Adding Confidence Intervals to the NESMA Functional Size Estimation Method

Luigi Lavazza
Università degli Studi dell'Insubria
Varese, Italy
email:luigi.lavazza@uninsubria.it

Angela Locoro
Università degli Studi di Brescia
Brescia, Italy
email:angela.locoro@unibs.it

Geng Liu
Hangzhou Dianzi University
Hangzhou, China
email:liugeng@hdu.edu.cn

Roberto Meli
DPO
Rome, Italy
email:roberto.meli@dpo.it

***Abstract*—In many projects, software functional size is measured via the IFPUG (International Function Point Users Group) Function Point Analysis method. However, applying Function Point Analysis using the IFPUG process is possible only when functional user requirements are known completely and in detail. To solve this problem, several early estimation methods have been proposed and have become *de facto* standard processes. Among these, a prominent one is the 'NESMA (NEtherlands Software Metrics Association) estimated' (also known as High-level Function Point Analysis) method. The NESMA estimated method simplifies the measurement by assigning fixed weights to Base Functional Components, instead of determining the weights via the detailed analysis of data and transactions. This makes the process faster and cheaper, and applicable when some details concerning data and transactions are not yet known. The accuracy of the mentioned method has been evaluated, also via large-scale empirical studies, showing that the yielded approximate measures are sufficiently accurate for practical usage. However, a limitation of the method is that it provides a specific size estimate, while other methods can provide confidence intervals, i.e., they indicate with a given confidence level that the size to be estimated is in a range. In this paper, we aim to enhance the NESMA estimated method with the possibility of computing a confidence interval. To this end, we carry out an empirical study, using data from real-life projects. The proposed approach appears effective. We expect that the possibility of estimating that the size of an application is in a range will help project managers deal with the risks connected with inevitable estimation errors.**

***Index Terms*—Function Point Analysis; Early Size Estimation; High-Level FPA; NESMA estimated; Confidence intervals.**

I. Introduction

This paper illustrates the extension of an initial study on the enhancement of the NESMA method with the computation of confidence interval [1].

In the late seventies, Allan Albrecht introduced Function Points Analysis (FPA) at IBM [2], as a means to measure the functional size of software, with special reference to the "functional content" delivered by software providers. Albrecht aimed at defining a measure that might be correlated to the value of software from the perspective of a user, and could also be useful to assess the cost of developing software applications, based on functional user requirements.

FPA is a functional size measurement method, compliant with the ISO/IEC 14143 standard, for measuring the size of a software application in the early stages of a project, generally before actual development starts. Accordingly, software size measures expressed in Function Points (FP) are often used for cost estimation.

The International Function Points User Group (IFPUG) is an association that keeps FPA up to date, publishes the official FP counting manual [3], and certifies professional FP counters. Unfortunately, in some conditions, performing the standard IFPUG measurement process may be too long with respect to management needs, because standard FP measurement can be performed only when relatively complete and detailed requirements specifications are available, while functional measures could be needed much earlier for management purposes.

To tackle this problem, the IFPUG proposes Simple Function Points (SFP). This is an alternative way of measuring the functional size of software: while the SFP method is based on the same concepts as FPA, it requires less detailed information than FPA, so that it is applicable before complete and detailed requirements specifications are available; besides, if is faster and cheaper to apply. As such, it is often presented as a lightweight functional measurement method, also suitable for agile processes. Although the SFP method provides measures that are quantitatively similar to those yielded by FPA, it is not an approximation method for FPA; instead, it is a different measurement method that yields different measures.

Before SFP was proposed, many methods were invented and used to provide *estimates* of functional size measures, based on fewer or coarser-grained information than required by standard FPA. These methods are applied very early in software projects, even before deciding what process (e.g., agile or waterfall) will be used. Among these methods, one of the most widely used is the "NESMA estimated" method [4], which was developed by NESMA [5]. Using this method for size estimation was then suggested by IFPUG [6], which renamed the method High-Level FPA (HLFPA).

The NESMA estimated method has been evaluated by several studies, which found that the method is usable in practice to approximate traditional FPA values, since it yields reasonably accurate estimates, although it has been observed that the NESMA method tends to underestimate size, which is potentially dangerous.

Many estimation methods provide a "confidence interval", meaning that instead of providing a single value, they compute

an interval in which the actual size is expected—with a given confidence level—to be. The greater the required confidence, the greater the interval. Knowing the confidence interval is considered very useful by project managers, because it helps managing the risk deriving from inevitable estimation errors and the inherent uncertainty of estimates. Unfortunately, the NESMA estimated method does not provide a confidence interval. Recently, a proposal for enhancing the NESMA estimated method with a mechanism confidence intervals has been published [1]. Specifically, that paper provided two main contributions: the correction of the NESMA method to eliminate underestimation, and the introduction of confidence intervals. The original study was based on a single dataset, and reported the hypothesis that different datasets may require different corrections and support different confidence intervals.

This paper extends the initial study [1] by verifying that a different dataset actually requires a different correction of the NESMA method and provides different confidence ranges. Therefore, the proposed numerical method is an instrument that lets software project managers get corrected size estimates, equipped with confidence intervals, which apply to specific data, in a context-aware manner.

In addition, this paper illustrates how to take advantage of confidence intervals in real-life situations, by introducing an example of how the proposed technique can be used in practice for effort estimation.

The remainder of the paper is organized as follows. Section II provides an overview of FPA and the NESMA method. Section III describes the empirical study and its results, which are discussed in Section IV. Section V illustrates the usage of the proposed techniques in practical project management, namely, for effort estimation. In Section VI, we discuss the threats to the validity of the study. Section VII reports about related work. Finally, in Section VIII, we draw some conclusions and outline future work.

