Conversion from IFPUG FPA to COSMIC: within- vs without-company equations

Filomena Ferrucci
Department of Management
and Information Technology
University of Salerno, Italy
Email: fferrucci@unisa.it

Carmine Gravino
Department of Management
and Information Technology
University of Salerno, Italy
Email: gravino@unisa.it

Federica Sarro
CREST, Department of
Computer Science
University College London
Email: f.sarro@ucl.ac.uk

***Abstract*—Companies have employed for years 1st generation Functional Size Measurement (FSM) methods, e.g., IFPUG Function Points Analysis (FPA), collecting IFPUG-based historical data useful for benchmarking and estimation purposes. With the advent of 2nd generation FSM methods (e.g., COSMIC) the need for resizing past projects utilizing these new measures arises. The adoption of 2nd generation FSM methods has been limited both by the costs for acquiring new know-how and the need for resizing the historical data. Conversion equations represent a useful mean to facilitate the migration towards 2nd generation FSM methods. Previous works showed a significant correlation between the COSMIC and the FPA sizes. In our study we applied the conversion equations found in those works (i.e., without-company equations) to resize 25 IFPUG-based projects coming from a single software company. We compared their use with respect to two conversion equations built by using two small subsets (i.e., 5 and 10 projects) from that company data (i.e., within-company equations). We aimed to verify whether the use of within-company equations built using few projects could provide more accurate conversions than those achieved by applying without-company equations. Our analysis revealed that the within-company equations performed significantly better than the without-company ones. Thus, companies should develop their own equations rather than using without-company conversions.**

***Keywords*-COSMIC; IFPUG; Conversion equations**

I. Introduction

To allow projects to be delivered on time and within budget it is crucial to have size measures to allow the proper allocation of resources, control costs and schedules, and monitor productivity. Indeed, the lack of accurate size estimations is recognized as the main cause of poor project management [16]. To this aim Functional Size Measurement (FSM) methods are widely used. Function Points Analysis (FPA) was the first approach, introduced by Albrecht of IBM in 1979 [4]. Since then several variants have been defined (e.g., MarkII and NESMA) with the aim of improving size measurement or extending the applicable domain [14]. They all fall in the first generation of FSM methods, distinguishing them from COSMIC, which is considered a second generation FSM method due to several distinguishing characteristics: it is based on fundamental principles of software engineering and measurement theory, it is applicable to business, real-time, and infrastructure software (or hybrids of these) [3].

A common problem for companies aiming at migrating from FPA to COSMIC is the lack of historical data based on COSMIC [7]. The reuse of FPA data could be very valuable to address this problem provided that there is a way to obtain the size in terms of COSMIC from the size in terms of FPA. Nevertheless, as it was pointed out [1], FPA and COSMIC measure different aspects of software systems since they are based on different basic functional components and a mathematical conversion between them cannot be conceived. A possible way to address the problem, also suggested in the COSMIC documentation [1], is to search for a statistical conversion. Some researchers have been studying the suitability and the effectiveness of such an approach by trying to build FPA-COSMIC conversion equations for different data sets. In particular, linear and non linear equations have been built by applying the linear regression analysis to the raw data and to the log-transformed data, respectively [7]. Also, more sophisticated techniques, such as piecewise regression, have been employed for building non linear models [24]. The results reported in the literature reveal that a statistical conversion is indeed possible thus supporting the suggestions provided in the COSMIC documentation [1]. The studies also show that both linear and non linear models should always be analyzed to identify the best correlation [24].

In order to build a FPA-COSMIC conversion equation a company needs to have at least a certain number of projects sized with both methods. However, in some cases, it might not want to spend time and money for this double sizing. This motivated us to investigate if it can be useful to exploit conversion equations built with other (without-company) data sets [27]. To this aim, we used some FPA-COSMIC conversion equations proposed in previous studies for other data sets (i.e., without-company equations) to resize historical data coming from a different single-company and compared their accuracy with respect to the one provided by two within-company FPA-COSMIC conversion equations. To the best of our knowledge this is the first study that investigated this issue, while other studies reported in the literature have addressed only the problem of building conversion models starting from the available data set. In particular, the research question of our study is:

RQ Are within-company FPA-COSMIC conversion equations significantly better than without-company FPA-COSMIC conversion equations?

To address the research question, we employed a data set of 25 applications and used linear regression to build conversion equations. As for the validation method we employed cross-validation, while we exploited descriptive statistics of absolute residuals and statistical tests to evaluate and compare the obtained size estimations.

The remainder of the paper is organized as follows. In Section II we briefly report on the size measures we have taken into account in our case study and present the related work. In Section III we describe the experimental method we exploited to perform the case study, while the results of the empirical analysis are reported and discussed in Section IV. Conclusion and future work concludes the paper.

II. Background

A. Employed Functional Size Measures

1) IFPUG: This method sizes an application starting from its FURs (or by other software artifacts that can be abstracted in terms of FURs). In particular, to identify the set of "features" provided by the software, each FUR is functionally decomposed into Base Functional Components (BFC), and each BFC is categorized into one of five Data or Transactional BFC Types. The Data functions are Internal Logical Files and External Interface Files, while the Transactional ones are External Inputs, External Outputs and External Inquiries.

