Evaluating the Accuracy of Estimates: the Case 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 –at a fraction of the cost and time required for standard measurement– size estimates affected by some estimation error, instead of accurate measures performed following the longer and more expensive standard measurement process. 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. The contribution of the paper is twofold: on the one hand we provide a reliable evaluation of the accuracy that can be obtained when estimating the functional size of software applications based on UML models; on the other hand, we get indications concerning the effectiveness and expressiveness of recently proposed accuracy estimation methods.**

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 also when requirements are still incomplete or still defined at a coarse level of detail, size estimation models have been proposed.

When applying a size estimation method, the method –being applied to incomplete or not thoroughly detailed software specifications– requires less time and effort than the standard measurement process. However, 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 [1].

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 [2][3][4]. 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.

There are different types of FSM and many estimation methods. Here, we concentrate on the COSMIC FSM [5] –one of the most widely used methods– and on model-based COSMIC size estimation [6]. 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. There is no specific reason why the COSMIC FSM –among the many functional size measurement methods– is addressed here. The proposed method can be applied to evaluate the accuracy of estimating functional size expressed in other measurement methods, e.g., IFPUG Function Point Analysis.

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 [6] 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 model-based simplified COSMIC measurement methods. 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 deals with threats to the validity of the study. Section VII accounts for related work. Finally, Section VIII 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 [7][8][9], also when FUR concern real-time applications [10]. 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 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.

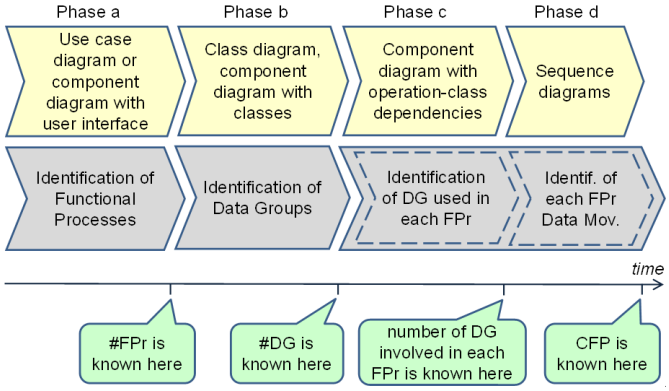

Figure 1. UML modeling process and COSMIC measurement process phases.

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 [6]. 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 (FPrs). 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.

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	$ECFP = 1.588\ \#FPr^{1.00357}\ AvDGperFPr^{1.0312}$

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

In [6], the accuracy of the models given in Table I was evaluated based on the traditional indicators MMRE 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 [6] 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 [11]. 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

Let us first define the problem of evaluating estimates.

We measured the value of interest from n software applications; in our case the value of interest is 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} application, expressed in CFP.

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

In general, there are several estimation methods that can be used to carry out an estimation. Hence, the problem is not just to evaluate the accuracy of a method or model, but to compare a given method or model against other methods and models. In industrial environments, the goal is usually to compare a new method against the method currently used, to evaluate whether it could be convenient to abandon the latter in favor of the former.

A. The Mean Magnitude of Relative Errors

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

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

where $y_i - \hat{y}_i$ is the estimation error (also named the residual).

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 [2], [3], [4]. In practice, MMRE is biased towards prediction systems that under-estimate [11].

B. The Mean Absolute Residual and the Standardized Accuracy

Shepperd and MacDonell [11] proposed that the accuracy of a given estimation method P be measured via the Mean Absolute Residual (MAR):

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

Unlike MMRE, MAR is not biased [11], therefore it is preferable to MMRE.

When we need to compare a new model P with a model P_0, we have MAR_P (the MAR of P) and MAR_{P0} (the MAR of P_0). Based on these 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 P_0, values close to zero indicate that P's accuracy is similar to P_0's accuracy, negative values indicate that P is worse than P_0, hence it should be rejected.