II. Background

Function Point Analysis was originally introduced by Albrecht to measure the size of data-processing systems from the point of view of end-users, with the goal of the estimating the value of an application and the development effort [2]. The critical fortunes of this measure led to the creation of the IFPUG (International Function Points User Group), which maintains the method and certifies professional measurers.

The "amount of functionality" released to the user can be evaluated by taking into account 1) the data used by the application to provide the required functions, and 2) the transactions (i.e., operations that involve data crossing the boundaries of the application) through which the functionality is delivered to the user. Both data and transactions are counted on the basis of Functional User Requirements (FURs) specifications, and constitute the IFPUG Function Points measure.

FURs are modeled as a set of Base Functional Components (BFCs), which are the measurable elements of FURs: each of the identified BFCs is measured, and the size of the application is obtained as the sum of the sizes of BFCs. IFPUG BFCs are: data functions (also known as logical files), which are classified into Internal Logical Files (ILF) and External Interface Files (EIF); and Elementary Processes (EP)—also known as transaction functions—which are classified into External Inputs (EI), External Outputs (EO), and External inQuiries (EQ), according to the activities carried out within the considered process and the primary intent.

The complexity of a data function (ILF or EIF) depends on the RETs (Record Element Types), which indicate how many types of variations (e.g., sub-classes, in object-oriented terms) exist per logical data file, and DETs (Data Element Types), which indicate how many types of elementary information (e.g., attributes, in object-oriented terms) are contained in the given logical data file.

The complexity of a transaction depends on the number of FTRs—i.e., the number of File Types Referenced while performing the required operation—and the number of DETs—i.e., the number of types of elementary data—that the considered transaction sends and receives across the boundaries of the application. Details concerning the determination of complexity can be found in the official documentation [3].

The core of FPA involves three main activities:

1) Identifying data and transaction functions.
2) Classifying data functions as ILF or EIF and transactions as EI, EO or EQ.
3) Determining the complexity of each data or transaction function.

The first two of these activities can be carried out even if the FURs have not yet been fully detailed. On the contrary, activity 3 requires that all details are available, so that FP measurers can determine the number of RET or FTR and DET involved in every function. Activity 3 is relatively time- and effort-consuming [7].

Note that IFPUG defines both unadjusted FP (UFP) and adjusted FP. The former are a measure of functional requirements. The latter are obtained by correcting unadjusted FP to obtain an indicator that is better correlated to development effort. Noticeably, the ISO standardized only unadjusted FP, recognizing UFP as a proper measure of functional requirements [8]. Following the ISO, in this paper we deal only with UFP, even when we speak generically of Function Points or FP.

The NESMA estimated method does not require activity 3, thus allowing for size estimation when FURs are not fully detailed: it only requires that the complete sets of data and transaction functions are identified and classified.

The SFP method [9] does not require activities 2 and 3: it only requires that the complete sets of data and transaction functions are identified.

Both the NESMA estimated method and SFP methods let measurers skip the most time- and effort-consuming activity, thus both are relatively fast and cheap. The SFP method does not even require classification, making size estimation even faster and less subjective (since different measurers can sometimes classify differently the same transaction, based on the subjective perception of the transaction's primary intent).

NESMA defined two size estimation methods: the 'NESMA Indicative' and the 'NESMA Estimated' methods. IFPUG acknowledged these methods as early function point analysis methods, under the names of 'Indicative FPA' and 'High-Level FPA,' respectively [6]. The NESMA Indicative method proved definitely less accurate [10], [11]. Hence, in this paper, we consider only the NESMA Estimated method.

The NESMA Estimated method requires the identification and classification of all data and transaction functions, but does not require the assessment of the complexity of functions: ILF and EIF are assumed to be of low complexity, while EI, EQ and EO are assumed to be of average complexity. Hence, estimated size is computed as follows:

$$EstSize_{UFP} = 7\ \#ILF + 5\ \#EIF + 4\ \#EI + 5\ \#EO + 4\ \#EQ$$

where *#ILF* is the number of data functions of type ILF, *#EI* is the number of transaction functions of type EI, etc.

III. Empirical study

In this section, the empirical study is described: Section III-A described the datasets used for the reported analysis; Section III-B illustrates some considerations concerning the accuracy of the NESMA method that affect the study, and introduces the correction of the NESMA method, to avoid size underestimation; Section III-C describes how the study was performed; Section III-D describes the obtained results. While the aforementioned sections use the same dataset used previously [1], the following Sections III-E and III-F use a second dataset, to replicate the previous study.

A. The datasets

In the empirical study, we used two datasets. One is the ISBSG dataset [12], which has been extensively used for studies concerning functional size [13]–[18].

The ISBSG dataset contains many data concerning software development projects. Of the many available data, we considered only the project size, expressed in UFP, and the components used to compute the size, i.e., *#ILF*, *#EIF*, *#EI*, *#EO* and *#EQ*.

The ISBSG dataset contains several small project data. As a matter of fact, estimating the size of small projects is not very interesting. Based on these considerations, we removed from the dataset the projects smaller than 100 UFP (Unadjusted Function Points). The resulting dataset includes data from 140 projects having size in the [103, 4202] range. Some descriptive statistics for this dataset are given in Table I.

TABLE I
Descriptive statistics for the ISBSG dataset (after removing small projects).

	UFP	*#ILF*	*#EIF*	*#EI*	*#EO*	*#EQ*	NESMA
Mean	801	22	20	35	37	37	730
Std	818	21	22	37	65	48	721
Median	475.5	14	14.5	22	10	20.5	463
Min	103	0	0	0	0	0	71
Max	4202	100	172	204	442	366	3755

The second dataset was provided by a Chinese company (whose identity we cannot disclose) that is active in the banking and finance domain. Although not popular as the ISBSG dataset, also the Chinese dataset was formerly used in a few studies concerning functional size measurement [19], [20].

