Journal of Object Technology
Published by AITO — Association Internationale pour les Technologies Objets
http://www.jot.fm/

Function Point Structure and Applicability: A Replicated Study

Christian Quesada-López[a] Marcelo Jenkins[a]

a. Center for ICT Research (CITIC)
University of Costa Rica, San José, Costa Rica
{cristian.quesadalopez,marcelo.jenkins}@ucr.ac.cr.

Abstract Background: The complexity of providing accurate functional software size and effort prediction models is well known in the software industry. Function point analysis (FPA) is currently one of the most accepted software functional size metrics in the industry, but it is hardly automatable and generally requires a lengthy and costly process. Objectives: This paper reports on a family of replications carried out on a subset of the International Software Benchmarking Standards Group dataset (ISBSG R12) to evaluate the structure and applicability of function points. The goal of this replication is to aggregate evidence about internal issues of FPA as a metric, and to confirm previous results using a different set of data. Methods: A subset of 202 business application projects from 2005 to 2011 was analyzed. FPA counting was analyzed in order to determine the extent to which the basic functional components (BFC) were independent of each other and thus appropriate for an additive model of size. The correlations among effort and BFCs and unadjusted function points (UFP) were assessed in order to determine whether a simplified sizing metric might be appropriate to simplify effort prediction models. Prediction models were constructed and evaluated in terms of accuracy. Results: The results confirmed that some BFCs of the FPA method are correlated. There is a relationship between BFCs and effort. That suggest that prediction models based on transactional functions (TF) or external inputs (EI) appears to be as good as a model based on UFP in this subset of projects. Conclusions: The results might suggest an improvement in the performance of the measurement process. Simplifying the FPA measurement process based on counting a subset of BFCs could allow savings in measurement effort, preserving the accuracy of effort estimates.

Keywords Function point Analysis, software effort prediction, family of replications, empirical evaluation

Christian Quesada-López, Marcelo Jenkins. *Function Point Structure and Applicability: A Replicated Study*. In Journal of Object Technology, vol. 15, 2016, pages 2:1–26. doi:10.5381/jot.2016.15.3.a2

1 Introduction

Software estimation process is a key factor for software project success [PAP10]. The complexity to provide accurate software size estimation and effort prediction models in software industry is well known. The need for accurate size estimates and effort predictions for projects is one of the most important issues in the software industry [MJ03]. Inaccurate estimates are often the main cause of a great number of issues related to low quality and missed deadlines [Boe84] [MJ03]. Software size measurement and effort prediction models based on software size have been studied for many years, but many software companies are still using expert judgment as their preferred estimation method, producing inaccurate estimations and severe schedule overruns in many of their projects [Boe84] [MJ03]. The use effort estimation models properly is complex and time consuming, an organization must therefore decide whether it should prioritize its limited resources on training formal estimation methods [JBR09]. The development of a proper model may be too complex or take too much effort [Jør07].

Software size measurement is an important part of the software development process [LJ90] [GH01]. Functional size measures are used to measure the logical view of the software from the users' perspective by counting the amount of functionality to be delivered. These measures can be used for a variety of purposes, such as project estimation [LJ90] [GH01] [Kit95] quality assessment, benchmarking, and outsourcing contracts [GH01]. According to [ISO07], functional size measurements can be used for budgeting software development or maintenance, tracking the progress of a project, negotiating modifications to the scope of the software, determining the proportion of the functional requirements satisfied, estimating the total software asset of an organization, managing the productivity of software development, operation or maintenance and analyzing and monitoring software defect density. The use of functional size measures has been extensively discussed in the literature. These measures can be used for generating a variety of productivity, financial and quality indicators in different phases of the software development process [GH01]. Software size has proved to be one of the main effort-and-cost drivers [Boe84] [Alb79] [AG83] [JYW+11]. It is widely accepted that software size is one of the key factors that has the potential to affect the effort and cost of software projects [Boe84] [Kit95] [AG83] [Jon07] [Kem87].

FPA measurement is based on a set of basic functional components (BFC). But some studies suggest that BFCs have inter-correlations with each other. BFCs inter-correlation is likely to involve two problems. First, from a practical point of view, correlation between BFCs implies that some aspects are measured twice, which represents a waste of measurement effort. Second, from the theoretical point of view, measuring a BFC that is already measured by another BFC could affect the reliability of FPA measurement method [KK93] [LMR13]. Practitioners use the BFCs relations useful to predict FPA count from single elements without applying the entire method [Lok99].

This paper reports on a family of replications [Car10] based on [KK93] [JLB93] [JS96] [LMR13] [QLJ14] [QLJ15] and carried out on a subset of the ISBSG R12 dataset to evaluate the structure and applicability of function points. The importance of a family of replications is that all studies are related and investigate related questions in different contexts [Car10]. The aggregation of replication results will be useful for software engineers to draw conclusions and consolidate findings about similar research questions. This paper evaluates structure and applicability of function point analysis (FPA) as a measure of software size. First, we examined FPA counting in order to determine which base functional components (BFC) were independent of each

other and thus appropriate for an additive model of size. Second, we investigated the relationship between size and effort.

Although it is well known in the literature that there are many drivers for software effort and cost estimation, and that many factors can influence the prediction models, we decided to work with functional size as an effort driver in order to compare previous results and then, use other known effort drivers in an attempt to improve the prediction model accuracy. We analyzed software project estimations data in order to evaluate function point counting as a measure of software size. In this study we compared our results with [KK93] [JLB93] [JS96] [Lok99] [LMR13] [QLJ14]. Our goal was to aggregate evidence and to confirm previous results reported using a different dataset. The structure of this paper follows the reporting guidelines for experimental replications proposed by Carver [Car10]. The remainder of the paper is structured as follows. section 2 provides the foundations about function point analysis as a measure of software functional size. section 3 provides information on the original studies that is useful for understanding the replication. section 4 describes the current replication. section 5 compares the results of the replication and the original studies. section 6 discusses threats to validity. Finally, section 7 outlines conclusions and future work.

2 Function Point Analysis

Many functional size measurement (FSM) methods have been proposed to quantify the size of software based on functional user requirements (user perspective). Functional size is defined as "a size of the software derived by quantifying the functional user requirements" [ISO07]. Function point analysis (FPA) [Alb79] [AG83] [Jon13] was the first proposal for a FSM and it is one of the most used FSM methods in the industry [Jon13]. The International Function Point User Group Function Point Analysis (FPA) [ISO09], is a refinement of the very first method for functionality measurement proposed by Allan Albrecht [Alb79]. FSM methods analyze software requirements, transactions and data that are meaningful to the final user, which are identified and classified in a set of basic functional size components (BFC) and counted according to a defined complexity criteria. The BFC are mapped to numerical values and the sum of them constitute the functional size. In FPA the user requirements are identified, classified and counted in a set of basic functional size components (BFC). These BFC are data functions and transactional functions. They represent data and operations that are relevant to the final users. Data functions (DF) are classified into internal logic files (ILF) and external interface files (EIF). Transactional functions (TF) are classified into external inputs (EI), external outputs (EO), and external inquires (EQ). A function point analysis in FPA involves the identification of these five BFC types: EI, EO, EQ, ILF, and EIF. Each BFC contributes in the FPA counting that depends on its complexity. Complexity weight is calculated according to given tables. Unadjusted Function Points is obtained by the summing of all BFCs. As described in Figure 1, the method identifies data inputs and outputs that go inside or outside of the application boundaries. Details about FPA method can be found in FPA manual [ISO09].

FPA is independent from technology based influences, for example, the functional size must be the same regardless of the programing language [AG83]. Moreover, functional size does not depend on a specific notation of requirements specification [FB97]. FPA can be used to develop a measure of productivity [LJ90] [Jon07].

FPA have been subject to a number of critiques: the reliability of FPA measurement [LJ90], the BFCs have inter correlations with each other [Kit95] [Kem87] [JS96], the

application and usefulness of the complexity adjustments [Jon07], among others. FPA is prone to different interpretations by different subjects [TOOD08] [Jon13]. Variation in the counts is expected and finally, the counting method is slow and expensive [Jon13]. Since FPA, other FSM methods have been proposed. All of these methods have contributed towards the measurement of functional size, and all of them have issues that should be analyzed in order to create a reliable and consistent method [LMR13].