Then, the "complexity" of each BFC is assessed. This step depends on the kind of function type and requires the identification of further attributes (such as the number of data fields to be processed). Once derived this information, a table provided in the IFPUG approach method [18] specifies the complexity of each function, in terms of Unadjusted Function Points (UFP). The sum of all these UFPs gives the functional size of the application. Subsequently, a Value Adjustment Factor (VAF), can be computed to take into account some non-functional requirements, such as Performances, Reusability, and so on. The final size of the application is terms of Function Points is given by $FP = UFP \cdot VAF$.

We are aware that there is a huge discussion about the acceptance of the VAF and the ISO standard never included the VAF since it measures non-functional aspects of the software [18]. However, even if the VAF is not part of the ISO standard, it is still part of the IFPUG method and is documented in IFPUG documentation. Since the company invoked in our study usually employ IFPUG with VAF we decided to consider it in our empirical study.

From this very brief description of the method, some of the appealing advantages of IFPUG over competitors (i.e., other software measures) becomes highlighted. In particular, the success of IFPUG can be mainly motivated by its early applicability and by its independence from the adopted programming language. It is widely exploited to estimate productivity, in terms of Function Points per person-month, and quality, in terms of the number of defects with respect to requirements, design, coding and user documentation phases. For more details about the application of the IFPUG method, readers are referred to the counting manual [18].

2) COSMIC: It stands for COmmon Software Measurement International Consortium [3]. It has been approved as an international standard by ISO (ISO/IEC 19761:2003). This method was originally conceived for business and real-time applications, but it can be applied in any other context, where software is mainly dominated by large amounts of data movements. Indeed, the basic idea underlying COSMIC is that, for these kinds of software, the majority of development efforts are devoted to handle data movements from/to persistent storage and users. Thus, the number of these movements can provide a meaningful insight of the system size [3].

COSMIC is composed of a set of models, principles, rules, and processes applicable to the FURs of a given piece of software. The first step to obtain the COSMIC measure is to define the COSMIC Software Context Model. This model introduces the principles and concepts needed to identify the FURs of the piece of software to be measured. Then, the COSMIC Generic Software Model is applied to the FURs to identify the components of the functionality that will be measured. In particular, the Software Context Model allows us to define why a measurement is required (Purpose) and to select the set of FURs to be included in a specific functional size measurement task (Scope). Furthermore, Functional User and Boundary concepts can be specified. A Functional User is a (type of) user that is a sender and/or an intended recipient of data in the FURs of a piece of software, while the Boundary is defined as a conceptual interface between the software being measured and its functional users. Then, the concepts of the Generic Software Model are applied to the FURs, to identify the set of "features" provided by the software. This model is the key of the COSMIC method, assuming that (i) each FUR can be mapped into a unique functional process, (ii) each functional process consists of sub-processes, and (iii) each sub-process may be either a data movement or a data manipulation. A Functional Process is an elementary component of a set of FURs comprising a unique, cohesive, and independently executable set of data movements. It is triggered by a data movement from a functional user that informs the piece of software that the functional user has identified a triggering event. A Functional Sub-Process may be a data movement or a data manipulation. As an approximation for measurement purposes, data manipulation sub-processes are not separately measured; the functionality of any data manipulation is assumed to be accounted for the data movement associated with it [3]. To identify and measure data movements other three concepts have to be specified. A Triggering Event is an event (something that happens) that causes a functional user of the piece of software to initiate (trigger) one or more functional processes. A Data Group is a distinct, non empty, non ordered, and non redundant set of data attributes describing a complementary aspect of the same object of interest. A Data Attribute is the smallest piece of information, within an identified data group, carrying a meaning from the perspective of the interested FURs. Data movements are defined as follows: an Entry data movement moves a data group from a functional user across the boundary into the functional process where it is required; an Exit data movement moves a data group from a functional process across the boundary to the functional user that requires it; a Read data movement moves a data group from persistent storage within each of the functional process that requires it; a Write data movement moves a data group lying inside a functional process to persistent storage.

In the measurement phase, the data movements of each functional process have to be identified and used to obtain a size measure of the software. Each data movement is counted as 1 COSMIC Function Point (CFP) that is the COSMIC measurement standard. Thus, the size of an application within a defined scope is obtained by summing the sizes of all the functional processes identified. For more details about the COSMIC method, readers are referred to the COSMIC Measurement Manual [3].

