# Accuracy Evaluation of Model-based COSMIC Functional Size Estimation

Luigi Lavazza

Dipartimento di Scienze Teoriche e Applicate
Università degli Studi dell'Insubria
Varese (Italy)
Email: luigi.lavazza@uninsubria.it

*Abstract*—**Functional Size Measurement is widely used, especially to quantify the size of applications in the early stages of development, when effort estimates are needed. However, the measurement process is often too long or too expensive, or it requires more knowledge than available when development effort estimates are due. To overcome these problems, early size estimation methods have been proposed, to get approximate estimates of functional measures. In general, early estimation methods adopt measurement processes that are simplified with respect to the standard process, in that one or more phases are skipped. So, the idea is that you get estimates affected by some estimation error, instead of accurate measures performed following the standard measurement process, but at a fraction of the cost and time required for standard measurement. In this paper, we consider some methods that have been proposed for estimating the COSMIC (Common Software Measurement International Consortium) size of software during the modeling stage. We apply the most recent methodologies for estimation accuracy, to evaluate whether early model-based estimation is accurate enough for practical usage.**

***Keywords*–*Functional size measurement; COSMIC Function Points; Measurement process; Functional size estimation; Accuracy estimation.***

## I. Introduction

Functional Size Measurement (FSM) is widely used. Among the reasons for the success of FSM is that it can provide measures of size in the early stages of software development, when they are most needed for cost estimation. However, FSM requires that the functional requirements of the application to be measured are available in a complete and quite detailed form. Often, this is not possible in the very early stages of development. Therefore, to get measures even when requirements are still incomplete or still defined at a coarse level of detail, estimation models have been proposed. There are different types of FSM and many estimation methods. Here, we concentrate on the COSMIC FSM [1] –one of the most widely used methods– and on model-based COSMIC size estimation [2].

When applying a size estimation method, we expect that the method –being applied to incomplete or not thoroughly detailed software specifications– requires less time and effort than the standard measurement process. However, we also expect that the size estimates so obtained contain some estimation error. In general, we are ready to accept a relatively small estimation error in exchange of being able to get size estimates without having to apply the standard measurement process. On the contrary, an excessively large estimation error would defeat the very reason for measuring. Hence, we are interested in knowing the likely accuracy of measure estimates. To this end, we need reliable methods to evaluate the accuracy of estimates.

Unfortunately, it has been shown that the most popular estimate accuracy statistic, the Mean Magnitude of Relative Errors (MMRE) is flawed, in that it is a biased estimator of central tendency of the residuals of a prediction system because it is an asymmetric measure [3][4][5]. So, MMRE and similar indicators are not suitable for providing practitioners who are potentially interested in applying estimate methods with reliable information upon which they can base informed decisions.

Luckily, sound estimate evaluation methods have been proposed recently (as described in Section III). It is thus possible to apply such new methods to size estimation methods.

The main purpose of this paper is the evaluation of the actual accuracy of model-based COSMIC size estimation method: to this end, we use the new sound evaluation methods (described in Section III), together with more traditional statistical tools.

It should be noted that the paper does not aim at introducing new COSMIC size estimation methods, rather the goal of the paper is (re)evaluating the accuracy of the formerly [2] proposed ones. However, by applying these new evaluation methods, as a side effect we also get some indications on their expressiveness.

The paper is structured as follows. Section II briefly illustrates the COSMIC FSM, and the model-based simplified COSMIC measurement method. Section III illustrates the methods used for evaluating the accuracy of estimates. Section IV describes the application of the accuracy evaluation methods to model-based simplified COSMIC measurement, while Section V illustrates and discusses the results of the analysis. Section VI accounts for related work. Finally, Section VII draws conclusions and briefly sketches future work.

## II. COSMIC Functional Size Measurement and Model-Based COSMIC Estimation

COSMIC measurement is based on the analysis of the specification of functional user requirements (FUR). The FUR can be described in various ways, including the Unified Modeling Language (UML): functional size measurement of UML models was widely studied [6][7][8], also when FUR concern real-time applications [9]. During the initial stage of development, UML models are built, progressively incorporating more knowledge concerning the software to be developed: this results in progressively more complete and detailed specifications. More specifically, the UML modeling

ISBN: 978-1-61208-590-6

process can be seen as organized in the phases described in Figure 1. The more complete and detailed the UML model, the more elements needed for COSMIC measurement become available. Figure 1 shows the relationship between the UML diagrams that are made available by each modeling phase and the COSMIC measurement elements. During the initial UML modeling phases –i.e., before the complete and detailed FUR specifications are available– it is often the case that size measures are needed anyway. In such cases –not being possible to measure the COSMIC size of the application– we can think of *estimating* the COSMIC size, based on the information that is present in the available UML diagrams.