Also with this dataset (which is called the "Chinese" dataset throughout the paper) we removed the data concerning projects smaller than 100 UFP. As a result, we obtained a dataset containing 424 project data: some descriptive statistics for the dataset are given in Table II. It can be noticed that the Chinese dataset includes data form much larger projects than the ISBSG dataset.

TABLE II
Descriptive statistics for the Chinese dataset (after removing small projects).

	UFP	*#ILF*	*#EIF*	*#EI*	*#EO*	*#EQ*	NESMA
Mean	3819	90	47	303	122	246	3670
Std	5877	180	129	516	319	470	5706
Median	1447	31	7	116	29	77	1484
Min	103	0	0	0	0	0	99
Max	35910	2169	1198	3551	4517	4231	37571

B. The accuracy of the NESMA estimated method when applied to the ISBSG dataset

As already observed in previous papers [18], [21], the NESMA estimated method tends to underestimate. Figure 1 shows that more than 75% of the NESMA estimates of ISBSG project size have positive error. Being the error defined as the actual size (i.e., the size measured via the ISBSG standard FPA process) minus the estimate, positive error indicate underestimation.

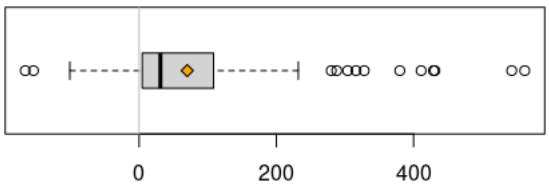

Fig. 1. Boxplot of estimation errors by the NESMA method, when applied to the ISBSG dataset.

In addition, Figure 1 suggests that the distribution of NESMA errors is skewed. The skewedness of NESMA errors is clearly visible in Figure 2, which illustrates the distribution of errors: it is easy to notice that most errors are positive.

For our purposes, the fact that the distribution of NESMA errors is skewed and not centered on zero means that we cannot evaluate confidence errors as is usually done. Specifically, given a confidence level C, we cannot select two error levels e_L and e_H that are symmetric with respect to the mean error $\bar{e}$ (i.e., $|e_H - \bar{e}| = |\bar{e} - e_L|$) such that the proportion of errors such that $e_H \geq error \geq e_L$ is C.

Since it makes hardly sense to provide confidence intervals for a method that underestimates systematically, we first

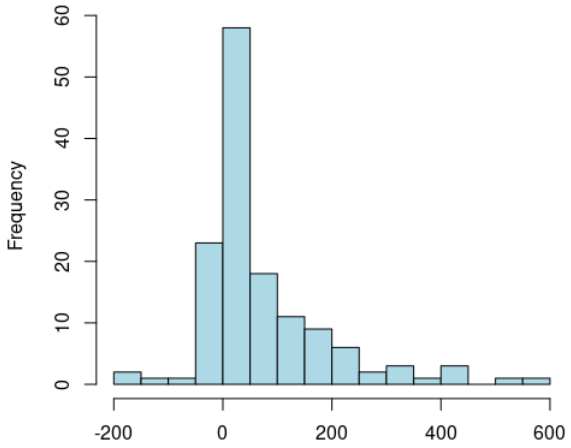

Fig. 2. Histogram of estimation errors by the NESMA method, when applied to the ISBSG dataset.

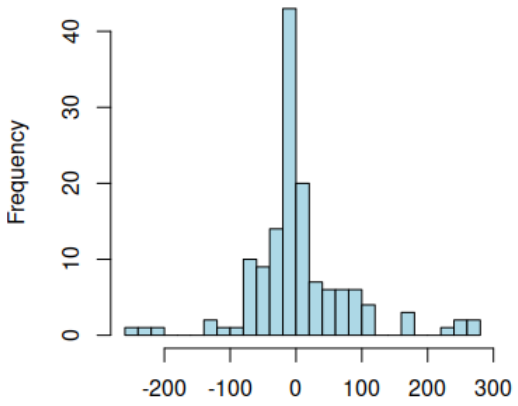

Fig. 4. Boxplot of estimation errors by the corrected NESMA method, when applied to the ISBSG dataset.

"correct" the NESMA estimated method. The mean actual size is 801 UFP, while the mean size estimated via the NESMA method is 730 UFP. The ratio between these two means is approximately 1.09. Accordingly, we need to correct NESMA estimates, multiplying them by 1.09, to make the means of the two distributions equal (in [1] the correction factor was 1.08; subsequent more accurate evaluations led to set the correction factor to 1.09). In this way, we obtain estimates that have a better error distribution (less skewed and centered around zero) and a smaller mean absolute error (49.7 UFP instead of 83.8 UFP).

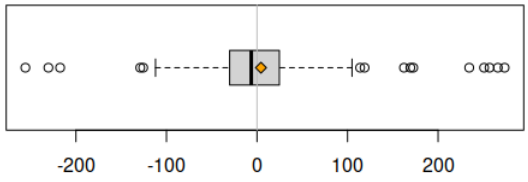

Fig. 3. Boxplot of estimation errors by the corrected NESMA method, when applied to the ISBSG dataset.

The boxplot of estimation errors obtained with the corrected NESMA method is shown in Figure 3: it can be noticed that the mean error is just above zero, while the median error is just below zero.

The error distribution is shown in Figure 4: it can be noticed that the distribution is much less skewed than in Figure 2.

Since the practical objective of this work is to provide project managers with reliable predictions of functional size, in what follows we consider only estimates provided by the original NESMA method and corrected as described above. In other words, we consider the following estimates:

$$EstSize_{UFP} = 1.09\ (7\ \#ILF + 5\ \#EIF + 4\ \#EI + 5\ \#EO + 4\ \#EQ)$$

We make reference to this estimation as the "Corrected NESMA" method.