C. Baselines

When a new estimation model P is proposed, we should first establish if it is a "good enough" model, independent of possible alternative models. To this end, we compare the proposed model with a "baseline" model P_0. If there is an estimation model that is in use and is generally believed to be "good enough," the problem is establishing if P is more accurate than such model, which will act as the baseline or reference model P_0. If no reference model is available, the problem is to establish if P satisfies some minimum accuracy conditions. To this end, we use as a baseline some fairly trivial model, that requires little or no knowledge of the phenomena being estimated.

In any case, if the estimates obtained using P are less accurate than the estimates provided by the baseline P_0, we can conclude that P does not yield any improvement, at least as far as accuracy is concerned, and can be rejected, if accuracy is the only acceptance criterion. Of course we may find that P is slightly less accurate than P_0, but is much faster and cheaper than P_0, thus it could be preferable when estimates are needed as soon as possible or the estimation budget is tight.

The random model

When no obvious baseline model exists, Shepperd and MacDonell suggest to use as a referenced model random estimation, based solely on the known (actual) values of previously measured applications. 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; therefore, to compute the MAR of the random model 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 this is not necessary, since the average of the random estimates can be computed exactly [12].

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, or even better ones.

Shepperd and MacDonell observed also that the value of the 5% quantile of the random estimate MARs can be interpreted like α for conventional statistical inference, that is, any accuracy value that is better than this threshold has a less than one in twenty chance of being a random occurrence. Accordingly, MAR_P should be compared with the 5% quantile of the random estimate MARs, rather than with $\overline{MAR_{rnd}}$, to be reasonably sure that P is actually more accurate than rnd.

The constant model

Lavazza and Morasca [13] 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} application is given by the average of the sizes of the other applications, 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.

Lavazza and Morasca [13] found that in real cases MAR_{CM} is quite close to the 5% quantile of random MARs. However, computing MAR_{CM} is much easier and faster than computing the 5% quantile of random MARs, thus ***CM*** can generally be preferred to rnd.

D. Statistical Significance

To establish if the estimates yielded by a method are significantly better than the estimations provided by another method, we need to test the statistical significance of the absolute errors achieved with different estimation methods [2]. To check for statistical significance we used the Wilcoxon Signed Rank Test [14].

The Wilcoxon Signed Rank test can be safely applied also to not normally distributed data, since it makes no assumptions about data distributions.

Via the Wilcoxon test we tested the following Null Hypothesis: "The absolute errors yielded by a model P_i are significantly less than those provided by a model P_j".

E. Size Effect

Suppose that we have two estimation methods P_1 and P_2, and $MAR_{P2} < MAR_{P1}$ (hence, $SA = 1 - \frac{MAR_{P2}}{MAR_{P1}} > 0$). We can conclude that P_2 is more accurate than P_1. Anyway, suppose that we are using P_1 and we are considering the possibility of switching to using P_2, which involves some effort, because P_2 requires some activity or data or programs that P_1 does not require. We would like to know if the improvement that P_2 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 Δ [15] or Hedges's g, which might be preferred when the sample size is small [16]. The effect size –which is scale-free– can be interpreted in terms of the categories proposed by Cohen [17] of small (≈ 0.2), medium (≈ 0.5) and large (≈ 0.8).

F. Estimate Comparison Based on Individual Absolute Residuals

Given a dataset and two models P_1 and P_2, following Shepperd and MacDonell we state that P_1 is more accurate than P_2 if $SA_{P1} > SA_{P2}$, i.e., if $MAR_{P1} < MAR_{P2}$. However, this is not the only criterion that can be used to compare the performances of P_1 and P_2. Let $\langle \hat{y}_{1,P1}, .., \hat{y}_{n,P1} \rangle$ and $\langle \hat{y}_{1,P2}, .., \hat{y}_{n,P2} \rangle$ be the estimates provided by the two models. We may say that P_1 is more accurate than P_2 if and only if there are at least $\lceil \frac{n+1}{2} \rceil$ distinct values of i such that $|y_i - \hat{y}_{i,P1}| < |y_i - \hat{y}_{i,P2}|$, that is, if P_1 provides smaller absolute residuals than P_2 in the majority of cases.

