minimize the expectation of the MMRE. For simplicity, we therefore omit the summation part of MMRE and focus on the MRE in the following.

As can be derived from [2, Theorem 5], the effort estimate that minimizes the expected value of MRE is the *median of the effort probability distribution (f) with density proportional to* , that is, for a normalising *C*, the median of a distribution .

It is not intuitive what this result means in terms of finding the point of the effort distribution that minimizes the expected value of MRE, and consequently that of the MMRE. To facilitate the analysis of properties of MRE, we therefore introduce the assumption that the effort probability distribution *f(act)* is lognormally distributed. Lognormal distributions are right-skewed distributions with a minimum of zero, flexible and well-known to give a good fit to many real-world phenomena, including a fit to probability distributions of effort usage [6].

With *f(act)* being a lognormal distribution with mean and standard deviation of the corresponding normal distribution we have that:

Applying Theorem 5 from [2] we know that the median *(m)* of minimizes the expected value of MRE. The median *m* is defined as the point of where the following is met:

We remove the constant C/, which occurs on both sides, and derive the following integral expression (by use of *wolframalpha.com*):

where *erf* is the Gauss error function.

Applying the integral expression in III) on the equation in II) and removing constants, the following must be true for *m* being the median of :

Since *erf (-)* = -1 and eq. IV implies that m must satisfy:

*V*) – (-1) = 1 – implying that must be zero. Since erf is monotonously increasing and erf(0) = 0, this imply that ln m = and thus m = which corresponds to the mode of the lognormal *f*. Due to the linearity of the expected values, the mode will also minimize the expected value of MMRE.

In spite of the widespread use of MMRE (MAPE) we are unaware of any prior research pointing out this property, which in many real-world contexts may be central for proper measurement of estimation error.

Knowing that the mode values of the lognormal distributions are lower than the median and mean values, we also know that evaluating estimates of the median and mean effort using MMRE will not be fair. Instead of rewarding accurate estimates of the median or the mode, MMRE will reward the under-estimation of these types of estimates, as pointed out in previous, simulation-based research [7]. For strongly right-skewed effort distributions the MMRE-minimizing effort estimates, which in the case of log-normal distributions is the mode, may be substantially lower than the median and the mean effort[[1]](#footnote-1).

# Illustrating the relevance on real-world software development effort data

As part of the study presented in [8], we collected the effort spent by twenty developers solving the same five tasks. Figure 1 shows a histogram of the distributions of effort, with a lognormal distribution fitted to the effort data (numbers in minutes).

We use this data to simulate a situation where estimates for five different tasks are required, but the tasks have not yet been assigned to a specific developer. We assume that the estimator has access to historical data for each task, as presented in Figure 1, and that these data properly reflect future software development effort for similar tasks in this company. Given the historical data, what should be stated as the intended type of estimate when evaluated by the MMRE? What about other estimates and error measures?

In Table 1, we illustrate the principle of matching the type of estimate to the error measure, and demonstrate that the MMRE can be a reasonable error measures when estimates are intended to reflect the mode of the outcome distribution, given that the effort distribution is approximately lognormal.



**Figure 1. Effort distributions of the five tasks with fitted log-normal distributions**

The analysis is based on the following procedure:

* The estimates (*esti(Mode)*) of the mode values of the five tasks (*i=1..5*) are derived from the mode from log-normal distributions fitted to the five effort distributions (using *@risk*). The estimates of the medians (*esti(Median)*) and the means (*esti(Mean)*) of the five tasks are calculated as the empirical median and mean values of the twenty actual efforts of each task.
* We assume a situation where the actual effort (*acti*) of each developer has the same probability to be sampled for each task.
* The error measures are selected to include one scale-free measure with expected error minimized by the mode when assuming lognormal distributions (MMRE), one absolute measure minimized by the median (Mean absolute Error, MAE = ), one scale-free measure minimized by the median (Mean Absolute Relative Log-error, MARL = ), and absolute measure minimized by the mean (Root Mean Square Error, RMSE=) . See [2] for documentation about the minimizing types of estimates for different error measures.