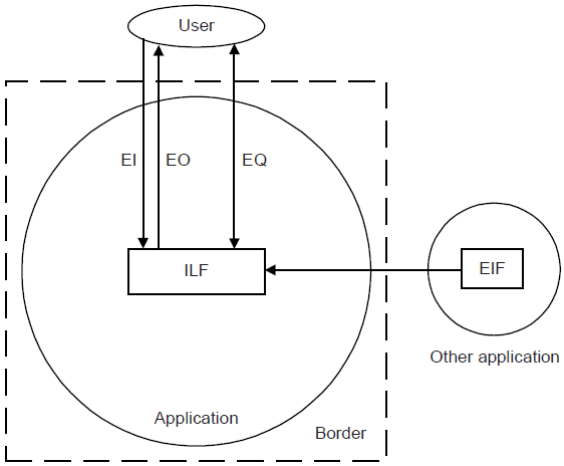


Figure 1 – Components of FPA [RBLB00].

3 Description of the Original Studies

The original studies [KK93] [JS96] [Lok99] [LMR13] [QLJ14] have evaluated the structure and applicability of function points as a measure of software size. Base functional components (BFC) inter-correlation implies that some aspects are measured twice and that some BFC are already measured by another BFC. The papers examined FPA counting in order to determine which BFCs were independent of each other and thus appropriate for an additive model of size and they investigated the relationship between functional size (UFP, AFP and BFC) and effort.

3.1 Goals and Research Question

Kitchenham and Kansala [KK93] analyzed the internal consistency of FPA and the use of FPA to predict effort. Jeffery, Low and Barnes [JLB93] investigated complexity adjustments in FPA and BFCs correlation. Jeffery and Stathis [JS96] empirically analyzed BFCs of unadjusted function count, and whether BFC size measures are statistically independent of each other and the relation between effort and BFC, UUFP, UFP and AFP. Lokan [Lok99] studied correlations between BFCs in FPA and analyzed how factors influenced the balance between BFCs. Quesada-López and Jenkins [QLJ14] [QLJ15], in previous studies, empirically investigated correlations between BFCs, UFP and effort. Lavazza, Morasca & Robiolo [LMR13] analyzed correlations between BFCs to evaluate the possibility of a simplified definition of function points. The goals and research questions from the original studies related with the replication are provided in Table 1.

Authors	Goals and Research Questions
Kitchenham & Kansala [KK93]	(1) To determine whether all the elements are required to provide a valid measure of size. (2) To determine whether all the sum of all the elements is a better predictor of effort than the constituent elements.
Jeffery & Stathis [JS96]	(1) To determine the extent to which the component elements of function points were independent of each other and thus appropriate for an additive model of size. (2) To investigate the relationship between effort and the function point components, and unadjusted function points; and (3) To determine whether the complexity weightings were adding to the effort explanation power of the metric.
Lokan [Lok99]	(1) To describe correlations between the FPA elements according to development type, language type, and program language.
Lavazza, Morasca & Robiolo [LMR13]	(1) To investigate whether it is possible to take into account only subsets of BFC as to obtain FSM that simplify FPA with the same effort estimation accuracy. They analyzed correlations between UFP and BFCs and effort and BFC.
Quesada-López & Jenkins [QLJ14] [QLJ15]	(1) To examine FPA counting in order to determine which BFC are independent from each other and thus appropriate for an additive model of size. (2) To investigate the relationship between size UFP, BFC and effort.

Table 1 – Goals and Research Questions.

3.2 Context and Variables

The original studies were run based on real project datasets from distinct software development organizations where the main types of applications were in the MIS domain. Table 2 shows relevant information about previous studies. Information about the dataset and the context of the data are mentioned. Table 3 summarizes the independent and dependent variables analyzed in the empirical analysis, taken directly from the datasets.

3.3 Summary of Analysis Techniques used in the Original Studies

Several number of effort estimation techniques have been previously applied and tested in literature, such as ordinary least squares regression (OLS), and stepwise regression (SLR) with or without transformation. According to [DVMB12] OLS in combination with a logarithmic transformation performs best. Moreover, when the assumptions for OLS linear regression are not satisfied, least squares regression (LSR) [SBJ13] and least median of squares (LMS) [LMR13] have been successfully used. Correlation between variables have been checked using parametric test such as Pearson's correlation and non-parametric tests such as Kendall's and Spearman's correlation. For the Pearson r correlation, both variables should be normally distributed. Pearson assumes that data is normally distributed about the regression line. Kendall rank correlation test the strength of dependence between two variables. Spearman rank correlation test does not make any assumptions about the distribution. The presence of correlation between

Authors	Dataset	Type	Domain
Kitchenham & Kansala [KK93]	40 projects from 9 software development organizations	Cross company	MIS
Jeffery, Low & Barnes [JLB93]	64 projects from 1 software development organization	Within-company	MIS
Jeffery & Stathis [JS96]	17 projects from 1 software development organization Within- company MIS		
Lokan [Lok99]	269 projects from the ISBSG R4 dataset	Cross company	MIS, DSS
Lavazza, Morasca & Robiolo [LMR13]	Over 600 projects from the ISBSG R11 dataset	Cross company	MIS
Quesada-López & Jenkins [QLJ14]	14 projects from the ISBSG R4 dataset	Cross company	MIS
Quesada-López & Jenkins [QLJ15]	72 projects from the ISBSG R12 dataset	Cross company	MIS

Table 2 – Information about Original Studies.

Independent		Dependent
Global	Specific	
BFC Size (UUFP and UFP)	Input count, Output count, Interface count, File count, Enquiry count	Work Effort
UUFP Size	Unadjusted and unweighted Functional size	
UFP Size	Unadjusted Functional size	
AFP Size	Adjusted Functional size	
Context	Development type, Language and Language type, Application group, Team Size	

Table 3 – Independent and Dependent Variables.

variables does not necessarily imply accurate prediction models [LMR13]. Regarding the assessment of the accuracy of prediction models, the common indicators used in software engineering are mean magnitude of relative error (MMRE) and prediction quality indicator Pred (25) [MA08b]. However, the usefulness of these indicators has been criticized [KPMS01], and other indicators have been proposed, such as balance relative error (BRE) [MTON94] [SBJ13] [MA08b]. Table 4 summarizes the applied modeling techniques and the evaluation techniques for each study.

3.4 Summary of Results

Kitchenham and Kansala [KK93] reported correlations among BFC size measures. BFC were not independent. They observed that FP does not have the characteristics of a valid additive size metric, because some elements seem to be counted more than once. Not all BFC were related to effort, an effort prediction model based on some

Authors	Techniques	Evaluation
Kitchenham & Kansala [KK93]	Simple linear regression, stepwise multivariate regression.	R^2, Kendall's correlation, Pearson correlation.
Jeffery & Stathis [JS96]	Simple linear regression.	R^2, Pearson correlation, Kendall's correlation.
Lokan [Lok99]	Simple linear regression.	Kendall's correlation.
Lavazza, Morasca & Robiolo [LMR13]	Ordinary Least Square Regression, Log transformation, Least Median of Squares Regression, analogy criteria.	Cook's distance, Kendall's correlation, Spearman's correlation, R^2, $MMRE$, $PRED(25)$.
Quesada-López & Jenkins [QLJ14]	Simple linear regression.	R^2, Pearson correlation, Kendall's correlation.
Quesada-López & Jenkins [QLJ15]	Stepwise regression.	R^2, Pearson correlation, Kendall's correlation.

Table 4 – Summary of Techniques.