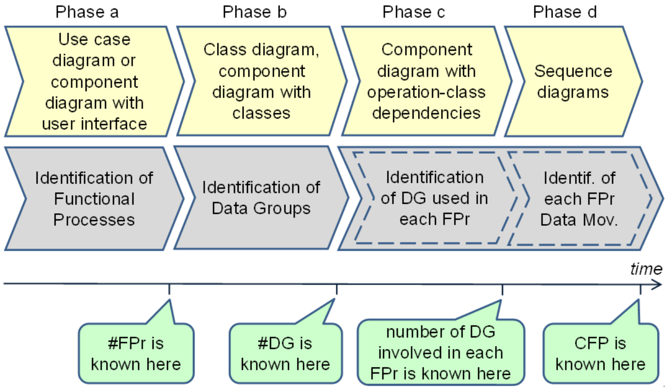
Phase a
Phase b
Phase c
Phase d
Use case diagram or component diagram with user interface
Class diagram, component diagram with classes
Component diagram with operation-class dependencies
Sequence diagrams
Identification of Functional Processes
Identification of Data Groups
Identification of DG used in each FPr
Identif. of each FPr Data Mov.
time
#FPr is known here
#DG is known here
number of DG involved in each FPr is known here
CFP is known here

Figure 1. UML modeling process and COSMIC measurement process phases.

Specifically, del Bianco et al. proposed a few families of statistical models that can be used to estimate COSMIC size based on information derived from UML diagrams [2]. These models are described in Table I.

A first family of COSMIC size estimation models requires only the knowledge of the number of functional processes (FPr's). These models have form ECFP = f(#FPr) where ECFP is the estimated size in CFP (COSMIC Function Points), and #FPr is the number of functional processes. As shown in Figure 1, the statistical model can be built after the completion of phase a), when class or component diagrams properly specifying the user interfaces are delivered.

Another family of COSMIC size estimation models requires also that the number of Data Groups (#DG) is known. These models can be built after phase b), when UML diagrams fully describing the involved classes are delivered. The models found by del Bianco et al. involve the parameter AvDGperFPr, namely the average number of data groups per functional process, which requires that both the functional process and the data groups (i.e., classes in UML diagrams) are known.

Figure 1 shows that potentially one could use also the knowledge of the number of data groups involved in each functional process, which is available after phase c). However, no statistically significant models of this type were found.

Finally, we observe that after phase d), i.e., when the complete UML models of FUR are available, the standard COSMIC measurement process is applicable, and proper COSMIC measures –instead of estimates– can be achieved.

It is expected that models based on more information are more accurate than models based on less information.

TABLE I. COSMIC SIZE ESTIMATION MODELS.

| *Name* | *Formula* |
|---|---|
| avg1 | $ECFP = 7.3 \ \#FPr$ |
| reg1 | $ECFP = -16.5 + 6.698 \ \#FPr$ |
| avg2 | $ECFP = AvDGperFPr \ 1.8 \ \#FPr$ |
| reg2 | $ECFP = -64.6 + 7.63 \ \#FPr + 9.71 \ AvDGperFPr$ |
| log2 | $ECFP = 1.588 \ \#FPr^{1.00357} \ AvDGperFPr^{1.0312}$ |

In [2], the accuracy of the models given in Table I was evaluated based on the traditional indicators MMRE –the Mean Magnitude of Relative Errors– and Pred(25) –the fraction of applications for which the absolute relative estimation error is less than 0.25. The evaluation of accuracy performed in [2] indicated that models using both #FPr and AvDGperFPr (that is, models avg2, reg2 and log2) are more accurate than models based only on #FPr (that is, models avg1 and reg1). However, it has been shown that indicators based on the magnitude of relative errors are biased [10]. Hence, we repeat here (in Section IV) the analysis of accuracy using more reliable methods (described in Section III).