C. Method used

In essence, given a confidence level C we aim at finding two values k_L and k_H such that a proportion C of the actual size measures (i.e., measures obtained via the official IFPUG FPA process) is in the range $[k_L \cdot EstSize_{UFP}, k_H \cdot EstSize_{UFP}]$, where $EstSize_{UFP}$ is the size estimates computed via the Corrected NESMA method.

Finding k_L and k_H would be straightforward if the estimation errors obtained via the Corrected NESMA method were normally distributed. Instead, it is not so, as shown by the Shapiro-Wilk test.

Therefore, we proceeded as follows:

1) We computed the ratio $\frac{ActualSize}{EstSize_{UFP}}$ for all projects in the dataset, obtaining a set of ratios; this set was then sorted and stored in vector *vRatios*.
2) We computed the quantiles from 0 to 1, with 0.01 steps, of *vRatios*, obtaining an ordered vector *vQuant*.
3) We looked for two indexes i_L and i_H in *vQuant* such that $i_H - i_L + 1 = C \cdot n$, where n is the number of projects in the dataset.
4) k_L and k_H are the values in *vRatios* having index i_L and i_H, respectively, i.e., $vRatios[i_L]$ and $vRatios[i_H]$.

In this way, we obtain a size estimate interval that contains a proportion C of all estimates, such that all estimation errors outside the interval are greater than those within the interval.

D. Results obtained for the ISBSG dataset

We applied the procedure described in Section III-C for various confidence levels. The results obtained are given in Table III. Note that these results depend on the dataset being used, in our case, the ISBSG dataset. In other contexts, a given confidence level could correspond to different confidence intervals. For instance, in the ISBSG dataset, the minimum and maximum ratios $\frac{ActualSize}{EstSize_{UFP}}$ are 0.758 and 1.343, respectively; in another dataset, a smaller minimum and a larger maximum ratios are clearly possible, as shown in Section III-F.

TABLE III
CONFIDENCE INTERVALS FOR VARIOUS CONFIDENCE LEVELS, FOR THE ISBSG DATASET.

conf. level	k_L	k_H
0.10	0.991	1.011
0.20	0.980	1.019
0.30	0.968	1.030
0.40	0.954	1.043
0.50	0.943	1.057
0.60	0.929	1.077
0.70	0.910	1.095
0.80	0.872	1.137
0.90	0.843	1.208
0.95	0.818	1.208
1.00	0.751	1.331

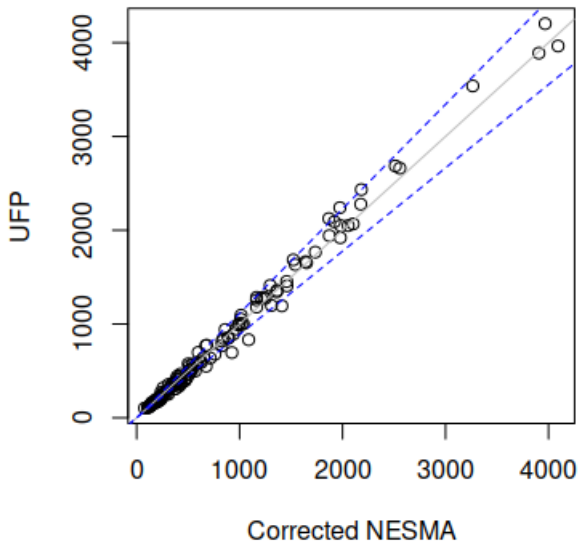

Fig. 5. Corrected NESMA estimates vs. actual size in UFP, with confidence $C = 0.75$, for the ISBSG dataset.

For illustration purposes, Figure 5 plots the ISBSG project data in the plan defined by actual size (the y axis) and the size estimated via the Corrected NESMA method (the x axis). In the plot, the dashed blue lines represent the $y = k_L\ x$ and $y = k_H\ x$ lines.

E. The accuracy of the NESMA estimated method when applied to the Chinese dataset

As observed in Section III-B, the NESMA estimated method tends to underestimate. Figure 6 shows that the majority of the NESMA estimates of the Chinese dataset project size have positive error (positive errors indicate underestimation).

As for the ISBSG dataset, the distribution of NESMA errors is skewed (although less evidently than for the ISBSG dataset), as shown in Figure 7. It is easy to notice that most errors are positive.

Therefore, we corrected NESMA estimation, as we did for the ISBSG dataset. As discussed above, we cannot just apply the same multiplier found for the ISBSG dataset. For the Chinese dataset, the mean actual size is 3819 UFP, while the mean size estimated via the NESMA method is 3670 UFP.

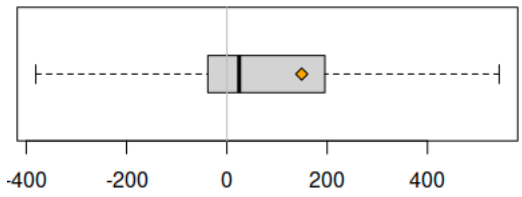

Fig. 6. Boxplot of estimation errors by the NESMA method, when applied to the Chinese dataset (outliers not shown).

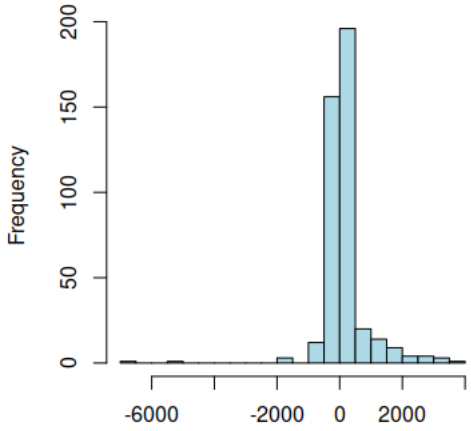

Fig. 7. Histogram of estimation errors by the NESMA method, when applied to the Chinese dataset.