BFC (EI and EO) was just as good as total FP. They expect that simpler counting would reduce the variability of the counting results because some BFC were as good at predicting effort as UFP. Jeffery, Low and Barnes [JLB93] also found that BFC are not independent. Furthermore, they concluded that processing complexity adjustment had no effect on the accuracy of the effort models. Jeffery and Stathis [JS96] found significant correlations between UFP and EI, EQ, ILF, and between BFC and effort. Also, they determine that the adjusted values in the counting did not improve the power of the measure and the effort prediction models. They also suggested that a simplified sizing metric may be appropriate. Lokan [Lok99] reported evidence of BFC inter-correlation as well after completing an experiment involving data from 269 projects where EI and ILF were correlated and EIF were rarely correlated to other BFCs. He confirmed previous results that some BFCs are counted more than once. He determined that specific context factors such as type of development and language type influence the balance between BFCs. Lavazza, Morasca & Robiolo [LMR13] determine correlations between BFCs and assess encouraging effort prediction models based on a simplified count. Quesada-López and Jenkins [QLJ14] found correlations between UFP and EI, EQ, ILF, and between BFC and effort. Besides, they found correlations between BFCs EI and EO, EQ and EQ and ILF. Finally, correlation between some BFCs and effort were found. In [QLJ15], the authors found that most of the BFCs appear to be correlated with UFP, these components are not independent because there are correlations between EI and EQ, EI and ILF, and EQ and ILF. Besides, BFCs are significantly correlated with effort, EI (0.531), ILF (0.588) and EQ (0.861) presented similar correlations as UFP (0.785).

The results showed that BFCs size measures were actually correlated between them and with effort. This suggests that a simplified form of function point sizing method (i.e. based on data) would be possible across different domains. It is expected that simpler counting would reduce the variability of the counting results. Several studies have explored the possibility of a simplified function point method. Lavazza et al. [LMR13] proposed a simplified definition of FP using only subsets of BFCs. The findings in

[QLJ15] confirmed this possibility, the analysis indicates that a prediction model based on TF or EI, EO and ILF appear to be as good as UFP. Moreover, the use of some context attributes in prediction models such as language type, language, platform, architecture and team size may improve the results. Other studies have proposed simplified definitions for FPA, as an example, Symons [Sym88] based Mark II on the basis of three BFC, Early & Quick Function Points (EQFP) [SCM05] measurement process leads to an approximate measure of size in IFPUG FP. An advantage of the method is that different parts of the system can be measured at different levels of detail. NESMA [ISO05] simplifies the process of counting function points by only requiring the identification of logic data from a data model. NESMA provides ways to estimate size in FPA based only on data functions. The function point size is then computed by applying predefined weights.

4 Replication

4.1 Motivation

Combined results from a family of replications are interesting because all studies are related and investigate related questions in different contexts. The aggregation of replication results will be useful for software engineers to draw conclusions and consolidate findings about similar research questions [Car10]. In this study, we compare results with [KK93] [JLB93] [JS96] [Lok99] [LMR13] [QLJ14] [QLJ15]. Correlations between the BFCs have been found in previous studies but their findings were different in some respects, but not in others (See section 5). Further research is needed to understand the relationships between BFCs. By replicating, with a different dataset, selected with specific characteristics, a better understanding about previous agreement and disagreement results is reached [Lok99]. The goal of this replication was to aggregate evidence about internal issues of FPA as a metric, and to confirm previous results reported using a different set of data. Moreover, an empirical evaluation was conducted to assess the accuracy of effort prediction models based on some basic functional components (BFC).

4.2 Level of Interaction with the Original Investigators

The authors of the original study did not take part in the replication process. Current replication is external [SCVJ08] regarding to [KK93] [JLB93] [JS96] [Lok99] [LMR13].

4.3 Changes to the Original Study

This section describes how the replication experiment changed. This study was designed to respect most of the analysis of the original experiments in order to assure that the results would be comparable. Three types of changes were made intentionally: the context and dataset selection, the independent variable selection, and some of the statistical techniques. This paper empirically evaluates the accuracy of effort prediction models based on relative error indicators. The analysis presented in this paper is based on a sample of software projects from the ISBSG R12 dataset. The ISBSG repository provides organizations with a broad range of project data from various industries and business areas [Hil11]. The data can be used for effort estimation, trend analysis, comparison of platforms and languages, and productivity benchmarking [MLHT05]. The ISBSG repository is a multi-organizational, multi-application, and

multi-environment data repository [CA13]. However, the ISBSG repository is a large heterogeneous dataset and suffers from missing data. A detailed data preparation process is required to obtain the appropriate subset for analysis that can be applied for organization [Hil11]. Different versions and subsets of this database have been used for many studies, such as [Lok99] [LMR13] [QLJ14] [DVMB12] [SBJ13] [QLJ15] [MA08a] [JRW00] [LM06] [ML08].

4.3.1 Overview of the Empirical Study

The overall process consisted of the following steps:

- Data selection and preprocessing: the original dataset was preprocessed to obtain the appropriate data points and variables for the experiment.
- UFP and BFCs evaluation: unadjusted function points and basic functional components are examined to determine which of these BFCs are independent from each other and thus appropriate for an additive model of size. Relationship between size in UFP and BFCs, and effort were evaluated.
- Estimation model evaluation: prediction models based on UFP and BFCs were evaluate to determine the accuracy of the estimates.

4.3.2 Data Set Selection

The subset of data projects for our study was selected from the ISBSG R12 database according to the criteria shown in Table 5 based on recommendations in [QLJ14] [MLHT05] [DVMB12] [SBJ13] [QLJ15] [LM06] [ML08]. Projects for which all functional components (UFP and BFC) of function points were missing were discarded. For our study, we selected the variables related with FPA functional size components (BFC), effort of software development, and context attributes according to recommendation in [GLdGFD14]. The list of selected variables is shown in Table 6.

As a result of the selection, a total of 204 project data were identified. Outliers were evaluated using a Cook's distance test [Coo77]. After the analysis, 2 outliers identified were removed because this reduction of the dataset improved the fit of the model. This study was conservative to remove the outliers from the data set to prevent the possibility of extreme values distorting the derived regression equations for effort models. The outliers were large projects (1,659 and 2,250 UFP) with abnormal productivity values (1.5 and 13.7 h/UFP). In total 202 projects were included in our analysis. Forty three of them are from 2005, 34 from 2006, 46 from 2007, 35 from 2008, 26 from 2009, 13 from 2006, and 5 from 2011. Table 7 summarizes descriptive statistics for the ratio-scale variables: size in unadjusted function points (UFP), effort in person hours (Effort), productivity on a simple ratio of product size to project effort (PDR), external inputs (EI), external outputs (EO), external inquires (EQ), internal logic files (ILF), external interface files (EIF), data functions (DF), and Transactional functions (TF). Statistics were selected according measures' scale to describe the data [WRH+12]. The smallest project size is 6 UFP, the average is 247 with a median of 184 UFP, and the largest project is 1,337 UFP. The average productivity for the dataset is 21 hours per UFP, with a median of 16 hours per UFP. The data set is positively skewed for all variables indicating the quantity of small and medium projects is higher than the number of large projects. Table 8 summarizes descriptive statistics grouped for the nominal-scale variables. This table details number of projects, size (UFP), and productivity (PDR=UFP/Effort) for categorical variables. In each case percentage (%)

Criteria	Value	Motivation
Count Approach	IFPUG 4+	Latest FPA standard and counting rules.
Data Quality Rating	A	Only data with a high level of quality and integrity.
Unadjusted Function Point Rating	A or B	Counting data with a high level of quality and integrity. Most studies using ISBSG selected 'A' or 'B' ranking.
Year of project	>= 2005	New projects using new technologies.
Application group	BA	Business Application is one of the mayor development area in the industry.
Resource Level	1	Only development team effort included.

Table 5 – Project Selection Criteria.

related to the number of projects and functional size (UFP) by categorical attribute is presented. Most of the projects were small (between 32-99 UFP) and medium size (between 102-993 UFP). Extra small projects (between 6-28 UFP) and large projects (between 1,018-1,337 UFP) show lower productivity than small and medium size projects. Data indicate that use small teams in small and medium size projects are better in terms of productivity. Nominal variables (context) influence effort and productivity. Similar influences were reported in previous studies and observed in [QLJ15].