## III. A Method for Evaluating the Accuracy of Estimates

The method we use for evaluating the accuracy of a given model's estimates involves two activities: 1) comparing the model's estimates with "baseline" estimates, and 2) evaluating the size effect. The former activity –described in Section III-A– is aimed at verifying that the given model's estimates are "good enough:" if they are less accurate than the estimates provided by the baseline, the given models does not yield any improvement, at least as far as accuracy is concerned, and can be rejected. The second activity –described in Section III-B– verifies whether the given model provides an increase in accuracy that is large enough to make the given model a desirable alternative with respect to the baseline.

### A. *Baselines*

Let us suppose that in n previous projects we measured the value of interest (in our case, the size of applications, measured in CFP). Accordingly, we have a set $Y = \{y_i\}$ (with $i \in [1, n]$) of observations (where $y_i$ is the actual size of the $i^{th}$ project, expressed in CFP).

A new estimation method $P$ is proposed: for the $n$ known projects, method $P$ yields $n$ estimates $\hat{y}_i$ with $i \in [1, n]$, and we need to evaluate the accuracy of these estimates.

The most popular way of evaluating estimation accuracy is the MMRE, the mean of the magnitude of absolute errors, which is defined as

$$MMRE = \frac{1}{n} \sum_{i=1..n} \frac{|y_i - \hat{y}_i|}{y_i} \qquad (1)$$

MMRE has been shown to be a biased estimator of central tendency of the residuals of a prediction system, because it is an asymmetric measure [3], [4], [5]. In practice, MMRE is biased towards prediction systems that under-estimate [10].

Shepperd and MacDonell [10] proposed that the accuracy of a given estimation method $P$ is compared to the accuracy of a reference estimation method $P0$. The indicator to be used

ISBN: 978-1-61208-590-6

is the Mean Absolute Residual (MAR), which, unlike MMRE, is not biased:

$$MAR = \frac{1}{n} \sum_{i=1..n} |y_i - \hat{y}_i|) \quad (2)$$

So we have $MAR_P$ (the MAR of the proposed method) and $MAR_{P0}$ (the MAR of the reference method). Based on the MAR values, Shepperd and MacDonell propose to compute a Standardized Accuracy measure (SA) for estimation method $P$:

$$SA = 1 - \frac{MAR_P}{MAR_{P0}} \quad (3)$$

Values of SA close to 1 indicate that $P$ outperforms $P0$, values close to zero indicate that $P$'s accuracy is similar to $P0$'s accuracy, negative values indicate that $P$ is worse than $P0$, hence it should be rejected.

As a referenced model, Shepperd and MacDonell suggest to use random estimation, based on the known (actual) values of previously measured projects. A random estimation $\hat{y}_i$ is obtained by picking at random $y_j$, with $j \neq i$. Of course, in this way there are $n-1$ possible estimates for $y_i$, so to compute the MAR of $rnd$ we need to average all these possible values. Shepperd and MacDonell suggest to make a large number of random estimates (typically, 1000), and then take the mean $\overline{MAR_{rnd}}$. Langdon et al. showed that it is not necessary to make 1000 guesses, since the average of the random estimates can be computed exactly [11].

So, a first evaluation consists in computing

$$SA = 1 - \frac{MAR_P}{\overline{MAR_{rnd}}}. \quad (4)$$

Achieving a value substantially greater than zero is clearly a sort of necessary condition that the estimation method $P$ must satisfy, otherwise we could simply guess (instead of estimating using $P$) and get similarly accurate estimates.

Lavazza and Morasca [12] observed that the comparison with random estimation is not very effective in supporting the evidence that $P$ is a good estimation model. Instead, they proposed to use a "constant model" (*CM*), where the estimate of the size of the $i^{th}$ project is given by the average of the sizes of the other projects, that is

$$\hat{y}_i = \frac{1}{n-1} \sum_{j \in Y - \{y_i\}} y_j \quad (5)$$

So, we can compute the $MAR_{CM}$ of these estimates, and then compute SA, but this time comparing $P$ with $CM$:

$$SA = 1 - \frac{MAR_P}{MAR_{CM}}. \quad (6)$$

Again, we require that SA is substantially greater than zero, to deem $P$ acceptable.