B. *Previous studies proposing conversion equations*

The most comprehensive study about the relationship between the size expressed in terms of (IFPUG) FPA and the size given in terms of COSMIC is the one by Cuadrado-Gallego *et al.* [7]. Their aim was to carry out a review of previous investigations that mainly exploited linear regression analysis to identify a sound mathematical basis for converting FPA in COSMIC. They employed six data sets obtained from previous studies (i.e., those named *fet99* [11], *fet99-2* [30], *ho99* [17], *vog03* [30], *abr05* [2], and *des06* [8] in Tables I and II) and three new data sets: the first two (i.e., those named *jjcg06* and *jjcg07*) contained 21 and 14 data points and were obtained in two different studies conducted with academic students while the third (i.e., the one named *jjcg0607*) one was obtained by merging the first two. Cuadrado-Gallego *et al.* also performed statistical analysis such as the confidence intervals (calculated at a 95% confidence level) of the coefficient (i.e., the "slope") for the conversion equations.

As for the estimation technique, these studies employed OLSR to build FPA-COSMIC conversion equations. In particular, the aim was to identify a mathematical conversion from (IFPUG) FPA into COSMIC, i.e., by constructing an equation as:

$$CFP = a + b \cdot FP \tag{1}$$

where the independent variable CFP represents the size in terms of COSMIC and the independent variable FP represents the size in terms of (IFPUG) FPA. Furthermore, as suggested in [7], starting from the observation that an important aspect when applying a conversion approach is that the equation must consider the origin of coordinates, we have also analyzed a non linear relationship between CFP and FP by exploiting the log transformation of the variables in the application of linear regression analysis. Thus, the equation obtained is of this form:

$$Log(CFP) = Log(a) + b \cdot Log(FP) \tag{2}$$

which, when transformed back to the original raw data scale, gives the equation:

$$CFP = a \cdot FP^b \tag{3}$$

Tables I and II show for each employed data set (i.e., 1st column), the equation parameters a and b (i.e., 2nd and 3rd columns) and their confidence intervals (i.e., 4th and 5th columns), and the R^2 value (i.e., 6th column). We can observe that the equations obtained for the first six data sets in Table I are characterized by a coefficient around 1 but with a most probable interval for b of (1.1, 1.2). On the other hand, for the last three data sets, a coefficient very close to 1 was obtained and the intersection of most probable interval for b was (0.7, 0.8). As for non linear equations (see Table II), they show a very similar behavior in the exponent (around 1) with a probable interval of (0.9, 1.1) [7].

It is worth to mention that the data set *fet99* was obtained by Fetcke [11] who measured 5 applications for warehouse management by exploiting IFPUG 4.1 and COSMIC 2.2 (the maximum number of FP was 77). They were business applications with few data entities and no conversion equation was formally presented due to the limited number of observations. Ho *et al.* further analyzed the data provided by Fetcke and corrected the measurement in terms of COSMIC obtaining the data set named *ho99* in Table I. However, they did not provide the conversion equation [17].

The data set *vog03* was obtained by Vogelezang and Lesterhuis by measuring 11 applications (from a financial services organization) with NESMA 2.0 and COSMIC 2.2 (the maximum number of FP was 1424). As observed by Cuadrado-Gallego *et al.* [7], even if IFPUG FPA unit[1] was not formally used, the relationship between the sizes in terms of NESMA 2.0 and IFPUG 4.1 was 1:1 that made it possible to carry out the IFPUG to COSMIC conversion study. Vogelezang and Lesterhuis were the first to propose a linear equation for converting the size in terms of FPA into the size given in terms of COSMIC. Furthermore, they also provided the conversion equation for the data set *fet99*.

[1] Functional size unit (FSU) is the the widely used ISO term for the unit of measure of any FSM method [7]

Table I
PARAMETERS OBTAINED IN [7], USING LINEAR EQUATIONS

Data set	b	a	Confidence Interval a	Confidence Interval b	R^2
fet99	1.1	-6.2	(-20.7, 8.2)	(0.8, 1.4)	0.98
fet99-2	1.1	-7.6	(-40.1, 24.9)	(0.6, 1.7)	0.97
ho99	1	6.5	(-19.4, 6.3)	(0.8, 1.3)	0.98
vog03	1.2	-86.8	(-144.2, -29.4)	(1.1, 1.3)	0.99
abr05	0.8	18	(-182.1, 218.1)	(0.5, 1.2)	0.91
des06	1	-3.2	(-66,59.5)	(0.8, 1.2)	0.93
jjcg06	0.8	-36.6	(-121.7, 48.5)	(0.6, 1.1)	0.70
jjcg07	0.9	0.2	(-29.5, 29.5)	(0.7, 1.1)	0.86
jjcg0607	0.7	-4.5	(-26.8, 17.9)	(0.7, 1.8)	0.90

Table II
PARAMETERS OBTAINED IN [7], USING NON LINEAR EQUATIONS

Data set	b	a	Confidence Interval a	Confidence Interval b	R^2
fet99	1.1	0.6	(0.2, 1.7)	(0.9, 1.4)	0.99
fet99-2	1.1	60.6	(0.1, 6.8)	(0.5, 1.7)	0.97
ho99	1.1	0.7	(0.2, 1.4)	(0.9, 1.4)	0.99
vog03	1.2	0.3	(0.1, 1)	(0.9, 1.4)	0.94
abr05	1.1	1.1	(0.1, 18.1)	(0.5, 1.4)	0.88
des06	1.1	0.7	(0.3, 1.5)	(0.9, 1.2)	0.95
jjcg06	1.2	0.3	(0.1, 1.3)	(0.9, 1.4)	0.82
jjcg07	1	0.8	(0.1, 4.8)	(0.6, 1.4)	0.73
jjcg0607	1	1	(0.9, 1.1)	(0.9, 1.1)	0.99