5 Comparison and Discussion of Results

5.1 Data Analysis

Scatter plot of actual work against UFP for the dataset (202 projects) shows evidence that there is a positive relationship between effort and UFP (Figure 2). The simple linear regression equation for the selected dataset was $effort = -1242.842 + 24.177 * UFP$, ($R^2 = 0.47, p < 0.000$). Assumptions were checked and residuals were normally distributed. All models presented in this paper were built and tested using the statistical tools R [Tea05], SPSS v21, and WEKA [HFH+09]. Statistical significance level was set at 0.05, unless otherwise stated. A comparison of simple linear regression results against previous studies is shown in Table 9. This data shows the sensibility of the results depending of the data selection. As in [Boe84], with a nonlinear regression model with the equation $effort = 48.395 * UFP^{0.782}$ the model gave a better fit for the dataset ($R^2 = 0.53, p < 0.000$). Although low regression values ($R^2 = 0.47, R^2 = 0.53$) were found for this dataset, similar correlation coefficients have been reported in the previous studies that studied effort prediction models.

5.2 Internal Consistency of Function Points

Table 10 shows the Kendalls's Tau correlation coefficients between all pairs of funcion point BFCs using the entire dataset (202 projects). Previous study results are also presented in Table 10 for comparison. Outliers were removed from datasets in [JS96] [Lok99] [QLJ14] [QLJ15]. The results showed that BFCs are not independent. Jeffery & Stathis [JS96] reports similarities and differences in results with [KK93] [LMR13]

Variable	Scale	Description
Input count	Ratio	Unadjusted function points (UFP) of External Input (EI).
Output count	Ratio	UFP of External Output (EO).
Interface count	Ratio	UFP of External Interface (EIF).
File count	Ratio	UFP of Internal Logical Files (ILF).
Enquiry count	Ratio	UFP of External Enquiry (EQ).
Functional size	Ratio	Application size in Unadjusted Function Point count (UFP).
Normalized Level 1 Work Effort	Ratio	The development team full life-cycle effort in person hours recorded against the project.
Normalized Level 1 Productivity Delivery Rate	Ratio	Productivity delivery rate in hours per functional size unit (UFP).
Development Type	Nominal	Enhancement, New Development, Re-development.
Relative Size	Ordinal	1. XXS, 2. XS, 3. S, 4. M1, 5. M2, 6. L.
Team Size Group	Ordinal	2, 3-4, 5-8, 9-14, 15-20, 21-30, 31-40, 61-70.
Development Platform	Nominal	Multi, MF, PC, MR.
Architecture	Nominal	Client server, Stand alone, Multi-tier with web public interface, Multi-tier, Stand-alone.
Language Type	Nominal	3GL, ApG, 4GL.
Program Language	Nominal	Java, COOL:Gen, ASP.Net, C#, JavaScript, ABAP, PL/I, Visual Basic, PowerBuilder, ASP, SQL, Visual Studio .Net, Datastage, .Net, IBM WTX, XML, COBOL, AB INITIO, A:G, C++.

Table 6 – ISBSG Dataset Variables used in this Study.

Variable	Min	First Q.	Median	Third Q.	Max	Mean	G. Mean	St. Dev.	Kurtosis	Skewness
UFP	6	90	184	312	1,337	247	170	224	4.67	2.00
Effort	167	1,303	2,752	4,979	71,729	4,727	2,683	7,921	40.35	5.74
PDR	2	10	16	24	191	21	16	20	29.84	4.50
EI	0	27	66	131	551	97		100	3.91	1.90
EO	0	5	21	49	287	38		50	8.20	2.64
EQ	0	14	43	90	419	65		71	5.02	1.99
ILF	0	7	21	51	403	37		53	18.31	3.65
EIF	0	0	0	10	261	10		27	40.83	5.43
DF	6	73	152	260	1,075	200		182	4.57	2.00
TF	0	7	27	59	606	47		69	25.66	4.18

Table 7 – ISBSG Sub Dataset Summary of Ratio-Scale Variables (202 projects).

Nominal	Ratio	Level	N	%	Min	First Q.	Median	Third Q.	Max	Mean	G. Mean	St. Dev.
Dev.	UFP	E	161	0.8	6	84	154	270	960	210	148	180
Type		N	39	0.19	24	231	303	421	1,337	394	291	311
		R	2	0.01	112	112	364	615	615	364	262	356
	PDR	E	161	0.8	2	10	14	24	191	20	15	21
		N	39	0.19	7	13	18	27	73	24	20	16
		R	2	0.01	5	5	12	19	19	12	9	10
DevType: Development Type (E=Enhancement, N=New Development, R= Re-development)												
Band	UFP	XXS	2	0.01	6	6	7	8	8	7	7	1
Size		XS	5	0.02	13	14	16	24	28	19	18	7
		S	49	0.24	32	53	72	84	99	68	66	19
		M1	92	0.46	102	140	185	241	299	192	183	59
		M2	51	0.25	303	336	411	624	993	497	465	196
		L	3	0.01	1,018	1,033	1,048	1,193	1,337	1,134	1,126	176
	PDR	XXS	2	0.01	137	137	164	191	191	164	162	38
		XS	5	0.02	7	20	39	41	106	43	30	38
		S	49	0.24	3	11	18	24	73	21	18	14
		M1	92	0.46	4	9	16	23	48	18	15	10
		M2	51	0.25	2	9	13	22	59	17	13	13
		L	3	0.01	7	31	54	57	60	40	29	29
Band Size: Relative Band Size (1.XXS, 2.XS, 3.S, 4.M1, 5.M2, 6.L)												
Lang.	UFP	3GL	112	0.55	8	103	199	327	1,337	266	185	241
Type		4GL	36	0.18	59	165	232	389	1,048	315	250	240
		ApG	52	0.26	6	54	99	238	635	155	104	135
		ND	2	0.01	349	349	361	372	372	361	360	16
	PDR	3GL	112	0.55	2	11	17	24	137	21	16	17
		4GL	36	0.18	4	8	12	21	59	16	13	13
		ApG	52	0.26	5	10	16	26	191	24	17	29
		ND	2	0.01	8	8	8	9	9	8	8	1
Lang. Type: Language Type (3GL, ApG, 4GL, ND=Not defined)												
Team	UFP	2	1	0	41		41		41	41	41	
Size		3-4	18	0.09	24	90	138	212	297	149	124	80
		5-8	56	0.28	51	93	162	274	1,048	236	177	214
		9-14	36	0.18	89	182	287	395	923	309	265	182
		15-20	16	0.08	121	280	336	676	993	480	412	274
		21-30	3	0.01	298	330	361	528	695	451	421	213
		31-40	2	0.01	381	381	859	1,337	1,337	859	714	676
		61-70	1	0	153		153		153	153	153	
		ND	69	0.34	6	56	103	237	1,018	173	108	177
	PDR	2	1	0	9		9		9	9	9	
		3-4	18	0.09	4	6	9	13	23	10	9	6
		5-8	56	0.28	2	8	11	18	73	14	11	11
		9-14	36	0.18	3	14	18	27	49	22	19	11
		15-20	16	0.08	5	15	21	25	59	23	19	13
		21-30	3	0.01	28	32	36	41	45	37	36	9
		31-40	2	0.01	54	54	56	59	59	56	56	4
		61-70	1	0	41		41		41	41	41	
		ND	69	0.34	5	11	19	27	191	26	19	29
Team Size: Team Size Group (2, 3-4, 5-8, 9-14, 15-20, 21-30, 31-40, 61-70, ND=Not defined)												

Table 8 – ISBSG Sub Dataset Nominal Variables Characteristics (202 projects).