Finally, note that SA can be used to compare a method $P$ against any other model $P1$ used as a reference method, simply by computing

$$SA = 1 - \frac{MAR_P}{MAR_{P1}}. \quad (7)$$

### B. Size Effect

Suppose that we have two estimation methods $P1$ and $P2$, and $MAR_{P2} < MAR_{P1}$ (hence, $SA = 1 - \frac{MAR_{P2}}{MAR_{P1}} > 0$). We can conclude that $P2$ is more accurate than $P1$. Anyway, suppose that we are using $P1$ and we are considering the possibility of switching to using $P2$, which involves some effort, because $P2$ requires some activity or data or programs that $P1$ does not require. We would like to know if the improvement that $P2$ offers in terms of accuracy is possibly so inconsequential as to not be worth the effort.

To judge the effect size, Shepperd and MacDonell suggest using Glass's $\Delta$ [13] or Hedges's $g$, which might be preferred when the sample size is small [14]. The effect size –which is scale-free– can be interpreted in terms of the categories proposed by Cohen [15] of small (≈ 0.2), medium (≈ 0.5) and large (≈ 0.8).

## IV. Experimental Evaluation

The five size estimation models given in Table I were applied to the projects in the dataset that was used to derive the models [2]. The MAR for each model was then computed. Similarly, the data from the same dataset were used to compute $\overline{MAR_{rnd}}$ and $MAR_{CM}$, as described in Section III. The values of the methods' MAR are given in Table II.

Note that here we do not explicitly compute SA. Instead, we give the values of MAR needed for the computation. The reason is that with 7 methods there are 21 possible comparison among methods, hence 21 values of SA. Listing all these SA values could create confusion, while to compare two methods' accuracies, we just need to compare their SA's: the model featuring the smaller SA is likely the best.

TABLE II. Mean absolute residuals of models.

| *Name* | *Formula* | *MAR* |
|---|---|---|
| rnd | – | 146 |
| CM | – | 114 |
| avg1 | $ECFP = 7.3\ \#FPr$ | 56 |
| reg1 | $ECFP = -16.5 + 6.698\ \#FPr$ | 48 |
| avg2 | $ECFP = AvDGperFPr1.8\ \#FPr$ | 28 |
| reg2 | $ECFP = -64.6 + 7.63\ \#FPr + 9.71\ AvDGperFPr$ | 40 |
| log2 | $ECFP = 1.588\ \#FPr^{1.00357}\ AvDGperFPr^{1.0312}$ | 25 |

Table II provides a first piece of evidence: model-based COSMIC size estimation are definitely more accurate than both the random and constant models.

Table II also confirms that the constant model is always more accurate than the random model, as demonstrated by Lavazza and Morasca [12]. For this reason, in the remainder of the paper the random model is no longer used.

To establish if the estimations of one method were significantly better than the estimations provided by another method, we tested the statistical significance of the absolute errors achieved with the two estimation methods [3]. Namely, we compared the absolute residuals provided by every pair of methods via Wilcoxon Sign Rank Test. To check for statistical significance we used the Wilcoxon Signed Rank Test [16] because it is a safe test to apply to both non-normally distributed (as are often MAR distributions) and normally distributed populations.

The results are given in Table III, where in each cell the sign ">" (respectively, "<", "=") indicates that the absolute residuals of the model on the cell's row are larger (resp., smaller, equal) than the absolute residuals of the model on the cell's column.

TABLE III. COMPARISON OF ABSOLUTE RESIDUALS USING WILCOXON SIGN RANK TEST.

| | CM | avg1 | reg1 | avg2 | reg2 | log2 |
|---|---|---|---|---|---|---|
| CM | | $>$ | $>$ | $>$ | $>$ | $>$ |
| avg1 | $<$ | | $>$ | $>$ | $>$ | $>$ |
| reg1 | $<$ | $<$ | | = | $>$ | $>$ |
| avg2 | $<$ | $<$ | = | | = | $>$ |
| reg2 | $<$ | $<$ | $<$ | = | | = |
| log2 | $<$ | $<$ | $<$ | $<$ | = | |

The results provided by Wilcoxon Sign Rank Test confirm the indications provided by MAR and SA in that the constant model is outperformed by all other models and that avg1 is outperformed by all other model-based size estimation methods. However, Wilcoxon Sign Rank Test provides further insights with respect to MAR and SA:

- There is no sufficient evidence to conclude that log2 is better than reg2 (this fact could be guessed, based on the fact that $MAR_{log2}$ and $MAR_{reg2}$ are quite close).
- Similarly, there is no evidence that reg2 (which has $MAR_{reg2} = 40$) is actually more accurate than reg1 (which has $MAR_{reg1} = 48$).
- Somewhat surprisingly, there is no evidence that avg2 (which has $MAR_{avg2} = 28$) is actually more accurate than reg2 (which has $MAR_{reg2} = 40$).