The data set *abr05* was obtained by Abran *et al.* [2] who exploited 6 Management Information System (MIS) applications from a governmental organization. These applications were measured using the documentation of completed projects (the maximum number of FP was 1146). They provided a linear equation for this data set. In their discussion they also analyzed the results obtained with the data sets *fet99* and *vog03*. They also performed a new analysis on the data set provided by Fetcke [11] by modifying the measurement and obtaining a new version of the data set, namely *fet99-2*, and a new conversation equation.

The data set *des06* was built by Desharnais *et al.* [8], who exploited data from a set of 14 MIS applications (the maximum number of FP was 647) developed by a single governmental organization.

III. DESIGN OF THE EMPIRICAL STUDY

A. Data set

The empirical study we present in this paper is based on a data set coming from a medium-sized software company operating in Italy, whose core business is the development of enterprise information systems, mainly for local and central government. The company is specialized in the development and management of solutions for enterprise intranet/extranet applications (such as Content Management Systems, e-commerce, work-flow managers, etc), and Geographical Information Systems. It has about 50 employees and a turnover of about 5M E and is certified ISO 9001:2000, and partner of Microsoft, Oracle, and ESRI.

Data used in the study are related to a set of 25 applications, including e-government, e-banking, and Intranet applications, developed with different technologies (e.g., J2EE, ASP.NET). Oracle has been the most commonly adopted DBMS, but also SQL Server, Access and MySQL were employed in some cases.

Table III
DESCRIPTIVE STATISTICS OF EFF, CFP, AND FP

Var	Obs	Min	Max	Mean	Median	Std. Dev.
EFF	25	782	4537	2577	2686	988.136
CFP	25	163	1090	602	611	268.473
FP	25	89	915	366.76	303.94	208.65

It is worth pointing out that one of the main difficulty for this kind of study is the availability of information on the size of the applications in terms of two or more size measurement methods (in our case COSMIC and FPA). We obtained this information from the software company involved in our study during a long term investigation aimed at verifying the effectiveness of COSMIC and FPA in estimating development effort (see e.g., [9]). In particular, two researchers followed the measurement process and cross-checked the measures (taking into account the documents exploited by measurers of the company (i.e., project managers). The projects managers were also involved in an adequate training program and the two researchers work on this topic for 10 years.

Table III provides some descriptive statistics about the considered data set, where CFP denotes the size expressed in terms of COSMIC and FP in terms of FPA.

B. Employed estimation technique

We employed Ordinary Least-Squares Regression (OLSR) to build FPA-COSMIC conversion equations for our data set, as done in previous similar studies (e.g., [2] [7] [8] [17]) [10]. In particular, we exploited simple linear regression [25] to obtain a model as the one shown in equation 1, where CFP is the dependent variable and FP is the independent variable. To evaluate the goodness of fit of a regression model, several indicators have to be considered. Among them, the square of the linear correlation coefficient, R^2, shows the amount of the variance of the dependent variable explained by the model related to the independent variable. Other useful indicators are the F value and the corresponding $p-value$ (denoted by $SignF$), which high and low values, respectively, denote a high degree of confidence for the prediction. We have also considered the $p-values$ and $t-values$ for the corresponding coefficients and the intercept. The p-values give an insight into the accuracy of the coefficients and the intercept, whereas their t-values allow us to evaluate their importance for the generated model. In particular, p-values less than 0.05 are considered an acceptable threshold, meaning that the variables are significant predictors with a confidence of 5%. As for the t-value, a variable is significant if its corresponding value is greater than 1.5.

Moreover, similarly to the study presented in [7], we also carried out an analysis by first exploiting a log transformation of the original data and then applying linear regression analysis. This strategy is also usually employed whenever variables are highly skewed and they are transformed before applying linear regression analysis. This is done in order to comply with the assumptions underlying linear regression [25] (i.e., residuals should be independent and normally distributed; relationship between dependent and independent variables should be linear).

C. Validation method and evaluation criteria

To verify whether or not the obtained prediction values are useful estimations of the actual values we carried out a cross validation, which means that the original data set is divided into different subsets of training and validation sets. Training sets are used to build models with OLSR and validation sets are used to validate the obtained models. In particular, we applied a hold-out cross validation that means that each test set is completely different from the corresponding training set [21]. Thus, for the two within-company FPA-COSMIC conversion equations, we used 5 and 10 applications in the original data set as training set while the remaining 20 and 15 were used as test sets, respectively. In the following, we name Train1 and Train2 the two training sets of 5 and 10 applications and Test1 and Test2 the corresponding test sets. On the other hand, in the case of the application of without-company FPA-COSMIC conversion equations, the models have been built on external data sets that are used as training sets, while we employed Test1 and Test2 as test sets, to allow comparison with the two within-company equations.