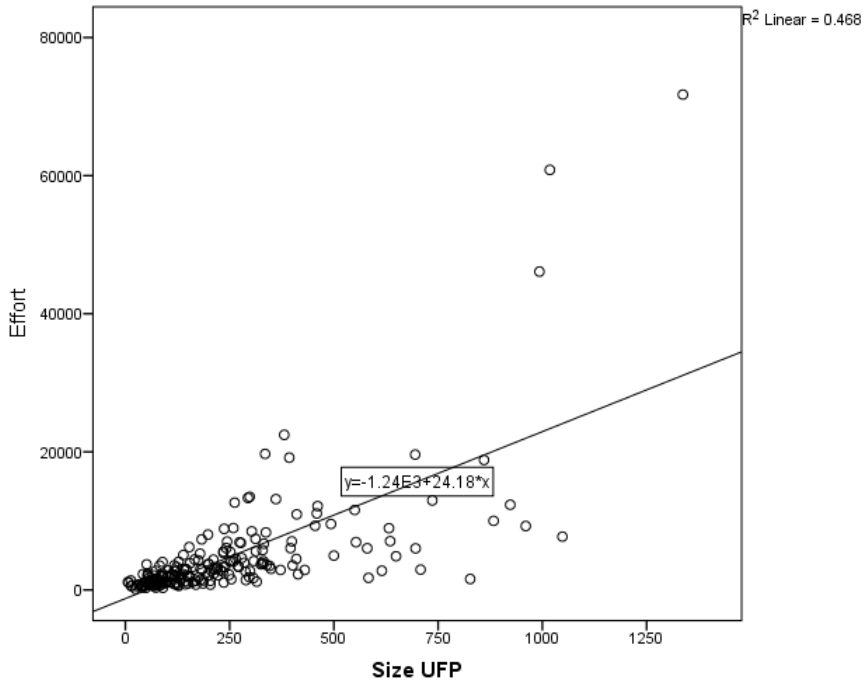


Figure 2 – Actual work against UFP (202 projects).

[Lok99] [QLJ14]. These studies found correlations in EO and EI, EO and EQ, EO and EIF, and EO and ILF not presented in [JS96]. Jeffery & Stathis [JS96] confirms [KK93] results in EI and EQ, EI and ILF, and EQ and ILF. These correlations are presented also in [Lok99] [QLJ14] and the current study. The results in the current study confirms correlations reported in all the studies between EI and EQ, and EQ and ILF as is presented in Table 10. Our results are consistent with authors regarding that differences could be caused by the nature of projects data (application types, design techniques, programming languages, and other causes). Regarding the correlation between UFP and BFCs, the results in all studies show that EI, EQ and ILF elements are significantly correlated with UFP, and this study shows that UFP are significantly correlated with all BFCs. Lavazza, Morasca & Robiolo [LMR13] reported that transaction functions (TF) is extremely correlated with UFP. Besides, there is also a correlation between TF and data functions (DF). In our study, we also find significant correlation between TF and UFP (Kendall's $s = 0.868, p < 0.000$), and a relation between TF and DF (Kendall's $s = 0.407, p < 0.000$).

5.3 Using UFP and BFCs to predict effort

Table 9 shows evidence of correlations between UFP and effort and Table 10 shows evidence of correlations between BFCs and UFP. The question to investigate is whether a better size/effort model exists instead of the sum of the BFCs. Table 11 shows that some BFCs are significantly correlated with effort. For the dataset in this study, EI, ILF and EQ presented similar correlations as UFP as in [QLJ15]. In addition, the results

Study	Projects	R^2	(p)
Albrecht, Gaffney [AG83]	24	0.90	< 0.001
Kemerer [Kem87]	15	0.54	< 0.001
Kitchenham, Kansala [KK93]	40	0.41	< 0.010
Jeffery, Low & Barnes [JLB93]	64	0.36	< 0.001
Jeffery & Stathis [JS96]	17	0.95	< 0.001
Jeffery & Stathis [JS96]	14	0.58	< 0.001
Quesada-López & Jenkins [QLJ14]	14	0.94	< 0.000
Quesada-López & Jenkins [QLJ14]	12	0.62	< 0.003
Quesada-López & Jenkins [QLJ15]	72	0.68	< 0.000
This Study	202	0.47	< 0.000

Table 9 – Previous Studies Comparison – UFP against Effort.

Study	BFC	UFP	EI	EO	EQ	EIF
[KK93]	EI	**0.67** $p < 0.001$				
[JS96]		**0.54** $p < 0.01$				
[Lok99]		(*n.r.*)				
[LMR13]		**0.658** (*n.r.*)				
[QLJ14]		**0.74** $p < 0.00$				
[QLJ15]		**0.64** $p < 0.00$				
This Study		**0.65** $p < 0.000$				
[KK93]	EO	0.53 $p < 0.001$	**0.47** $p < 0.001$			
[JS96]		0.27 (*n.s.*)	0.03 (*n.s.*)			
[Lok99]		(*n.r.*)	**0.37** $p < 0.001$			
[LMR13]		**0.597** (*n.r.*)	**0.438** (*n.r.*)			
[QLJ14]		**0.45** $p < 0.04$	**0.55** $p < 0.01$			
[QLJ15]		0.34 $p < 0.00$	0.19 $p < 0.19$			
This Study		**0.36** $p < 0.000$	**0.247** $p < 0.00$			
[KK93]	EQ	**0.47** $p < 0.001$	**0.47** $p < 0.001$	**0.32** $p < 0.01$		
[JS96]		**0.68** $p < 0.001$	**0.72** $p < 0.001$	-0.06 (*n.s.*)		
[Lok99]		(*n.r.*)	**0.48** $p < 0.001$	**0.29** $p < 0.001$		
[LMR13]		**0.528** (*n.r.*)	**0.448** (*n.r.*)	**0.288** (*n.r.*)		
[QLJ14]		**0.80** $p < 0.00$	**0.61** $p < 0.00$	**0.25** $p < 0.27$		
[QLJ15]		**0.54** $p < 0.00$	**0.38** $p < 0.00$	0.03 $p < 0.66$		
This Study		**0.575** $p < 0.000$	**0.384** $p < 0.00$	0.086 $p < 0.76$		
[KK93]	EIF	0.32 $p < 0.01$	0.14 (*n.s.*)	**0.31** $p < 0.01$	0.60 (*n.s.*)	
[JS96]		-0.37 (*n.s.*)	-0.56 $p < 0.05$	0.03 (*n.s.*)	-0.53 $p < 0.05$	
[Lok99]		(*n.r.*)	-0.02 (*n.s.*)	0.10 (*n.s.*)	0.00 (*n.s.*)	
[LMR13]		0.264 (*n.r.*)	0.072 (*n.r.*)	0.194 (*n.r.*)	0.097 (*n.r.*)	
[QLJ14]		0.42 $p < 0.07$	0.16 $p < 0.50$	0.00 $p < 1.00$	0.41 $p < 0.08$	
[QLJ15]		-0.04 $p < 0.69$	-0.15 $p < 0.11$	-0.27 $p < 0.77$	-0.02 $p < 0.80$	
This Study		**0.572** $p < 0.000$	**0.387** $p < 0.00$	**0.183** $p < 0.00$	**0.427** $p < 0.00$	
[KK93]	ILF	**0.60** $p < 0.001$	**0.51** $p < 0.001$	**0.30** $p < 0.01$	**0.31** $p < 0.01$	0.17 (*n.s.*)
[JS96]		**0.73** $p < 0.001$	**0.44** $p < 0.05$	0.11 (*n.s.*)	**0.65** $p < 0.001$	-0.39 (*n.s.*)
[Lok99]		(*n.r.*)	**0.48** $p < 0.001$	**0.33** $p < 0.001$	**0.41** $p < 0.001$	0.08 $p < 0.02$
[LMR13]		**0.619** (*n.r.*)	**0.449** (*n.r.*)	**0.417** (*n.r.*)	**0.327** (*n.r.*)	0.195 (*n.r.*)
[QLJ14]		**0.66** $p < 0.00$	**0.44** $p < 0.05$	0.19 $p < 0.40$	**0.51** $p < 0.02$	0.56 $p < 0.02$
[QLJ15]		**0.58** $p < 0.00$	**0.38** $p < 0.00$	0.11 $p < 0.21$	**0.41** $p < 0.00$	0.60 $p < 0.52$
This Study		**0.225** $p < 0.000$	0.04 $p < 0.478$	0.104 $p < 0.06$	**0.15** $p < 0.005$	**0.24** $p < 0.00$

Kitchenham & Kansala [KK93], Jeffery & Stathis [JS96], Lokan [Lok99],
Lavazza, Morasca & Robiolo [LMR13], Quesada-López & Jenkins [QLJ14] [QLJ15].
(n.s.) not significant. (n.r.) not reported.