The latter result is especially interesting, in that by just looking at the MAR values we could have concluded that avg2 is more accurate than reg2, while –according to Wilcoxon Sign Rank Test– there is no statistically significant evidence of this fact. The explanation of why MAR can be somewhat misleading in this case is given in Figure 2, where the boxplots of the absolute residuals of models avg2 and reg2 are given: it is easy to see that the distributions are similar, but reg2 has a greater MAR because of three applications, whose size estimation error is quite large.

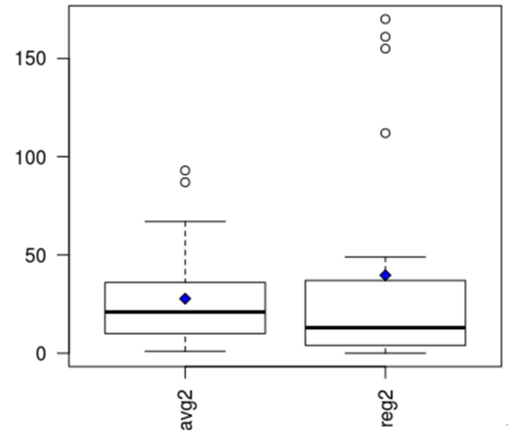

Figure 2. Absolute residuals of models avg2 and reg2.

Now, as recommended by Shepperd and MacDonell (see Section III-B) we evaluate the size effect. To this end, we computed Hedges's $g$ for all model pairs. The results are given in Table IV.

TABLE IV. EFFECT SIZE (HEDGES'S $g$).

| | CM | avg1 | reg1 | avg2 | reg2 | log2 |
|---|---|---|---|---|---|---|
| CM | – | 0.75 | 0.82 | 1.30 | 0.98 | 1.36 |
| avg1 | -0.75 | – | 0.12 | 0.58 | 0.27 | 0.66 |
| reg1 | -0.82 | -0.12 | – | 0.38 | 0.13 | 0.44 |
| avg2 | -1.30 | -0.58 | -0.38 | – | -0.26 | 0.12 |
| reg2 | -0.98 | -0.27 | -0.13 | 0.26 | – | 0.33 |
| log2 | -1.36 | -0.66 | -0.44 | -0.12 | -0.33 | – |

It is easy to see that all model-based size estimation methods appear definitely preferable with respect to the constant model. Models avg2 and log2 appear preferable to the other model-based estimation methods, with log2 only marginally better than avg2.

The indications provided by Hedges's $g$ are also consistent with the indications obtained from Wilcoxon Sign Rank Test, e.g., according to Hedges's $g$ avg2 is only marginally better than reg2.

## V. DISCUSSION OF RESULTS

With reference to Figure 1, at the end of phase a), we know the number of Functional Processes (#FPr), thus models avg1 and reg1 are applicable. At the end of phase b), the other models are also applicable.

According to the analysis of experimental data, we have that the models that are applicable at the end of phase b) are –to different extents– more accurate than the models that are applicable at the end of phase a). This was expected, since by progressing from phase a) to phase b), more information concerning the application is made available through UML models, thus we can exploit this information to achieve more accurate size estimates. However, having reliable empirical evidence that progressing trough application modeling phases enable the construction of progressively more accurate models of the functional size is quite important. It also indicates that collecting measures of COSMIC elements (especially #FPr and #DG, hence AvDGperFPr) and building several statistical models of COSMIC size is useful to get a progressively more accurate notion of the size of the application being built. Actually, the size effect indicators (see Table IV) suggest that the models available at the end of phase b) allow only for a medium-small improvement over the best model available at the end of phase a), especially as far as reg1 is concerned. However, to achieve this moderate improvement, all you have to do is counting the data group (i.e., classes in UML models): since this counting is very easy (it can even be automated) building more accurate models at the end of phase b) is not only possible, but most probably always convenient.

Like in any empirical study, we have to deal with some threats to the validity of our analysis.

We see no construction issues with our analysis, since all the used techniques are statistically sound; in fact, they have been proposed to correct the problems with previous indicators, such as MMRE.