Observe that we decided to exploit the estimation models provided in [7] and reported in Tables I and II as without-company FPA-COSMIC conversion equations to answer research question. The comparison has been performed by applying the without-company equations to the subsets of Test1 and Test2 exploited as test sets for our within-company FPA-COSMIC conversion equations.

Regarding the evaluation criteria, we used descriptive statistics of absolute residuals (i.e., |*Actual - Predicted*|), namely Mean, Median, and Sum of Absolute Residuals (MAR, MdAR, and SAR). Moreover, we tested the statistical significance of the results, by using absolute residuals, to compare estimates obtained with different approaches (e.g., to establish if one model provided significantly better estimates than another one [22]). In particular, we performed Wilcoxon signed rank test [6] to verify the following null hypothesis "the two considered population of absolute residuals have identical distributions". This kind of test is used to verify the hypothesis that the mean of the differences in the pairs is zero.

In order to have also an indication of the practical/managerial significance of the results we verified the effect size [19]. Effect size is a simple way of quantifying the standardized difference between two groups. It has many advantages over using only the tests of statistical significance, since "whereas p-values reveal whether a finding is statistically significant, effect size indicates practical significance" [19]. In particular, we employed the point-biserial correlation r because it is suitable to compute the magnitude of the difference when a non parametric test is used [12]. In the empirical software engineering field [19], the magnitude of the effect sizes measured using the point-biserial correlation is classified as follows: small (0 to 0.193), medium (0.193 to 0.456), and large (0.456 to 0.868).

Table IV
RESULTS OF THE TESTS TO VERIFY ASSUMPTIONS FOR OLSR

Subset	Pearson's correlation statistic/p-value	Breush-Pagan Test statistic/p-value	Shapiro-Wilk Test statistic/p-value
Train1	0.989/0.001	0.882/0.347	0.849/0.191
Train2	0.840/0.002	0.324/0.569	0.880/0.131

IV. RESULTS AND DISCUSSION

We performed the OLSR analysis to build the linear conversion equations on Train1 and Train2. To this, aim we first verified the underlying assumptions:

- Linearity. For the two models, the linear relationship was confirmed by the Pearson's correlation test (see Table IV) [13]. This result suggested that there was a high and significant correlation between size expressed in terms of CFP and FP, also encouraging to find a mathematical conversation among them.
- Homoscedasticity. We investigated the homoscedasticity assumption by performing a Breush-Pagan Test [5], with the homoscedasticity of the error terms as null hypothesis. As we can see from Table IV, the p-values obtained are greater than 0.05 and thus the null hypothesis cannot be rejected for both the models.
- Normality. To verify this assumption we used the Shapiro-Wilk Test [29], by considering as null hypothesis the normality of error terms. Again, we cannot reject the null hypothesis since the p-values of the statistics are greater than 0.05 for all the models (see Table IV).

The first part of Table V shows some statistics about the models obtained with OLSR employing CFP and FP as dependent and independent variables, respectively. We can observe that the model obtained for Train1 is characterized by high R^2 and F values (i.e., 0.978 and 132, respectively) and a low Sign F. (i.e., 0.001), indicating that the prediction is indeed possible with a high degree of confidence. The t-values and p-values for the corresponding coefficient and the intercept present values greater than 1.5 and less than 0.05, respectively, revealing their significance and importance for the built model. The model built on Train2 is characterized by lower R^2 and F values. However, it has a low Sign F (i.e., 0.002), indicating that the prediction is indeed possible with a high degree of confidence. The t-value and p-value for the coefficient are greater than 1.5 and less than 0.5, revealing that it is important and significant in the model.

Table V
THE (LINEAR AND NON LINEAR) WITHIN-COMPANY CONVERSION EQUATIONS, BUILT ON THE TWO TRAINING SETS

Model	Variable	Value	Std. Err. Err	t-value	p-value	Confidence Interval	R^2	Std. Err	F	Sign. F
Linear on Train1	Coefficient	1.470	0.128	11.508	0.001	(1.064-1877)	0.978	49.170	132	0.001
	Intercept	165.104	41.299	3.998	0.028	(33.671-296.536)				
Linear on Train2	Coefficient	1.383	0.317	4.370	0.002	(0.653-2.113)	0.705	135.900	19	0.002
	Intercept	154.271	101.120	1.526	0.166	(-78.912-387.454)				
Non Linear on Train1	Coefficient	0.631	0.111	5.704	0.011	(0.279-0.983)	0.916	0.157	33	0.011
	Intercept	2.834	0.603	4.701	0.018	(0.915-4.752)				
Non Linear on Train2	Coefficient	0.572	0.168	3.398	0.009	(0.184-0.959)	0.591	0.276	12	0.009
	Intercept	3.071	0.937	3.280	0.011	(0.912-5.231)				

As for the intercept, the t-value revealed that it is important but the p-value suggested that it is not significant.