The ratio between these two means is approximately 1.04. Accordingly, we correct NESMA estimates, multiplying them by 1.04, to make the mean value of the estimates sizes equal to the mean value of the actual sizes. In this way, we obtain estimates that have a better error distribution (less skewed and centered around zero) and a smaller mean absolute error (319 UFP instead of 340 UFP).

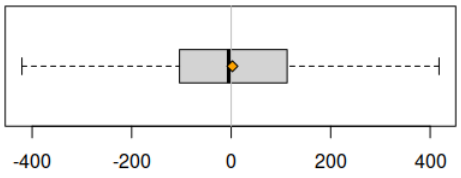

Fig. 8. Boxplot of estimation errors by the corrected NESMA method, when applied to the Chinese dataset (outliers not shown).

The boxplot of estimation errors obtained with the corrected NESMA method is shown in Figure 8: it can be noticed that the mean error is just above zero, while the median error is just below zero.

The error distribution is shown in Figure 9: it can be noticed that the distribution is less skewed than in Figure 7.

In what follows we consider the "Corrected NESMA" estimates, obtained as follows:

$$EstSize_{UFP} = 1.04\ (7\ \#ILF + 5\ \#EIF + 4\ \#EI + 5\ \#EO + 4\ \#EQ)$$

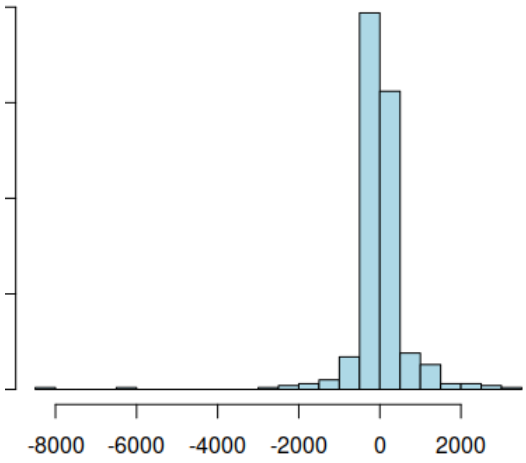

Fig. 9. Boxplot of estimation errors by the corrected NESMA method, when applied to the Chinese dataset.

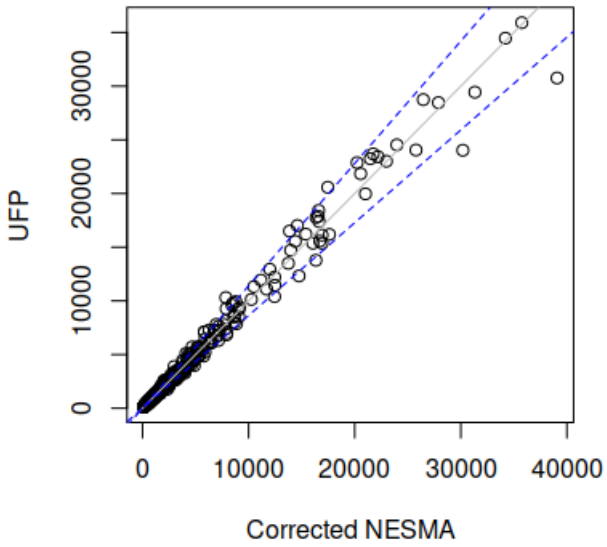

Fig. 10. Corrected NESMA estimates vs. actual size in UFP, with confidence $C = 0.75$, for the Chinese dataset.

F. Results obtained for the Chinese dataset

We applied the procedure described in Section III-C for various confidence levels. The results obtained are given in Table IV. As expected, the results obtained for the Chinese datasets are different from those derived from the ISBSG dataset: for a a given confidence level, we obtained different confidence intervals. Specifically, the confidence intervals are larger for the Chinese dataset than for the ISBSG dataset. For instance, the minimum and maximum ratios $\frac{ActualSize}{EstSize_{UFP}}$ are 0.741 and 1.597, respectively, while they were 0.758 and 1.343, respectively, for the ISBSG dataset.

TABLE IV
CONFIDENCE INTERVALS FOR VARIOUS CONFIDENCE LEVELS, FOR THE CHINESE DATASET.

conf. level	k_L	k_H
0.10	0.984	1.018
0.20	0.967	1.033
0.30	0.950	1.051
0.40	0.933	1.066
0.50	0.921	1.082
0.60	0.905	1.102
0.70	0.882	1.119
0.80	0.843	1.165
0.90	0.791	1.214
0.95	0.741	1.248
1.00	0.741	1.597

For illustration purposes, Figure 10 plots the ISBSG project data in the plan defined by actual size (the y axis) and the size estimated via the Corrected NESMA method (the x axis). In the plot, the dashed blue lines represent the $y = k_L\ x$ and $y = k_H\ x$ lines.

IV. DISCUSSION OF RESULTS

In the previous sections, we exploited two datasets that collect measures from real-life projects to determine i) a correction of the estimates provides by the NESMA method, and ii) confidence intervals for the corrected estimates.

The results of the study show that organizations that own historical data like those we used can apply the procedure illustrated in Sections III-B and III-C to derive the correction constant and the confidence intervals that suite best their development process.