**Table 1. Estimation error** (best performing type of point estimate(s) in bold)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Task (i)** | **Type of estimate** | **Estimate** | **MMRE** | **MAE** | **MARL** | **RMSE** |
| Task 1 | Mode | 303 | **0.32** | 119 | 0.34 | 159 |
| Median | 355 | 0.34 | **110** | **0.31** | 145 |
| Mean | 373 | 0.37 | 114 | 0.32 | **144** |
| Task 2 | Mode | 176 | 0.33 | 120 | 0.45 | 178 |
| Median | 233 | **0.31** | **98** | **0.34** | 149 |
| Mean | 287 | 0.41 | 108 | 0.38 | **140** |
| Task 3 | Mode | 494 | **0.31** | 289 | 0.41 | 408 |
| Median | 622 | 0.34 | **262** | 0**.36** | 347 |
| Mean | 741 | 0.41 | 269 | 0.37 | **326** |
| Task 4 | Mode | 305 | **0.41** | 135 | 0.41 | 169 |
| Median | 326 | 0.43 | **133** | **0.40** | 162 |
| Mean | 367 | 0.50 | 136 | 0.41 | **157** |
| Task 5 | Mode | 200 | 0.51 | 312 | 0.83 | 393 |
| Median | 505 | **0.48** | **192** | **0.40** | **239** |
| Mean | 509 | 0.49 | 193 | **0.40** | **239** |

From Table 1, we observe that the estimates of the mode gave better MMRE for Tasks 1, 3 and 4, but not for Tasks 2 and 5 (although close to the best). By inspecting the distributions in Figure 1, we observe that the shapes of the effort distributions for Tasks 1, 3 and 4 are similar to those of lognormal distributions, whereas those of Tasks 2 and 5 deviate more. While a lognormal distribution is flexible, it may not be optimal fit for all effort outcome distributions.

The theoretically minimizing types of estimates for MMRE (mode for lognormal distributions), MAE (median), MARL (median) and RMSE (mean), correspond well with what is actually minimized in Table 1. This support the use of different types of accuracy measures for different types of estimates, and underlines the importance of knowing what types of estimates one is evaluating and select a proper error measure. Evaluating estimates of the mean effort using MMRE will typically not enable a fair evaluation, while evaluating estimates of the most likely (mode) may frequently be acceptable.

While the estimates of the mode, the median and mean effort in Table 1 tend to be quite different (the mode is for example on average only 66% of the mean effort), the MMRE-values are deviating less (the MMRE of the mode is on average 86% of the mean effort). This observation, does not imply that the use of the MMRE can safely be extended to be used for the evaluation of estimates of the median or the mean effort. The MMRE may in other contexts differ substantially more.[[2]](#footnote-2) In addition, it is unfortunate to create estimation situations that reward underestimation.

# References

1. Idri, A., M. Hosni, and A. Abran, *Systematic literature review of ensemble effort estimation.* Journal of Systems and Software, 2016. **118**: p. 151-175.

2. Gneiting, T., *Making and evaluating point forecasts.* Journal of the American Statistical Association, 2011. **106**(494): p. 746-762.

3. Jørgensen, M. *Communication of software cost estimates*. in *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*. 2014.

4. Samset, K.F., et al., *Governance schemes for major public investment projects: A comparative study of principles and practices in six countries*. 2016, Ex ante akademisk forlag.

5. Golenko-Ginzburg, D., *On the distribution of activity time in PERT.* Journal of the Operational Research Society, 1988. **39**(8): p. 767-771.

6. Halkjelsvik, T. and M. Jørgensen, *Time Predictions: Understanding and Avoiding Unrealism in Project Planning and Everyday Life*. Vol. 5. 2018: Springer.

7. Foss, T., et al., *A simulation study of the model evaluation criterion MMRE.* IEEE transactions on software engineering, 2003. **29**(11): p. 985-995.

8. Jørgensen, M. and T.M. Gruschke, *The impact of lessons-learned sessions on effort estimation and uncertainty assessments.* IEEE Transactions on Software Engineering, 2009. **35**(3): p. 368-383.

1. The robustness of our finding was evaluated by simulating a variety of right-skewed distributions of the types: beta, gamma and triangular. For these distributions, even the mode values tended to be too high to minimize the MMRE. Even in these cases, however, the MMRE was lower for the mode than for the median or the mean. The implication that MMRE should mainly be used as a measure of mode-accuracy is consequently to some extent supported, even for non-lognormal, right-skewed distributions. [↑](#footnote-ref-1)
2. As an illustration of a situation where the MMRE may differ more, we applied the Cocomo81-dataset and built an OLS-based linear regression model (which aims at estimating the mean effort). The MMRE of the effort estimates of this model was compared with estimates of the mode, assuming that the mode was the same as the proportion of the mean effort as in Table 1, i.e., 66%. The use of the estimates of the mode gave in this case a reduction of the MMRE as high as 33%. [↑](#footnote-ref-2)