Moreover, as done in [7], we also carried out the OLSR analysis by first exploiting a log transformation of the original data and then applying linear regression analysis. The results are reported in Table V. We can observe that the models are characterized by R^2 and F values less than those obtained with the corresponding linear equations. These results are not surprising since the analysis of OLSR assumptions revealed that there was a significant (linear) correlation between CFP and FP and transformation could be considered not necessary. However, we also investigated the accuracy of the model obtained by log transforming the variables since we want to also consider the property that the equation must verify the origin of the coordinates [7]. In all the two models, the t-values and p-values for the corresponding coefficient and the intercept present values greater than 1.5 and less than 0.05, respectively. Thus, they are important and significant in the obtained models.

To allow a comparison with previous studies mentioned in Section II-B, in Table V we have also reported the confidence intervals related to the coefficient and the intercept. We can observe that, for linear conversion equations, we obtained a coefficient not so close to the ones obtained in previous studies [7] and reported in Tables I. Indeed, we obtained coefficients greater than 1.3 while the coefficients provided in [7] range from 0.8 to 1.2. This is confirmed by the analysis of confidence intervals. Our two linear conversion equations are characterized by intervals of confidence of the coefficient quite different from the intersection of the most likely intervals shown in Tables I. As for the intercept we obtained greater values in all the models with respect to the ones characterizing the models in previous studies. Furthermore, the models presented in Table I are characterized by greater values for R^2 with respect to our linear models, except for Train1. Thus, taking into account the analysis in terms of coefficient and intercept and R^2 it seems that the conversion equation built on Train2 are less accurate than those obtained in previous studies for different application types.

Concerning non linear conversion equations, we can observe that, differently from the results achieved in [7] and reported in Table II, the exponent is not very close to 1 and the confidence intervals of the coefficient are quite different from those obtained in the previous studies. Thus, we cannot converge to a proposal of a conversation equation based on the assumption of $CFP \approx FP$ as done in [7]. Furthermore, as in the case of linear equations, the conversion equations reported in Table II are characterized by greater values for R^2 with respect to our non linear model built on Train2, while the results achieved with the model built on Train1 has a R^2 value comparable with those of previous studies.

As designed, to evaluate the prediction accuracy of the obtained conversion models, we performed a cross validation as described in Section III-C, whose results for test sets Test1 and Test2 are reported in Tables VI and VII, respectively. We can observe that the results we achieved with the linear equations are better than the ones achieved with the non linear equations for all the two test sets.

Regarding the comparison with without-company conversion equations, we can note that within-company (linear and non linear) conversion equations provided better MdAR, MAR, and SAR than those achieved with without-company (linear and non linear) conversion equations for both the test sets Test1 and Test2.

To verify whether the differences in the results achieved by the considered conversion equations were significant, we performed tests on the statistical significance of the results by using absolute residuals (see Tables VIII and IX). The results of the Wilcoxon test revealed that the estimates achieved with the (linear and non linear) within-company were not statistically significant better than those obtained

Table VI
PREDICTION ACCURACY INDICATORS OBTAINED FOR TEST1

Model built on	Linear equations			Non Linear equations		
	MdAR	MAR	SAR	MdAR	MAR	SAR
fet99	214.30	205.39	4107.70	215.71	204.52	4090.40
fet99-2	214.30	205.39	4107.70	215.71	204.52	4090.40
ho99	256.00	226.10	4522.00	177.94	174.70	3943.98
vog03	262.90	238.34	4766.80	247.31	229.05	4580.97
abr05	337.70	285.40	5708.00	146.44	235.80	4715.91
desh06	265.70	233.61	4672.20	177.94	174.70	3493.98
jjcg06	392.30	335.18	6703.60	247.31	229.05	4580.97
jjcg07	308.90	265.45	6703.60	355.70	301.60	6032.00
jjcg0607	394.35	342.10	6841.90	262.50	231.05	4621.00
Train1	117.90	159.53	3190.66	86.36	130.76	2615.21

with (linear and non linear) without-company conversion equations, except for some in cases. In particular, for Test1 we found a statically significant difference (i.e., p-value less than 0.05) in the cases of abr05, jjcg06, and jjcg0607 for linear models, and in the cases of abr05 and jjcg07 for non linear models, with a large effect size (i.e., r greater than 0.456). The results achieved on Test2 are quite similar to the ones obtained for Test1 for linear models, except for jjcg06, where we also obtained a statically significant difference in favor of the within-company equation. Differently, in the case of non linear models, we found in 4 cases a statistically significant difference: vog03, jjcg06, jjcg07, and jjcg0607.

Thus, we can partially positively answer RQ, i.e., *Within-company FPA-COSMIC conversion equations are significantly better than without-company FPA-COSMIC conversion equations*. Indeed, the results achieved in terms of absolute residuals with within-company equations are better than those obtained with without-company equations, and in many cases the difference is statistically significant.

Implications. The results of our empirical study suggest that the company should build its within-company FPA-COSMIC conversion equations. Indeed, resizing few applications (i.e., just 5 or 10) in terms of COSMIC allowed us to obtain FPA-COSMIC conversion equations with better estimations than those achieved employing without-company FPA-COSMIC conversion equations.