Table 10 – Kendall Tau Correlation Coefficients Comparison between BFCs.

of this study, showed that TF (Kendall's $s = 0.520$) and DF (Kendall's $s = 0.452$) presented similar correlations as UFP. These results support the findings of previous studies where ILF and EQ have correlation with effort. The results provide additional evidence to suggest that some subset of FPA UFP base functional components (BFC) could offer an effort prediction models at least as good as the sum of all the BFCs. For example, Kitchenham & Kansala [KK93] found that a combination of EI and EO offers better correlation with effort than UPF. Lavazza, Morasca & Robiolo [LMR13] reported that a prediction model based on EI, EO and transactional function (TF) were as good as a model based on UFP.

Effort models were built using simple and stepwise regression techniques. First, we used only size measures (UFP and BFCs) and later nominal context variables were added to evaluate the improvement of the prediction models. Since over fitting is a concern of our study, we decide to only include the variables in the model that explain variance. To achieve this, we applied stepwise regression and part of the principal components analysis (PCA) that are variable reduction techniques to reduce a larger set of variables into a smaller set of variables which account for most of the variance in the original variables [Rid02].

Based on the list of communalities for each variable provided in the PCA analysis, the stronger variables were selected. Communalities represent the proportion of variability for a given variable that is explained by the factors and allows to examine how individual variables reflects the sources of variability. The values represent the proportion of each variable's variance that can be explained by the principal components. Variables with high values are well represented in the common factor space. In our dataset, best values for each variable were: TF (0.990), DF (0.963), EI (0.881), and ILF (0.792). These variables were used to find the best prediction models because our aim is to try to simplify the initial data collection of functional size components.

In the dataset, nominal variables were transformed by dummy coding where each variable was coded 0 and 1. For example, DevType had 3 levels, each was replaced by one dummy variable. The final set of variables used for the data set is presented in Table 12. Table 13 showed effort models based on UPF, BFCs and nominal variables. For each regression model, residuals versus fitted values were normally distributed and outliers were removed according recommendations in [KM09].

Results showed a relation between UFP and effort, and BFCs and effort. The results from this study support the findings of the previous studies. There is evidence to suggest that a subset of BFCs may offer an effort prediction model at least as good as UFP. It is known that context attributes such as development type, language type, language, platform, architecture, and team size affect effort prediction models [DVMB12]. Results shows that the use of these context attributes in prediction models may improve the model fitness. The prediction accuracy of models was tested on the raw data and the statistics used in [LMR13] [LM06] [SBJ13] were applied: magnitude of relative error (MRE), magnitude of error relative (MER), balanced relative error (BRE), and number of predictions within % of the actuals (Pred (25)). Pred is simply the percentage of estimates that are within m% of the actual value (the % of the estimates with MRE $<= 0.25$). Typically m is set to 25 so the indicator reveals what proportion of estimates are within a tolerance of 25%. This evaluation is conducted because the presence of a correlation does not necessarily imply that an accurate predictive model can be built. These accuracy indicators are defined as follows:

$$MRE_i = \frac{|ActualEffort_i - EstimatedEffort_i|}{ActualEffort_i}$$

$$MER_i = \frac{|ActualEffort_i - EstimatedEffort_i|}{EstimatedEffort_i}$$

$$BRE_i = \frac{|ActualEffort_i - EstimatedEffort_i|}{\min(ActualEffort_i, EstimatedEffort_i)}$$

$$Pred(n) = \tfrac{100}{N} \sum_{i=1}^{N} \begin{cases} 1 & if MRE_i \leq n/100 \\ 0 & otherwise \end{cases}$$

The regression models' accuracy results are summarized in Table 14. To compare the models and accuracy of models based on UFP and BFCs, we studied residuals between actual values and estimated values. To evaluate the equivalence of models, paired t-test, Wilcoxon signed rank test, and Mann-Whitney had been applied on absolute residuals [LMR13] [KPME02], and MRE values[SM98]. In our case, we use the non-parametric test Wilcoxon signed rank, and Mann-Whitney because MRE, MER, and BRE results were not normally distributed. We applied the Wilcoxon signed rank test and the Mann-Whitney test to the distributions of MRE, MER, and BRE results. We test the following null hypotheses to evaluate the prediction models shown in Table 13 to determine if the accuracy of the estimates for a model based on UFP and a model based on BFCs is the same:

$$H_0 : \textit{There is no difference in accuracy between } Model_i \textit{ and } Model_j$$

The Wilcoxon signed rank test indicated for each comparison except (UFP versus EI and ILF), we cannot reject the null hypothesis that the models based on BFCs are equivalent to the model based on UFP in terms of MRE, MRE, and BRE. Results were shown in Table 15.

In practice, results indicate that using effort models based on BFCs instead of UFP does not cause the accuracy of the estimates decrease. These results confirms previous results in [LMR13]. Regarding the prediction accuracy, it is difficult to compare results across different studies due to differences in empirical setup and data preprocessing, but a typical Pred(25) lies in the range of 10 to 60 percent, while the MdMRE typically attains values between 30 and 100 percent [DVMB12]. In our study, the models are short of the typical industry target of MMRE=25% and PRED(25)=75%, but this results are similar to reported in previous studies with this dataset [LMR13] [LM06] [DVMB12]. However, based on these results, future studies evaluating and improving accuracy of prediction models based on function points could consider the evaluation of models based on a subset of basic functional components (BFC).

6 Threats to Validity

This section analyses the threats to the validity for this study and the actions undertaken to mitigate them. There are several threats to the validity of this work.

Study	BFC	Pearson	Kendall Tau	Spearman
[KK93]	UFP	0.65 $p < 0.001$	$(n.r.)$	$(n.r.)$
[JS96]		0.58 $p < 0.01$	$(n.r.)$	$(n.r.)$
[QLJ14]		0.785 $p < 0.003$	$(n.r.)$	$(n.r.)$
[QLJ15]		0.825 < 0.000	0.607 $p < 0.000$	0.793 $p < 0.000$
This Study		0.684 $p < 0.000$	0.550 $p < 0.000$	0.749 $p < 0.000$
[KK93]	EI	0.60 $p < 0.001$	$(n.r.)$	$(n.r.)$
[JS96]		0.37 $p < 0.001$	$(n.r.)$	$(n.r.)$
[QLJ14]		0.531 $p < 0.076$	$(n.r.)$	$(n.r.)$
[QLJ15]		0.720 $p < 0.000$	0.484 $p < 0.000$	0.667 $p < 0.000$
This Study		0.582 $p < 0.000$	0.417 $p < 0.000$	0.590 $p < 0.000$
[KK93]	ILF	0.44 $p < 0.01$	$(n.r.)$	$(n.r.)$
[JS96]		0.73 $p < 0.001$	$(n.r.)$	$(n.r.)$
[QLJ14]		0.588 $p < 0.05$	$(n.r.)$	$(n.r.)$
[QLJ15]		0.622 $p < 0.000$	0.456 $p < 0.000$	0.613 $p < 0.000$
This Study		0.530 $p < 0.000$	0.443 $p < 0.000$	0.612 $p < 0.000$
[KK93]	EQ	0.28 $(n.s)$	$(n.r.)$	$(n.r.)$
[JS96]		0.63 $p < 0.001$	$(n.r.)$	$(n.r.)$
[QLJ14]		0.861 $p < 0.001$	$(n.r.)$	$(n.r.)$
[QLJ15]		0.596 $p < 0.000$	0.416 $p < 0.000$	0.561 $p < 0.000$
This Study		0.516 $p < 0.000$	0.400 $p < 0.000$	0.552 $p < 0.000$
[KK93]	EO	0.66 $p < 0.001$	$(n.r.)$	$(n.r.)$
[JS96]		0.03 $(n.s)$	$(n.r.)$	$(n.r.)$
[QLJ14]		0.277 $p < 0.383$	$(n.r.)$	$(n.r.)$
[QLJ15]		0.525 $p < 0.000$	0.320 $p < 0.000$	0.431 $p < 0.000$
This Study		0.469 $p < 0.000$	0.291 $p < 0.000$	0.406 $p < 0.000$
[KK93]	EIF	0.31 $(n.s)$	$(n.r.)$	$(n.r.)$
[JS96]		0.005 $(n.s)$	$(n.r.)$	$(n.r.)$
[QLJ14]		0.857 $p < 0.00$	$(n.r.)$	$(n.r.)$
[QLJ15]		0.233 $p < 0.049$	0.040 $p < 0.659$	0.057 $p < 0.632$
This Study		0.254 $p < 0.000$	0.262 $p < 0.000$	0.343 $p < 0.000$

Kitchenham & Kansala [KK93], Jeffery & Stathis [JS96].
Quesada-López & Jenkins [QLJ14] [QLJ15].
(n.s.) not significant. (n.r.) not reported.