Unfortunately, organizations that do not own historical data like those we used cannot derive the correction constant nor confidence intervals. However, they can adopt some rule of thumb to improve the performances of NESMA estimates. Specifically, the correction constant can be set to a value between 1.04 and 1.09, based on our findings. Similarly, confidence intervals can be defined, based on Tables III and IV. However, these organizations should be aware that the data we used might not match their situations, hence both the correction constant and the confidence intervals might not be perfectly suited for their case.

The confidence interval can be used to perform risk analysis. For instance, Table III shows that, given an estimate already corrected with respect to the NESMA original prediction, there are 30% probabilities that the actual size is more than 10% different (greater or smaller) than estimated. Most likely, half of these 30% probabilities concern underestimation: as a result, a project manager should consider that the probability of underestimating functional size of 10% or more is around 15%. The risk concerning the underestimation of cost can be then computed, if the relationship between size and cost is known.

Finally, being the estimates obtained via the Corrected NESMA method proportional to the estimates obtained via the original NESMA method, the confidence intervals for the Corrected NESMA method can be easily converted into confidence intervals for the original NESMA method.

V. Practical Usage of Size Estimate Intervals

In this section, the practical utility of the proposed method for computing confidence intervals for size is illustrated via an example, concerning the most typical usage of functional size metrics, i.e., effort estimation.

Suppose that Jane, a software project manager, has to estimate the effort required for developing a new application.

For effort estimation, she is using a model, shown in Figure 11, which estimates effort based on the size of the software to be developed. Note that the model shown in Figure 11 includes confidence intervals, which must not be confused with the confidence intervals discussed above: these are the confidence intervals embedded in the *effort* estimation model. That is, assuming that the provided size measure represents exactly the amount of software to be developed, the effort estimation model shows that the required development effort can vary because of many reasons, not connected with size: e.g., the characteristics of non functional requirements, the adopted process, the characteristics of developers, etc. Specifically, minimum and maximum effort values are provided, corresponding to some confidence level.

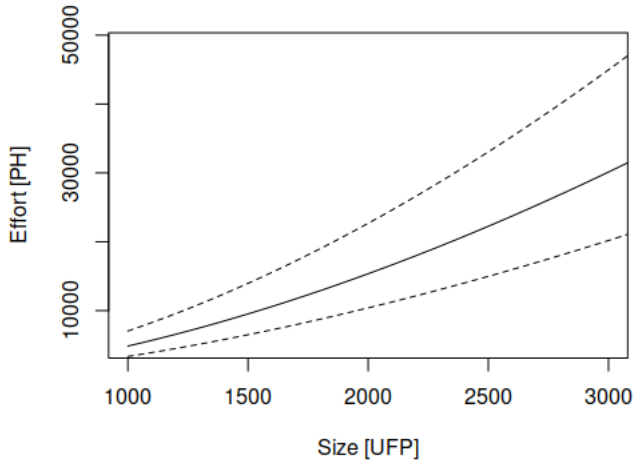

Fig. 11. An effort model with confidence intervals.

Jane, who read this paper, estimated the size of the application to be developed using the corrected NESMA method, and obtained that the estimated size is 2000 UFP. According to the model, developing an application of that size will likely take 15350 PH. With the confidence level being used, the effort will be in the [10393, 22669] PH range, as shown in Figure 12.

Now, Jane computes the confidence interval for the estimated size, using the procedure described in Section III-C. Since the size of the application is around 2000 UFP, Jane decides to use her company's data concerning projects in the [1000, 4000] UFP range to compute the confidence interval (this example uses ISBSG data; that is, for illustration purposes, we assume that Jane's company data are identical to ISBSG data). In this way, she finds that at a 0.75 confidence level the size of the application to be developed is in the [1760, 2248] UFP range.

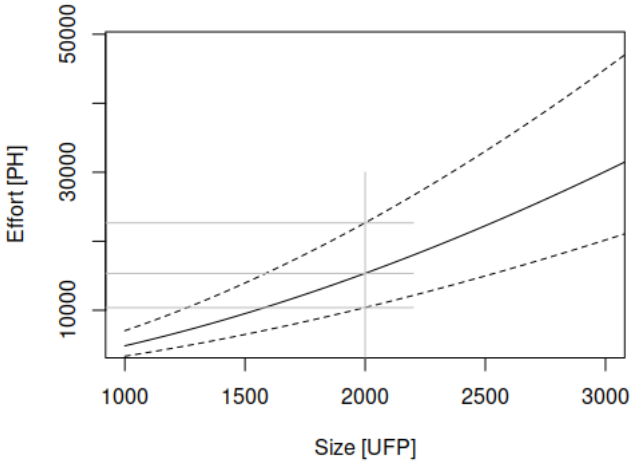

Fig. 12. Estimated effort based on the model and on the NESMA estimation of size.

With these values, Jane can now recompute the estimated effort. In the optimal case, i.e., when the effort model is given by the lower line in Figure 11 and the size of the application is 1760 UFP, the estimated effort is 8243 PH. In the worst case, i.e., when the effort model is given by the upper line in Figure 11 and the size of the application is 2248 UFP, the estimated effort is 27595 PH. The computations are illustrated graphically in Figure 13.

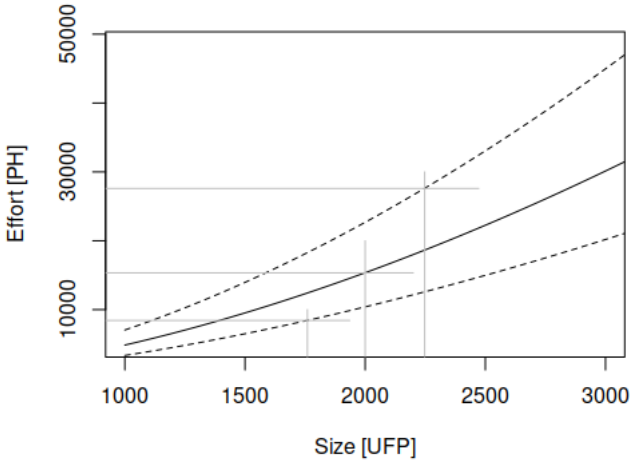

Fig. 13. Estimated effort considering both the confidence intervals of the model and those of the size estimate.