V. Threats to validity

It is widely recognized that several factors can bias the construct, internal, external, and conclusion validity of empirical studies [23] [26]. As for the construct validity, the collection of information about the size measures represents the main difficulty to carry out this kind of study [20]. As described in Section III-A, we were very conscious about performing the collection task in a controlled and uniform way. Of course we have to take into account that empirical studies do not ensure the level of confidence achieved with controlled experiments.

Table VII
Prediction accuracy indicators obtained for Test2

Model built on	Linear equations			Non Linear equations		
	MdAR	MAR	SAR	MdAR	MAR	SAR
fet99	216.80	197.63	2964.50	198.95	195.64	2934.53
fet99-2	216.80	197.63	2964.50	198.95	195.64	2934.53
ho99	279.50	222.10	3331.50	213.17	165.96	2498.47
vog03	245.00	230.37	3455.60	237.06	221.51	3322.65
abr05	351.20	287.41	3455.60	144.20	262.78	3941.74
desh06	289.20	230.17	4311.60	213.17	165.96	2489.47
jjcg06	405.80	335.59	5033.80	237.20	303.01	4545.20
jjcg07	335.20	265.03	3975.50	369.20	303.01	4545.20
jjcg0607	408.70	345.35	5180.20	286.00	227.40	3411.00
Train2	92.08	148.83	2232.45	83.45	119.45	1793.24

Table VIII
Comparison in terms of absolute residuals for Test1

Train1 vs	Linear equations		Non Linear equations	
	WT p-value	Effect size r	WT p-value	Effect size r
fet99	0.192	0.196	0.123	0.263
fet99-2	0.192	0.196	0.123	0.263
ho99	0.147	0.234	0.115	0.271
vog03	0.123	0.263	0.077	0.321
abr05	0.024	0.438	0.028	0.426
desh06	0.139	0.246	0.115	0.271
jjcg06	0.005	0.555	0.077	0.321
jjcg07	0.062	0.346	0.007	0.530
jjcg0607	0.004	0.572	0.054	0.358

Table IX
Comparison in terms of absolute residuals for Test2

Train2 vs	Linear equations		Non Linear equations	
	WT p-value	Effect size r	WT p-value	Effect size r
fet99	0.165	0.257	0.054	0.418
fet99-2	0.165	0.257	0.054	0.418
ho99	0.138	0.286	0.151	0.271
vog03	0.126	0.301	0.028	0.491
abr05	0.032	0.477	0.068	0.389
desh06	0.126	0.301	0.152	0.271
jjcg06	0.015	0.550	0.028	0.491
jjcg07	0.042	0.447	0.006	0.623
jjcg0607	0.018	0.535	0.032	0.476

Factors that should be taken into account for the internal validity are: reliability of the data and lack of standardization [26]. However, the questionnaires used were the same for all the applications and the project managers were instructed on how to use the questionnaires, to correctly provide the required information. Furthermore, instrumentation effects in general did not occur in this kind of studies.

As for the conclusion validity we carefully applied the estimation methods and the statistical tests, verifying all the required assumptions [28]. With regards to the external validity, a threat could be related to the fact that we used only applications from one company. Even if we can image that with a comparable organization, in terms of project size, quality of teams, experience of team members, and application domain, we might expect similar observations, we intend to conduct further studies with data sets from other companies [15].

VI. Conclusion and Future Work

Conversion equations could represent a useful mean to facilitate the migration from (IFPUG) FPA to COSMIC. Indeed, previous studies have shown that a mathematical conversion could be possible between sizes expressed in terms of COSMIC and FPA [7]. The main goal of the study presented here was to verify whether companies could exploit conversion models built on other data sets instead of resizing in terms of COSMIC previous applications measured with FPA and then build their own conversion equations. The results of our empirical analysis revealed that the within-company FPA-COSMIC conversion equations performed better than the without-company FPA-COSMIC conversion equations. The differences in many cases were statistically significant.

As future work, we plan to collect and analyze data from other companies also to investigate whether the kind of applications considered can influence the conversion factors and thus our findings.

References

[1] A. Abran, B.Londeix, M. O'Neill, L. Santillo, F. Vogelezang, J.-M. Desharnais, P. Morris, T. Rollo, C. Symons, A. Lesterhuis, S. Oligny, G. Rule, and H. Toivonen. The COSMIC Functional Size Measurement Method, Version 3.0, Advanced and Related Topics, 2007.

[2] A. Abran, J. Desharnais, and F. Azziz. Measurement convertibility: from function points to COSMIC. In *Procs of the International Workshop on Software Measurement*, page 227240. Shaker-Verlag, 2005.

[3] A. Abran, J.-M. Desharnais, S. Oligny, D. St-Pierre, and C. Symons. The COSMIC Functional Size Measurement Method Version 3.0.1, Measurement Manual, The COSMIC Implementation Guide for ISO/IEC 19761:2003, 2009.