Table 11 – Correlation Coefficients between UFP, BFCs and Effort.

Variable	Meaning
UFP	Size in unadjusted function points
EI	Size of external inputs
EO	Size of external outputs
EQ	Size of external inquires
ILF	Size of internal logic files
EIF	Size of external interface files
TF	Size of Transactional functions, TF = (EI+EO+EQ)
DF	Size of data functions, DF = (ILF+EIF)
Effort	Effort in person hours
Dev. Type	Development Type with 3 levels (E=Enhancement, N=New Development, R= Re-development)
Lang. Type	Language Type with 4 levels (3GL, ApG, 4GL, ND=Not defined)
Language	Programming Language with 21 levels (Java, COOL:Gen, ASP.Net, C#, JavaScript, ABAP, PL/I, Visual Basic, PowerBuilder, ASP, SQL, Visual Studio .Net, Datastage, .Net, IBM WTX, XML, COBOL, AB INITIO, A:G, C++)
Band Size	Band Size: Relative Band Size with 6 levels (1. XXS, 2. XS, 3. S, 4. M1, 5. M2, 6. L)
Dev. Platform	Development Platform with 5 levels (Multi, MF, PC, MR, ND=Not defined)
Team Size	Team Size: Team Size Group with 9 levels (2, 3-4, 5-8, 9-14, 15-20, 21-30, 31-40, 61-70, ND=Not defined)

Table 12 – Variables used in Effort Models.

6.1 Internal validity

These threats reflect to what extent the operational measures that are studied really represent what the researcher has in mind and what is investigated according to the research questions. The threats to the validity for this study are related to the ISBSG repository and correlation studies. First, the limited size and characteristics of the sub dataset used in some analysis may be one threat to internal validity. Data was filtered to make sure only desirable and high level quality information were used in the analysis and robust techniques were used to investigate correlations and prediction models. We filtered out outliers, to make sure that the results are not unduly influenced by a very small number of high leverage points. Even a large dataset as ISBSG does not offer representative data for several factors. Data preprocessing process and variables were chosen based on previous studies. We used nonparametric and robust techniques whenever the preconditions of parametric techniques were not supported by evidence. We also tried to identify homogeneous samples as possible. The whole dataset was divided in historical data, and new data. This division was repeated according to $N - PASS$. The cross validation technique used in effort estimation studies was applied to evaluate the effort estimation models. We selected the instances randomly avoiding the over fitting and optimistic models.

6.2 External validity

These threats are concerned with to what extent it is possible to generalize the findings. The ISBSG repository contains numerous projects from different domains and technologies. Projects of interest were filtered following a specific inclusion criteria in order to reduce the threat to external validity. This selection may improve the

Study	Id		Based on	R^2	Model
[KK93]			UFP	0.42	SR
			EI and EO	0.5	SR
[JS96]			UFP	0.58	SR
			EI and EO	$(n.s)$	SR
[LMR13]			TF	0.74	LMSR
			EI	0.41	LMSR
[QLJ15]			UFP	0.68	SR
			EI and EO	0.56	SR
			EI	0.52	SR
			TF	0.63	SR
			EI, EO and ILF	0.65	SR
			UFP, DevType, Language, Architecture and TeamSize	0.87	SR
			EI, EO, ILF, LangType, Language, Platform, Architecture and TeamSize	0.89	SR
This Study	(1)	(a)	UFP	0.47	LR
		(b)	TF	0.42	LR
		(c)	EI, EO, EQ and ILF	0.47	SR
		(d)	EI, EO and ILF	0.44	SR
		(e)	EI and ILF	0.42	SR
		(f)	EI	0.34	SR
	(2)	(a)	UFP	0.53	NLR
		(b)	TF	0.49	NLR
	(3)	(a)	UFP, BandSize, TeamSize, Languaje, DevType, DevPlatform	0.81	SR
		(b)	TF, BandSize, TeamSize, Languaje, DevType	0.79	SR
		(c)	EI, BandSize, TeamSize, DevType, Languaje, DevPlatform	0.79	SR

Kitchenham & Kansala [KK93], Jeffery & Stathis [JS96].
Lavazza, Morasca & Robiolo [LMR13].
Quesada-López & Jenkins [QLJ15].
SR: Stepwise regression, LMSR: LMS Regression Log transformation.
LR: Linear Regression, NLR: Non Linear regression (power).

Table 13 – Effort Models based on UFP and BFCs.

Id		MMRE	MdMRE	MMER	MdMER	MBRE	MdBRE	Pred (25)	Pred (50)
(1)	(a)	0.82	0.55	3.56	0.51	3.94	0.74	0.25	0.47
	(b)	0.89	0.55	5.00	0.54	5.45	0.78	0.26	0.48
	(c)	0.85	0.52	1.85	0.52	2.25	0.81	0.23	0.48
	(d)	0.88	0.54	1.14	0.52	1.57	0.86	0.23	0.48
	(e)	0.94	0.61	2.00	0.56	2.47	0.82	0.19	0.41
	(f)	1.07	0.59	0.98	0.54	1.59	0.90	0.20	0.42
(2)	(a)	0.63	0.43	0.64	0.41	0.90	0.55	0.33	0.59
	(b)	0.67	0.44	0.67	0.44	0.96	0.55	0.29	0.58
(3)	(a)	0.79	0.46	1.45	0.38	1.85	0.53	0.28	0.55
	(b)	0.91	0.46	0.94	0.38	1.47	0.54	0.28	0.54
	(c)	0.89	0.55	1.17	0.47	1.63	0.75	0.25	0.47

Table 14 – Effort Models Accuracy Evaluation.

Model		MRE		MER		BRE	
Id	Pair	Z	p	Z	p	Z	p
(1)	(a)(b)	-1.643	< 0.100	-1.500	< 0.134	-1.626	< 0.104
	(a)(c)	-0.285	< 0.776	-0.285	< 0.776	-0.354	< 0.724
	(a)(d)	-0.352	< 0.725	-0.282	< 0.778	-0.633	< 0.527
	(a)(e)	-2.836	< 0.005	-3.041	< 0.002	-2.899	< 0.004
	(a)(f)	-1.843	< 0.650	-0.570	< 0.569	-1.129	< 0.259
(2)	(a)(b)	-1.500	< 0.134	-0.495	< 0.621	-1.701	< 0.089
(3)	(a)(b)	-1.282	< 0.200	-1.357	< 0.175	-1.418	< 0.156
	(a)(c)	-2.203	< 0.028	-2.578	< 0.010	-2.333	< 0.020

Table 15 – Wilcoxon Signed Rank Test Results.

models for the analysis between UFP and BFCs. The software project data sets used in our experiments were based on the business application domain. The small size of some of the samples may make the models we evaluated of limited external validity. We also tried to mitigate the potential threats due to the changes in the development technologies by selecting only recent projects. This could make the data we used more applicable to current projects.

6.3 Construct validity

The ISBSG repository contains numerous projects for which variances in quality are beyond our control. To reduce this threat, only projects checked in the database as high quality were selected. The use of MMRE, MdMRE, and Pred(25) as accuracy indicators have been subjected to multiple criticisms in the previous literature [KPMS01], we provided other accuracy indicators to detail results about the accuracy of our results. The linear regression equation of actual work against UFP, for the selected dataset, shows evidence that there is a positive relationship. However, the ($R^2 = 0.47, p < 0.000$) were not significant, and therefore prediction models based on this data should be constructed carefully.