The probability that P_1 provides better results than P_2 in the majority of estimates is measured by the IARA (Individual Absolute Residual Assessment) indicator [13]. The IARA index is defined as the number of estimates for which the residuals of P_1 are less than those of P_2 divided by the total number of estimates. The statistical significance of IARA can be tested via the binomial test.

IV. Experimental Evaluation

The five size estimation models given in Table I were applied to the applications in the dataset that was used to derive the models [6]. As baseline models we also estimated the size of the applications in the dataset using the constant and random models.

While carrying out the analysis, we realized that model *ECFP = 7.3 #FPr* is similar to the Average Functional Process (AFP) estimation method proposed by COSMIC [18]. In fact, the AFP estimation model is *ECFP = MSFP×#Fpr*, where *MSFP* is the mean size of functional processes. Therefore, we computed *MSFP* and applied the AFP method as well.

A. The dataset

The dataset we used to evaluate the accuracy of the considered models included data from 23 projects of different nature. More specifically, we used data from 5 projects proposed by the COSMIC consortium to illustrate the counting process, 7 projects from academia, 10 Web-based Management Information Systems (from the same company) and a project management tool. Additional information on the dataset (including the actual data) can be found in [6].

Some descriptive statistics of the dataset are given in Table II.

TABLE II. Descriptive statistics of the dataset

	Size[CFP]	#FPr	AvDGperFPr
Mean	174.1	24.3	4.1
Median	116.0	19.0	3.8
Min	31.0	7.0	2.1
Max	514.0	74.0	8.2
StDev	139.8	16.8	1.5

B. Analysis of errors: Mean Absolute Error

The MARs of the estimates obtained using the models mentioned above are given in Table III, together with the MARs of the constant model (MAR_{CM}) and the random model ($\overline{MAR_{rnd}}$).

In Table III, column rnd 5% gives the value of the 5% quantile of the random estimate MARs. This practice is suggested by Shepperd and MacDonell [11] because the 5% quantile can be interpreted like α for conventional statistical inference, that is, any accuracy value that is better than this threshold has a less than one in twenty chance of being a random occurrence. Therefore, to have reasonable confidence that a given model is actually predicting and not guessing, we expect a value of MAR that is lower than this threshold value. We have that both AFP and the UML-based estimation methods yield MAR values that are well below the proposed threshold, hence we can regard them as not due to chance.

TABLE III. Mean absolute residuals of models.

Name	*Formula*	*MAR*
rnd	–	146
rnd 5%	–	106
CM	–	114
AFP	$ECFP = MSFP\ \#FPr$	52
avg1	$ECFP = 7.3\ \#FPr$	54
reg1	$ECFP = -16.5 + 6.698\ \#FPr$	48
avg2	$ECFP = AvDGperFPr1.8\ \#FPr$	27
reg2	$ECFP = -64.6 + 7.63\ \#FPr + 9.71\ AvDGperFPr$	40
log2	$ECFP = 1.588\ \#FPr^{1.00357}\ AvDGperFPr^{1.0312}$	25

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 8 methods there are 28 possible comparison among methods, hence 28 values of SA. Listing all these SA values could create confusion, while to compare two methods' accuracies, we just need to compare their MARs: the model featuring the smaller MAR is likely the best.

C. Distribution of Estimation Errors

In the following sections we shall see that MAR is a quite synthetic indicator, which "hides" important information. To get a deeper insight into estimation errors, in this section

the distribution of errors (alias residuals), absolute errors and relative absolute errors are given.

Figure 2 illustrates the boxplots of the errors yielded by each one of the evaluated models. The blue diamonds indicate the mean error for each model. It is interesting to note that AFP and avg1 tend to overestimate, reg1 tends to underestimate, while all the other model neither overestimate nor underestimate.

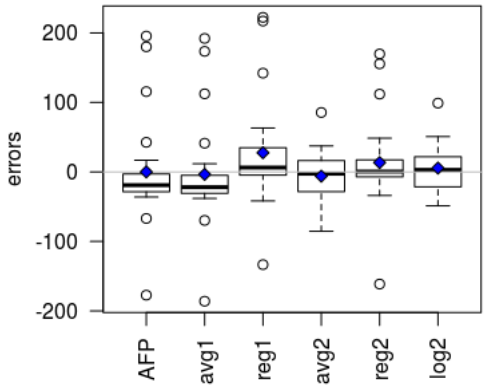

Figure 2. Estimation residuals of the evaluated models.