In conclusion, Jane finds out that a project that, according to 'one shot' estimation (using the likely effort model and the likely size estimate) requires 15350 PH, could instead require 27595 PH, i.e., 180% the one shot effort estimate (!) or 8243 PH, i.e., only 54% of the one shot estimate. This knowledge will allow Jane to devise a proper risk management strategy.

VI. Threats to validity

The proposed approach is empirical. In fact, the context itself suggests that a strong theoretical basis is not very

relevant. The definition of the NESMA estimated method itself has no theoretically strong basis: the method is based on the simple hypothesis that, on average, data have low complexity (in FPA terms) and transactions have mid complexity. So, we looked for reasonable confidence intervals, although these intervals are not statistically linked to confidence levels in a rigorous way.

Another typical concern in this kind of studies is the generalizability of results outside the scope and context of the analyzed dataset. We replicated the study with two datasets that are quite representative of real-life projects (the ISBSG dataset is deemed the standard benchmark among the community, and it includes data from several application domains, while the Chinese dataset collects data from a large set of banking and financial software projects). Therefore, our results may be representative of a fairly comprehensive situation. The general result we got is that the amount by which the NESMA method underestimates depends on the considered dataset; similarly, the confidence interval depends on the dataset. At any rate, it is worth underlying that while the numeric results we obtained are not applicable to datasets from different organizations, the proposed method is generally applicable as-is, in any context, provided that representative data are available.

VII. Related Work

Measures for early software estimation were conceived since the last decades [22]–[24]. The present study aims to advance this field by providing statistical foundations to some of these measures, by using confidence intervals where approaches not based on probability distributions were adopted. For example, the "Early & Quick Function Point" (EQFP) method [25] estimates an error of $\pm 10\%$ of the real size of software, for most of the times, but fails to indicate a more robust indicator of this estimate, such as a confidence interval. Several other early estimation methods were proposed: Table V lists the most popular ones.

TABLE V
Early estimation methods: definitions and evaluations

Method name	Definition	Used functions	Weight	Evaluation
NESMA indicative	[26] [27]	data	fixed	[4] [21], [28]–[31] [11]
NESMA estimated	[26] [27]	all functions	fixed	[4] [21], [28]–[31] [11]
Early & Quick FP	[24] [32] [25]	all functions	statistics	[11] [33]
simplified FP (sFP)	[34]	all functions	fixed	[11]
ISBSG average weights	[35]	all functions	statistics	[11]
SiFP	[36]	data and trans.	statistics	[13] [15]

Recently, comparisons based on the accuracy of the NESMA estimated (alias HLFPA) method and statistical modelling methods were carried out in order to assess whether standard measures fail in underestimating or overestimating software size [18].

A survey [37] reports how machine learning techniques were used for software development effort estimation, reporting accuracy as a comparison criterion for all the methods analysed. To the best of our knowledge, confidence intervals are overlooked as robust indicators of the estimates done in software size. In this respect, this study aims to emphasize the importance of providing robust indicators for a more reliable comparison and precision of reporting.

VIII. Conclusion

The "NESMA estimated" method was proposed to estimate the functional size of software (expressed in IFPUG Function Points). The NESMA method assigns fixed weights to base functional components (i.e., ILF, EIF, EI, EO and EQ), so that it is not necessary to analyze in depth every logic data file or transaction. This makes the method both easier and faster, and applicable when the details needed to characterize and weight base functional components are not yet available.

Previous studies showed that the NESMA method is sufficiently accurate to be used in practice. However, it has two possibly relevant limitations: 1) it tends to underestimate the "real" (i.e., as obtained via the IFPUG FPA process) size of software, and 2) it yields a single estimate, with no confidence intervals. Both these characteristics can be be problematic for software project managers. In fact, planning a project based on underestimated size and, consequently, on underestimated effort estimates usually leads to unrealistic plans. Besides, getting a confidence interval for size estimates allows for evaluating the risks connected with imprecise size estimates.

In this paper, we have proposed a correction for the estimates yielded by the NESMA method, to avoid underestimation, and a procedure to compute the confidence interval. Both these contributions are expected to make project managers' life easier.

It is important to remark that both the amount by which the NESMA method underestimates and the confidence intervals depend on the considered dataset. Hence, it is quite advisable that organizations that want to use the proposed techniques do so with their own data, which are expected to represent well the organization's projects.

Acknowledgment

The work reported here was partly supported by Fondo per la Ricerca di Ateneo, Università degli Studi dell'Insubria.

References

[1] L. Lavazza, A. Locoro, and R. Meli, "Estimating functional size of software with confidence intervals," in Proceedings of SOFTENG 2023: The Ninth International Conference on Advances and Trends in Software Engineering, 2023, pp. 14–19.

[2] A. J. Albrecht, "Measuring application development productivity," in Proceedings of the joint SHARE/GUIDE/IBM application development symposium, vol. 10, 1979, pp. 83–92.

[3] International Function Point Users Group (IFPUG), "Function point counting practices manual, release 4.3.1," 2010.

[4] 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.

[5] nesma, "nesma site," https://nesma.org/ [retrieved: March, 2023].

[6] A. Timp, "uTip – Early Function Point Analysis and Consistent Cost Estimating," 2015, uTip # 03 – (version # 1.0 2015/07/01).