[4] A. Albrecht. Measuring Application Development Productivity. In *Procs of the Joint SHARE/GUIDE/IBM Application Development Symposium*, pages 83–92, 1979.

[5] T. Breush and A. Pagan. A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, 47:1287–1294, 1992.

[6] W. J. Conover. *Practical Nonparametric Statistics*. Wiley, 3rd edition edition, 1998.

[7] J. J. Cuadrado-Gallego, L. Buglione, M. J. Domínguez-Alda, M. F. d. Sevilla, J. Antonio Gutierrez de Mesa, and O. Demirors. An experimental study on the conversion between IFPUG and COSMIC functional size measurement units. *Inf. Softw. Technol.*, 52(3):347–357, 2010.

[8] J. Desharnais, A. Abran, and J. Cuadrado. Convertibility of Function Points to COSMIC-FFP: Identification and Analysis of Functional Outliers. In *Int. Conference on Software Process and Product Measurement*, 2006.

[9] F. Ferrucci, C. Gravino, and S. Di Martino. A Case Study Using Web Objects and COSMIC for Effort Estimation of Web Applications. In *Procs of Euromicro Conference on Software Engineering and Advanced Applications (SEAA'08)*, pages 441–448. IEEE press, 2008.

[10] F. Ferrucci, C. Gravino, and F. Sarro. A Case Study on the Conversion of Function Points into COSMIC. In *EUROMICRO-SEAA*, pages 461–464, 2011.

[11] Fetcke. The Warehouse Software Portfolio, A Case Study in Functional Size Measurement. Technical report, Technical Report No. 199920, Département dinformatique, Université du Quebec á Montréal, Canada, 1999.

[12] A. Field and G. Hole. *How to Design and Report Experiments*. Sage publications Limited, 2003.

[13] J. Freund. *Mathematical Statistics*. Prentice-Hall, Upper Saddle River, NJ, 1992.

[14] C. Gencel and O. Demirors. Functional size measurement revisited. *ACM Trans. Softw. Eng. Methodol.*, 17(3):15:1–15:36, 2008.

[15] S. Ghaisas, P. Rose, M. Daneva, K. Sikkel, and R. Wieringa. Generalizing by similarity: Lessons learnt from industrial case studies. In *In Conducting Empirical Studies in Industry (CESI), 2013*, pages 37–42, May 2013.

[16] R. L. Glass. *Facts and Fallacies of Software Engineering*. Addison Wesley, 2002.

[17] V. Ho, A. Abran, and T. Fetcke. A Comparative Study Case of COSMIC, Full Function Point and IFPUG Methods. Technical report, Département dinformatique, Université du Quebec á Montréal, Canada, 1999.

[18] IFPUG. Intern. Function Point Users Group - www.ifpug.org.

[19] V. Kampenes, T. Dyba, J. Hannay, and I. Sjoberg. A systematic review of effect size in software engineering experiments. *Infor. and Soft. Tech.*, 4(11-12):1073–1086, 2007.

[20] C. Kaner and W. Bond. Software Engineering Metrics: What Do They Measure and How Do We Know? In *Procs of the International Software Metrics Symposium*. IEEE press, 2004.

[21] B. Kitchenham, E. Mendes, and Travassos. Cross versus Within-Company Cost Estimation Studies: A systematic Review. *IEEE Trans. on Soft. Eng.*, 33(5):316–329, 2007.

[22] B. Kitchenham, L. Pickard, S. MacDonell, and M. Shepperd. What accuracy statistics really measure. *IEE Proceedings Software*, 148(3):81–85, 2001.

[23] B. Kitchenham, L. Pickard, and S. Pfleeger. Case studies for method and tool evaluation. *IEEE Software*, 12(4):52–62, 1995.

[24] L. Lavazza and S. Morasca. Convertibility of Function Points into COSMIC Function Points: A study using Piecewise Linear Regression. *Information & Software Technology*, 53(8):874–884, 2011.

[25] K. Maxwell. *Applied Statistics for Software Managers*. Software Quality Institute Series, Prentice Hall, 2002.

[26] E. Mendes, S. Counsell, N. Mosley, C. Triggs, and I. Watson. A Comparative Study of Cost Estimation Models for Web Hypermedia Applications. *Empirical Software Engineering*, 8(23):163–196, 2003.

[27] E. Mendes, M. Kalinowski, D. Martins, F. Ferrucci, and F. Sarro. Cross- vs. Within-Company Cost Estimation Studies Revisited: An Extended Systematic Review. In *International Conference on Evaluation and Assessment in Software Engineering*, 2014.

[28] D. Montgomery, E. Peck, and G. Vining. *Introduction to Linear Regression Analysis*. John Wiley and Sons, Inc., 1986.

[29] P. Royston. An extension of Shapiro and Wilk's W test for normality to large samples. *App. Stat.*, 31(2):115–124, 1982.

[30] F. Vogelezang, A. Lesterhuis, and Sogeti. Applicability of COSMIC Full Function Points in an administrative environment: Experiences of an early adopter. In *Procs of Inter. Workshop on Softw. Meas.*, pages 23–25. Verlag, 2003.