Figure 3 illustrates the boxplots of the absoluter errors yielded by each one of the evaluated models. The blue diamonds indicate the MARs. Also from this figure we get some interesting result. For instance, the MAR of reg2 is larger than the MARs of avg2 and log2, but the distributions of absolute errors are very similar: the distribution of reg2 is even better than the other two, except for three outliers, which feature absolute errors larger than 150 CFP.

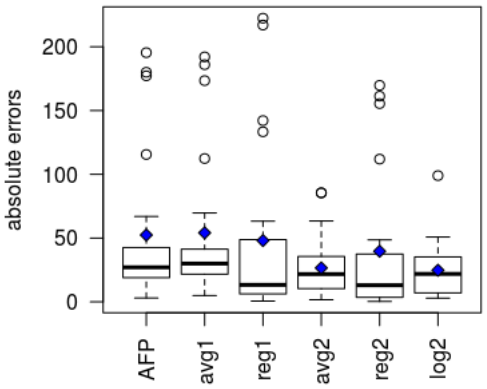

Figure 3. Absolute estimation errors of the evaluated models.

Figure 3 illustrates the boxplots of the relative absolute errors yielded by each one of the evaluated models. The blue diamonds indicate the MMREs.

Relative absolute errors are useful to assess the importance of errors. As a matter of facts, the boxplots show that some models yield some errors that are close to 100%.

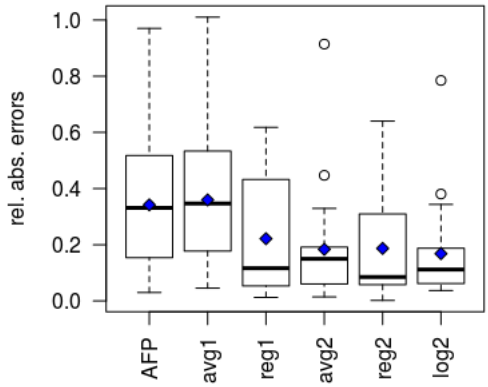

Figure 4. Relative absolute estimation errors of the evaluated models.

D. Analysis of errors: Wilcoxon Signed Rank Test

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

Table III also confirms that the constant model is more accurate than the random model, as demonstrated by Lavazza and Morasca [13]. 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 [2]. Namely, we compared the absolute residuals provided by every pair of methods via Wilcoxon Sign Rank Test.

The results are given in Table IV, 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 IV. COMPARISON OF ABSOLUTE RESIDUALS USING WILCOXON SIGN RANK TEST.

	const	AFP	avg1	reg1	avg2	reg2	log2
const		>	>	>	>	>	>
AFP	<		<	>	>	>	>
avg1	<	>		>	>	>	>
reg1	<	<	<		=	>	>
avg2	<	<	<	=		=	=
reg2	<	<	<	<	=		=
log2	<	<	<	<	=	=	

The results provided by Wilcoxon Sign Rank Test confirm the results given in Section IV-B in that the constant model is less accurate than all other models and AFP is more accurate than avg1 but less accurate than model-based size estimation methods. However, Wilcoxon Sign Rank Test provides further insights with respect to MAR:

- There is no sufficient evidence to conclude that log2 is more accurate than avg2: this fact could be guessed, based on the fact that MAR_{log2} (25) and MAR_{avg2} (27) are quite close.

- Somewhat surprisingly, there is no evidence that either avg2 (which has $MAR_{avg2} = 27$) or log2 (which has $MAR_{log2} = 25$) is more accurate than reg2 (which has $MAR_{reg2} = 40$).
- Similarly, There is no evidence that avg2 (which has $MAR_{reg2} = 27$) is more accurate than reg1 (which has $MAR_{reg1} = 48$).