The main problem we face is probably the generalizability of results. In fact, our results derive from the analysis of a dataset that collects data from only 21 projects. It is possible that other datasets could support somewhat different conclusions. However, the fact that our dataset includes several industrial projects, and that the size of the dataset is not excessively small (especially in the context of empirical software engineering studies) supports the hypothesis the results presented here are sufficiently representative in general. Also, the logical coherence of the results –namely the fact that the more information is available from UML models, the more accurate is size estimation– supports the hypothesis the results presented here are valid.

## VI. Related Work

The accuracy evaluation techniques used in this paper are being increasingly used by researchers that need to evaluate the accuracy of new effort estimation proposals. For instance, Sarro et al. used the Mean Absolute Error and the Standardized Accuracy to assess the accuracy of a bi-objective effort estimation algorithm that combines confidence interval analysis and assessment of mean absolute error [17]. To establish if the estimations of one method were significantly better than the estimations provided by another method, they tested the statistical significance of the absolute errors achieved with different estimation methods via the Wilcoxon Signed Rank Test, as we did in Section IV.

The techniques used here are becoming quite popular, but there are also several alternative proposal, actually too many to be mentioned here. As an example of an alternative to SA, Tofallis proposed to use the logarithm of the accuracy ratio: $log\frac{prediction}{actual}$ [18]. As an example of an alternative to Hedges's $g$, Vargha and Delaney proposed the A12 statistic, a non-parametric effect size measure: given a performance measure M, A12 indicates the probability that running algorithm A yields higher M values than running another algorithm B [19]. Finally, a quite different but interesting proposal is StatREC, a Graphical User Interface statistical toolkit designed to provide a variety of graphical tools and statistical hypothesis tests to facilitate strategies for an intelligent decision-making [20].

Concerning the assessment of accuracy of functional size estimation methods, to the best of the author's knowledge, very little work has been done. In general, some evaluation is done when a method is proposed, as in [21], where the NESMA estimated method is proposed and its accuracy is evaluated on the training set. A noticeable exception is [22], where several early estimation methods for Function Point measures are evaluated via an empirical study.

## VII. Conclusion

In this paper, the accuracy of a set of model-based methods to estimate the COSMIC size of software applications has been evaluated. The relevance of the paper is based on two factors:

- For practitioners (as well as for researchers) knowing the accuracy that can be achieved via size estimation methods is very important. Consider for instance that the application of the considered size estimation methods could provide the most important piece of information upon which the cost of software is estimated.
- To evaluate the accuracy of estimates, you need reliable indicators. Traditional indicators like MMRE have been proved to be biased. So, finding and testing more reliable indicators is necessary. Consider for instance a new estimation technique proposed by some researchers: how can they confidently claim that their new technique is good, and possibly even better than existing techniques? They need reliable accuracy evaluation techniques and indicators.

According to our empirical study, we can recommend that the accuracy of estimates be evaluated by

- Computing the mean of absolute residuals (MAR) of all the models to be tested.
- For any estimation method, doing better than the baseline models (the constant model and the random model) is a must. Hence, one should always test models against the constant model. In addition, one should also evaluate new estimation methods against the currently used estimation technique, to see it the change is worthwhile.
- Using Wilcoxon Sign Rank Test is advisable, since it can give statistically significant indications that are particularly informative when two methods' MAR values are close.
- Also looking at the boxplots of absolute residuals can help, especially when a few outliers affect the MAR at a great extent (as in Figure 2).
- Finally, assessing the effect size using Hedges's $g$ (or similar indicators) is useful to assess the extent of the improvement that a new technique can guarantee over another one.

When evaluating the accuracy of model-based COSMIC size estimation methods, we got easily quite representative indications via the MAR, as shown in Table II. By means of more sophisticated statistical tools –such as the Wilcoxon Sign Rank Test and Hedges's $g$– we achieved indications that are slightly more informative, e.g., that there is no statistically significant evidence that the log 2 model is more accurate than the reg2 model.

As a final observation, we note that the analyses reported in this paper were carried out quite easily via simple R [23] programs. So, practitioner and researchers that need to evaluate estimation accuracy can invest a small amount of effort to program a few hundred lines of R code that will make the analysis reported here totally automatic.

Future work includes:

- Further evaluating model-based COSMIC size estimation methods via additional evaluation methods and against additional datasets.
- Experimenting the accuracy evaluation methods used in this paper with other estimation techniques and using other datasets.