[7] L. Lavazza, "On the effort required by function point measurement phases," International Journal on Advances in Software, vol. 10, no. 1 & 2, 2017, pp. 108–120.

[8] International Standardization Organization (ISO), "ISO/IEC 20926: 2003, Software engineering – IFPUG 4.1 Unadjusted functional size measurement method – Counting Practices Manual," 2003.

[9] IFPUG, "Simple Function Point (SFP) Counting Practices Manual Release 2.1," 2021.

[10] nesma, "Early Function Point Analysis," https://nesma.org/themes/sizing/function-point-analysis/early-function-point-counting/ [retrieved: March, 2023].

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

[12] International Software Benchmarking Standards Group, ""Worldwide Software Development: The Benchmark, release 11," ISBSG, 2009.

[13] L. Lavazza and R. Meli, "An evaluation of simple function point as a replacement of IFPUG function point," in IWSM–MENSURA 2014. IEEE, 2014, pp. 196–206.

[14] L. Lavazza, S. Morasca, and D. Tosi, "An empirical study on the effect of programming languages on productivity," in Proceedings of the 31st Annual ACM Symposium on Applied Computing, 2016, pp. 1434–1439.

[15] F. Ferrucci, C. Gravino, and L. Lavazza, "Simple function points for effort estimation: a further assessment," in 31st Annual ACM Symposium on Applied Computing. ACM, 2016, pp. 1428–1433.

[16] L. Lavazza, S. Morasca, and D. Tosi, "An empirical study on the factors affecting software development productivity," E-Informatica Software Engineering Journal, vol. 12, no. 1, 2018, pp. 27–49.

[17] L. Lavazza, G. Liu, and R. Meli, "Productivity of software enhancement projects: an empirical study." in IWSM-Mensura, 2020, pp. 1–15.

[18] G. Liu and L. Lavazza, "Early and quick function points analysis: Evaluations and proposals," Journal of Systems and Software, vol. 174, 2021, p. 110888.

[19] L. Lavazza, A. Locoro, G. Liu, and R. Meli, "Using locally weighted regression to estimate the functional size of software: an empirical study," International Journal on Advances in Software, vol. 15, no. 3-4, 2022, pp. 211–223.

[20] ——, "Estimating software functional size via machine learning," ACM Transactions on Software Engineering and Methodology, 2023.

[21] L. Lavazza and G. Liu, "An Empirical Evaluation of the Accuracy of NESMA Function Points Estimates," in ICSEA, 2019, pp. 24–29.

[22] D. B. Bock and R. Klepper, "FP-S: a simplified function point counting method," Journal of Systems and Software, vol. 18, no. 3, 1992, pp. 245–254.

[23] G. Horgan, S. Khaddaj, and P. Forte, "Construction of an FPA-type metric for early lifecycle estimation," Information and Software Technology, vol. 40, no. 8, 1998, pp. 409–415.

[24] L. Santillo, M. Conte, and R. Meli, "Early & Quick Function Point: sizing more with less," in 11th IEEE International Software Metrics Symposium (METRICS'05). IEEE, 2005, pp. 41–41.

[25] DPO, "Early & Quick Function Points Reference Manual - IFPUG version," DPO, Roma, Italy, Tech. Rep. EQ&FP-IFPUG-31-RM-11-EN-P, April 2012.

[26] NESMA–the Netherlands Software Metrics Association, "Definitions and counting guidelines for the application of function point analysis. NESMA Functional Size Measurement method compliant to ISO/IEC 24570 version 2.1," 2004.

[27] International Standards Organisation, "ISO/IEC 24570:2005 – Software Engineering – NESMA functional size measurement method version 2.1 – definitions and counting guidelines for the application of Function Point Analysis," 2005.

[28] F. G. Wilkie, I. R. McChesney, P. Morrow, C. Tuxworth, and N. Lester, "The value of software sizing," Information and Software Technology, vol. 53, no. 11, 2011, pp. 1236–1249.

[29] J. Popović and D. Bojić, "A comparative evaluation of effort estimation methods in the software life cycle," Computer Science and Information Systems, vol. 9, no. 1, 2012, pp. 455–484.

[30] P. Morrow, F. G. Wilkie, and I. McChesney, "Function point analysis using nesma: simplifying the sizing without simplifying the size," Software Quality Journal, vol. 22, no. 4, 2014, pp. 611–660.

[31] S. Di Martino, F. Ferrucci, C. Gravino, and F. Sarro, "Assessing the effectiveness of approximate functional sizing approaches for effort estimation," Information and Software Technology, vol. 123, July 2020.

[32] T. Iorio, R. Meli, and F. Perna, "Early&quick function points® v3. 0: enhancements for a publicly available method," in SMEF, 2007, pp. 179–198.

[33] R. Meli, "Early & quick function point method-an empirical validation experiment," in Int. Conf. on Advances and Trends in Software Engineering, Barcelona, Spain, 2015, pp. 14–22.

[34] L. Bernstein and C. M. Yuhas, Trustworthy systems through quantitative software engineering. John Wiley & Sons, 2005, vol. 1.

[35] R. Meli and L. Santillo, "Function point estimation methods: A comparative overview," in FESMA, vol. 99. Citeseer, 1999, pp. 6–8.

[36] R. Meli, "Simple function point: a new functional size measurement method fully compliant with IFPUG 4.x," in Software Measurement European Forum, 2011, pp. 145–152.

[37] M. N. Mahdi, M. H. Mohamed Zabil, A. R. Ahmad, R. Ismail, Y. Yusoff, L. K. Cheng, M. S. B. M. Azmi, H. Natiq, and H. Happala Naidu, "Software project management using machine learning technique—a review," Applied Sciences, vol. 11, no. 11, 2021, p. 5183.