These results are interesting, in that by just looking at the MAR values we could have concluded that some model (e.g., avg2) is more accurate than another model (e.g., 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 5, where the boxplots of the absolute residuals of models avg2 and reg2 are given: it is easy to see that reg2 has a greater MAR because of three applications, whose size estimation error is quite large. Apart from these three applications, the distributions are similar: accordingly, the MARs of avg2 and reg2 are not significantly different.

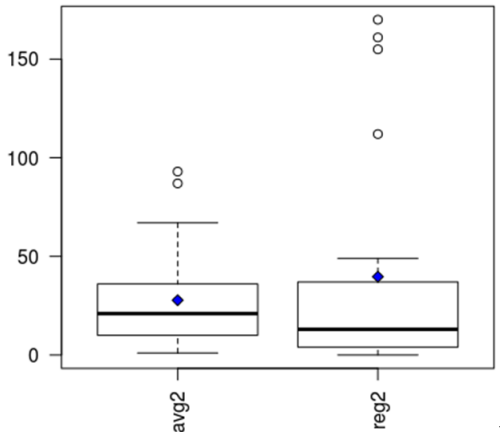

Figure 5. Absolute residuals of models avg2 and reg2.

E. Analysis of errors: Effect Size

Now, as recommended by Shepperd and MacDonell (see Section III-E) we evaluate the effect size. To this end, we computed Hedges's g for all model pairs. The results are given in Table V. where the value for the i^{th} row and j^{th} column is Hedges's g for the pair of models indicated on the i^{th} row and j^{th} column. For instance, the value on row avg1 and column reg1 is 0.09: this indicates that using reg1 instead of avg1 involves a negligible effect.

TABLE V. Effect size (Hedges's g).

	CM	AFP	avg1	reg1	avg2	reg2	log2
CM	–	0.80	0.79	0.82	1.32	0.98	1.36
AFP	-0.80	–	-0.03	0.07	0.55	0.21	0.60
avg1	-0.79	0.03	–	0.09	0.60	0.24	0.64
reg1	-0.82	-0.07	-0.09	–	0.40	0.13	0.44
avg2	-1.32	-0.55	-0.60	-0.40	–	-0.29	0.07
reg2	-0.98	-0.21	-0.24	-0.13	0.29	–	0.34
log2	-1.36	-0.60	-0.64	-0.44	-0.07	-0.34	–

Table V essentially confirms the findings given in the previous sections. It is easy to see that all model-based size estimation methods appear definitely preferable with respect to the constant model. AFP and avg1 appear essentially equivalent. All models involving two independent variables appear better than those based on one independent variable. Models avg2 and log2 appear preferable to the other model-based estimation methods, with log2 only marginally better than avg2.

F. Analysis of errors: IARA

The accuracy of the size estimation models was evaluated via the IARA index described in Section III-F. Specifically, the IARA index was computed for every pair of models. The results are given in Table VI. For instance, the value on row AFP and column reg1 indicates the ratio between the number of estimates for which reg1 was more accurate than AFP and the total number of estimates: 0.62 indicates that reg1 was more accurate than AFP in 62% of the estimates. Although this is a quite straightforward indication, the F in parenthesis indicates that the binomial test fails (i.e., its p-value is not < 0.05), hence the indication is not reliable.

TABLE VI. Comparison of models via IARA indexes

	AFP	avg1	reg1	avg2	reg2	log2
AFP		0.24(F)	0.62(F)	0.67(F)	0.81(S)	0.71(S)
avg1	0.76(S)		0.62(F)	0.71(S)	0.81(S)	0.76(S)
reg1	0.38(F)	0.38(F)		0.52(F)	0.67(F)	0.48(F)
avg2	0.33(F)	0.29(F)	0.48(F)		0.52(F)	0.48(F)
reg2	0.19(F)	0.19(F)	0.33(F)	0.48(F)		0.48(F)
log2	0.29(F)	0.24(F)	0.52(F)	0.52(F)	0.52(F)	

It can be observed that few of the indexes in Table VI are marked "S" for success (i.e., p-value < 0.05). This is largely due to the fact that the dataset used for the evaluation is relatively small. However, it is interesting to note that the IARA index provides indications that are consistent with the Wilcoxon test. In fact, the Wilcoxon signed rank test found that reg2 is not significantly worse than avg2 and log2, although its MAR (40) is larger that avg2 and log2 MARs (27 and 25, respectively). The IARA index tells us that both avg2 and log2 provide more accurate estimates than reg2 only in 52% of cases.