## Acknowledgment

The work presented here has been partly supported by the "Fondo di Ricerca d'Ateneo" of the Università degli Studi dell'Insubria.

## References

[1] The COSMIC consortium, Functional Size Measurement Method Version 4.0.1 Measurement Manual (The COSMIC Implementation Guide for ISO/IEC 19761:2011), 2015.

[2] V. Del Bianco, L. Lavazza, G. Liu, S. Morasca, and A. Z. Abualkishik, “Model-based early and rapid estimation of cosmic functional size–an experimental evaluation,” Information and Software Technology, vol. 56, no. 10, 2014, pp. 1253–1267.

[3] B. A. Kitchenham, L. M. Pickard, S. G. MacDonell, and M. J. Shepperd, “What accuracy statistics really measure,” IEE Proceedings-Software, vol. 148, no. 3, 2001, pp. 81–85.

[4] T. Foss, E. Stensrud, B. Kitchenham, and I. Myrtveit, “A simulation study of the model evaluation criterion MMRE,” IEEE Transactions on Software Engineering, vol. 29, no. 11, 2003, pp. 985–995.

[5] I. Myrtveit, E. Stensrud, and M. Shepperd, “Reliability and validity in comparative studies of software prediction models,” IEEE Transactions on Software Engineering, vol. 31, no. 5, 2005, pp. 380–391.

[6] K. Berg, T. Dekkers, and R. Oudshoorn, “Functional size measurement applied to UML-based user requirements,” 2005, pp. 69–80.

[7] L. A. Lavazza, V. Del Bianco, and C. Garavaglia, “Model-based functional size measurement,” in Proceedings of the Second ACM-IEEE international symposium on Empirical software engineering and measurement. ACM, 2008, pp. 100–109.

[8] A. Živkovič, I. Rozman, and M. Heričko, “Automated software size estimation based on function points using UML models,” Information and Software Technology, vol. 47, no. 13, 2005, pp. 881–890.

[9] L. Lavazza and V. Del Bianco, “A case study in COSMIC functional size measurement: The rice cooker revisited,” Software Process and Product Measurement, 2009, pp. 101–121.

[10] M. Shepperd and S. MacDonell, “Evaluating prediction systems in software project estimation,” Information and Software Technology, vol. 54, no. 8, 2012, pp. 820–827.

[11] W. B. Langdon, J. Dolado, F. Sarro, and M. Harman, “Exact mean absolute error of baseline predictor, MARP0,” Information and Software Technology, vol. 73, 2016, pp. 16–18.

[12] L. Lavazza and S. Morasca, “On the evaluation of effort estimation models,” in Proceedings of the 21st International Conference on Evaluation and Assessment in Software Engineering. ACM, 2017, pp. 41–50.

[13] R. Rosenthal, H. Cooper, and L. Hedges, “Parametric measures of effect size,” The handbook of research synthesis, 1994, pp. 231–244.

[14] P. D. Ellis, The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results. Cambridge University Press, 2010.

[15] J. Cohen, “A power primer.” Psychological bulletin, vol. 112, no. 1, 1992, pp. 155–159.

[16] ——, Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Lawrence Earlbaum Associates, 1988.

[17] F. Sarro, A. Petrozziello, and M. Harman, “Multi-objective software effort estimation,” in Proceedings of the 38th International Conference on Software Engineering. ACM, 2016, pp. 619–630.

[18] C. Tofallis, “A better measure of relative prediction accuracy for model selection and model estimation,” Journal of the Operational Research Society, vol. 66, no. 8, 2015, pp. 1352–1362.

[19] A. Vargha and H. D. Delaney, “A critique and improvement of the cl common language effect size statistics of mcgraw and wong,” Journal of Educational and Behavioral Statistics, vol. 25, no. 2, 2000, pp. 101–132.

[20] N. Mittas, I. Mamalikidis, and L. Angelis, “A framework for comparing multiple cost estimation methods using an automated visualization toolkit,” Information and Software Technology, vol. 57, 2015, pp. 310–328.

[21] H. van Heeringen, E. van Gorp, and T. Prins, “Functional size measurement-accuracy versus costs-is it really worth it?” in Software Measurement European Forum (SMEF), 2009.

[22] L. Lavazza and G. Liu, “An empirical evaluation of simplified function point measurement processes,” International Journal on Advances in Software, vol. 6, no. 1 & 2, 2013, pp. 1–13.

[23] R Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, 2014.