The IARA index appears to add a new dimension to the analysis, complementing the indications (often quite rough) given by the MAR.

V. Discussion of Results

The results of the empirical investigation described in Section IV support consideration concerning a few aspects: the accuracy of model-based methods for estimating the size of software applications expressed in CFP; the practical application of the proposed evaluation methods; the evaluations that can be obtained by directly observing the results yielded by estimation method, without performing statistical analyses.

A. Evaluation of estimation models

With reference to Figure 1, at the end of phase a), we know the number of Functional Processes (#FPr), thus models AFP, 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, and we can exploit this information to achieve more accurate size estimates. However, having reliable empirical evidence that progressing through 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 effect size indicators (see Table V) suggest that the models available at the end of phase b) allow only for a medium-small improvement over the best models 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 data groups (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.

B. On the practical application of the proposed indicators

The practical application of the proposed method depends largely on how easily practitioners can derive the indicators mentioned in the sections above. Luckily, the proposed technique are supported by open-source tools, and can be automated quite easily. Specifically, the language and programming environment R [19] supports the computation of all of the proposed indicators; you just need to install the 'effsize' package [20] to compute Hedges's g.

So, once the actual and estimated values are available, computing the indicators described in Section III (as well as plots given in Section V-C below) can be completely automatic, and quite fast.

In conclusion, the proposed indicators are easily obtained and are also fairly easy to interpret; therefore, practitioners should have no difficulty to assess the accuracy of estimation methods as proposed.

C. Direct observation of residuals

The proposed approach is statistically sound and appears to provide reliable indications. Moreover, as discussed in Section V-B, it can be automated. Nonetheless, one should not forget that in general the direct observation of the estimates and their comparison with actuals can provide quite enlightening indications.

For instance, suppose that you are interested in evaluating the accuracy of the avg2 and log2 models. To this end, you could look at a plot like the one in Figure 6. In Figure 6, the x axis reports the actual size of applications (in CFP), and the y axis reports the estimated size (in CFP); applications appear as small circles, estimates obtained using avg2 appear as green crosses, and estimates obtained via log2 appear as red xs.

By looking at Figure 6, appreciating the accuracy of the estimation models is easy. It is also easy to compare the estimates yielded by avg2 and log2. For instance, it can be noticed that log2 estimates are always smaller than avg2 estimates: this observation may be very relevant for a practitioner who has to choose whether to use avg2 or log2.

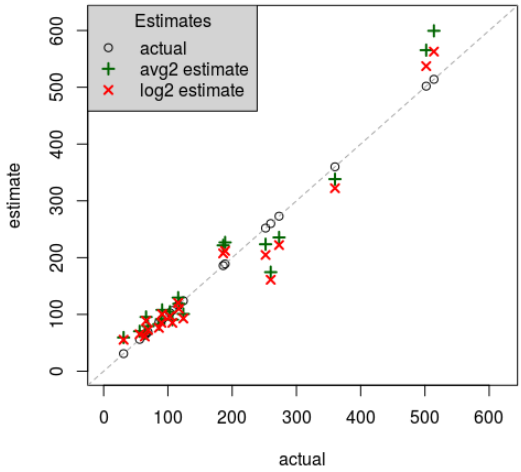

Figure 6. avg2 and log2 estimates vs. actuals.

As is often the case, there are many relatively small applications. In our case, it is difficult to assess the estimates of applications in the [50,150] CFP range. To overcome this difficulty, you can draw another plot that accounts for the applications smaller than 150 CFP only. Such plot is shown in Figure 7: the plot confirms that log2 estimates are always smaller than avg2 estimates also for relatively small applications.

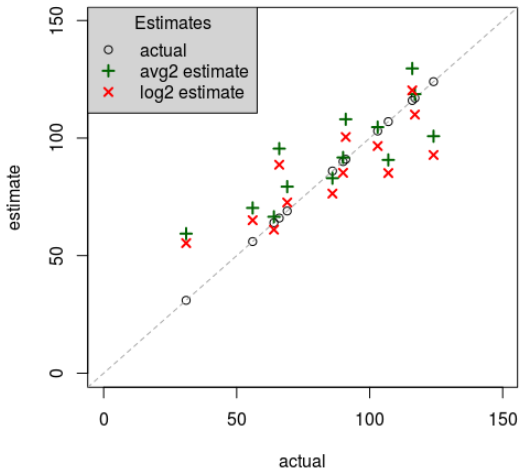

Figure 7. avg2 and log2 estimates vs. actuals (detail on applications smaller than 150 CFP).

VI. Threats to Validity

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 applications. It is possible that other datasets could support somewhat different conclusions. However, the fact that our dataset includes several industrial applications, 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.

VII. 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 [21]. 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: $log\frac{prediction}{actual}$ [22]. 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 [23]. 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 [24].

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 [25], where the NESMA estimated method is proposed and its accuracy is evaluated on the training set. A noticeable exception is [26], where several early estimation methods for Function Point measures are evaluated via an empirical study.

The method used in this paper was partly applied to evaluate the accuracy of measure conversions (from IFPUG Function Points to COSMIC Function Points) in [27].

VIII. 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 observations.

First, practitioners are very keen to know the accuracy that can be achieved via size estimation methods. In fact, they often use size estimation methods to derive the most important piece of information upon which the cost of software is estimated; therefore, accurate size estimation is essential to get accurate cost estimates, hence to allocate the correct amount of resources and prepare reliable development plans.

Second, to evaluate the accuracy of estimates, you need reliable indicators. Traditional indicators like MMRE have been proved to be biased: thus, finding and testing more reliable indicators is necessary. Consider for instance what happens when researchers propose a new estimation technique: how can we decide that the new technique is good, and possibly even better than existing techniques? Reliable accuracy evaluation techniques and indicators –like those proposed here– can answer such question.

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.
- Any estimation method must prove more accurate than the trivial models –like the constant model and the random model– that do not require any knowledge concerning the application to be estimated. Hence, one should always test models against the random and constant models.
- Having established that the considered estimation method is better than the trivial methods, one should also evaluate whether the considered method is more accurate than the currently used estimation technique, to see if abandoning the latter to adopt the new method is worthwhile.
- Using Wilcoxon Sign Rank Test is advisable, since it can give indications concerning the statistically significance of the comparisons based on two methods' MAR values.
- 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).
- 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.
- Finally, the IARA index shows in how many cases a method is more accurate than another one.

Overall, this paper shows that assessing the accuracy of estimates can hardly be based on a single indicator. Instead, using the proposed set of indicators provides a quite detailed and complete picture of the merits of the evaluated estimation models.

As a final observation, we note that the analyses reported in this paper were carried out in the R environment [19]. Practitioner and researchers that need to evaluate estimation accuracy can apply the proposed approach quite easily and derive the indicators described in Section III very quickly.

Future work includes:

- Further evaluating model-based COSMIC size estimation methods 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] L. Lavazza, "Accuracy Evaluation of Model-based COSMIC Functional Size Estimation," in International Conference on Software Engineering Advances (ICSEA), 2017.

[2] 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.

[3] 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.

[4] 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.

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

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

[14] J. Cohen, Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Lawrence Earlbaum Associates, 1988.

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

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

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

[18] F. Vogelezang, C. Symons, A. Lesterhuis, R. Meli, and M. Daneva, "Guideline for Early or Rapid COSMIC Functional Size Measurement by using approximation approaches," July 2015.

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

[20] M. Torchiano, "Efficient Effect Size Computation – package 'effsize'," March 2017. [Online]. Available: https://cran.r-project.org/web/packages/effsize/effsize.pdf

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] A. Z. Abualkishik and L. Lavazza, "IFPUG Function Points to COSMIC Function Points convertibility: A fine-grained statistical approach," Information and Software Technology, vol. 97, 2018, pp. 179–191.