7 Conclusions and Future Work

This paper reports an empirical study of a family of replications applying the guidelines proposed by Carver [Car10]. The study evaluates the structure and applicability of function points in a project dataset from the ISBSG repository. The results presented above corroborate some of the findings of the original studies. First, most of the BFCs appear to be correlated with UFP. The results showed that BFCs are not independent because there are correlations between EI and EQ, EI and ILF, and EQ and ILF. Some BFCs are significantly correlated with effort. EI, ILF and EQ presented similar correlations as UFP. These results support the findings of previous studies where ILF and EQ have correlation with effort. In addition, ILF and EI are found to be always correlated, and EIF is found to be uncorrelated with the others. The results provide additional evidence to suggest that some subsets of FPA base functional components (BFC) could offer effort prediction models at least as good as the sum of all the BFCs. Effort models based on UPF, BFCs and nominal variables support previous findings. Preliminary results in this study suggest that the use of some context attributes in

prediction models may improve the results. Results showed a correlation between UFP and effort, and BFCs and effort.

The findings confirm previous results that suggest that a simplified counting method, based for example solely on some BFCs, could provide the same estimates as UFP. The analyses indicates that a prediction model based on TF or EI, EO and ILF appear to be as good as UFP. There is evidence to suggest that a subset of BFCs may offer an effort prediction model at least as good as UFP. The prediction accuracy of models was tested with MRE, MER, BRE, and Pred (25) as indicators. Wilcoxon signed rank test and the Mann-Whitney was applied to test to the distributions of MRE, MER, and BRE results. The Wilcoxon signed rank test indicated that the models based on BFCs are equivalent to the model based on UFP in terms of MRE, MRE, and BRE. Further research on evaluating and improving accuracy of prediction models based on a subset of basic functional components (BFC) is needed.

The results might suggest an improvement in the performance of the measurement activities. Organizations counting only a subset of BFCs could reduce duration, effort and cost of measurement process with respect to UFP. As [LMR13] mentioned, this could help organizations to collect historical data, and to build simpler effort prediction models. The results of this study are a starting point for further research in FSM methods and their base functional components. To improve this work and prove some of the theories, we would like to assess some simplified effort predictions models based on the preliminary results using BFCs, and context nominal attributes. Additionally, an analysis of correlations between the FPA BFC according to development types, industry sectors, organization types, application types, language types, program languages, and different technologies will be conducted in order to examine differences with related studies. Based on these results, future work could investigate the correlation between FPA, FFP and NESMA and their BFCs.

References

[ACG09] A. Abran and J.J. Cuadrado-Gallego. Software Estimation: Universal Models or Multiple Models? In *Seke*, pages 625–630, 2009. doi:10.1.1.409.7833.

[AG83] A.J. Albrecht and J.E. Gaffney. Software functions, source lines of codes and development effort prediction: a software science validation. *IEEE Transactions on Software Engineering*, 9(11):639–648, 1983. doi:10.1109/TSE.1983.235271.

[Alb79] A.J. Albrecht. Measuring application development productivity. In *Proceedings of the Joint SHARE/GUIDE/IBM Application Development Symposium*, volume 10, pages 83–92, 1979.

[Boe84] B.W. Boehm. *Software Engineering Economics*, volume SE-10. Prentice-hall Englewood Cliffs (NJ), 1984. doi:10.1109/TSE.1984.5010193.

[CA13] L. Cheikhi and A. Abran. PROMISE and ISBSG software engineering data repositories: A survey. In *Proceedings - Joint Conference of the 23rd International Workshop on Software Measurement and the 8th International Conference on Software Process and Product Measurement, IWSM-MENSURA 2013*, pages 17–24. IEEE, 2013. doi:10.1109/IWSM-Mensura.2013.13.

[Car10] J. Carver. Towards reporting guidelines for experimental replications: a proposal. In *1st International Workshop on Replication in Empirical Software Engineering Research*, pages 2–5, 2010. URL: `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.455.6270&rep=rep1&type=pdf`.

[Coo77] D.R. Cook. Detection of Influential Observation in Linear Regression. *Technometrics*, 19(1):15–18, 1977. `doi:10.1080/00401706.2000.10485981`.

[DTOB08] O. Demirors, O. Turetken, O. Ozcan, and O. Baris. The Impact of Individual Assumptions on Functional Size Measurement. In *Software Process and Product Measurement*, volume NA, pages 155–169. Springer, 2008.

[DVMB12] K. Dejaeger, W. Verbeke, D. Martens, and B. Baesens. Data mining techniques for software effort estimation: a comparative study. *Software Engineering, IEEE Transactions on*, 38(2):375–397, 2012. `doi:10.1109/TSE.2011.55`.

[FB97] N. Fenton and J. Bieman. *Software Metrics: A Rigorous and Practical Approach*, volume 2. CRC Press, 1997. `doi:10.1201/b17461`.

[GH01] D. Garmus and D. Herron. *Function point analysis: measurement practices for successful software projects*. Addison-Wesley Longman Publishing Co., Inc., 2001.

[GLdGFD14] F. González-Ladrón-de Guevara and M. Fernández-Diego. ISBSG Variables Most Frequently Used for Software Effort Estimation: A Mapping Review. In *Proceedings of the 8th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, pages 42:1—-42:4. ACM, 2014. `doi:10.1145/2652524.2652550`.

[HFH+09] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten. The WEKA data mining software. *ACM SIGKDD Explorations Newsletter*, 11(1):10, nov 2009. `doi:10.1145/1656274.1656278`.

[Hil11] P.R. Hill. *Practical software project estimation; a toolkit for estimating software development effort & duration*, volume 65. McGraw Hill Professional, 2011.

[ISO05] ISO/IEC. 24570: Software engineering – NESMA functional size measurement method version 2.1 – Definitions and counting guidelines for the application of Function Point Analysis. Standard, International Organization for Standardization, Geneva, CH, 2005.

[ISO07] ISO. Information technology – Software measurement – Functional size measurement. Part 1: Definition of concepts. Standard, International Organization for Standardization, Geneva, CH, 2007.

[ISO09] ISO. ISO/IEC 20926:2009 Software and systems engineering - Software measurement - IFPUG functional size measurement method 2009. Standard, International Organization for Standardization, Geneva, CH, 2009.

[JBR09] M. Jorgensen, B. Boehm, and S. Rifkin. Software Development Effort Estimation: Formal Models or Expert Judgment? *IEEE Computer Society*, 26(02):14–19, 2009. doi:10.1109/MS.2009.47.

[JLB93] R. Jeffery, G. Low, and M. Barnes. A Comparison of Function Point Counting Techniques. *IEEE Transactions on Software Engineering*, 19(5):529–532, may 1993. doi:10.1109/32.232016.

[Jon07] C. Jones. *Estimating Software Costs*. McGraw-Hill, Inc., 2007.

[Jon13] C. Jones. Function points as a universal software metric. *ACM SIGSOFT Software Engineering Notes*, 38(4):1, 2013. doi:10.1145/2492248.2492268.

[Jør07] M. Jørgensen. Forecasting of Software Development Work Effort: Evidence on Expert Judgment and Formal Models 2 . Software Development Effort Estimation. *International Journal of Forecasting*, 23(3):449–462, 2007. doi:10.1016/j.ijforecast.2007.05.008.

[JRW00] R. Jeffery, M. Ruhe, and I. Wieczorek. A comparative study of two software development cost modeling techniques using multi-organizational and company-specific data. *Information and Software Technology*, 42(14):1009–1016, 2000. doi:10.1016/S0950-5849(00)00153-1.

[JS96] R. Jeffery and J. Stathis. Function point sizing: Structure, validity and applicability. *Empirical Software Engineering*, 1(1):11–30, 1996. doi:10.1007/BF00125809.

[JS07] M. Jorgensen and M. Shepperd. A systematic review of software development cost estimation studies. *IEEE Transactions on Software Engineering*, 33(1):33–53, 2007. doi:10.1109/TSE.2007.256943.

[JYW+11] B. Jeng, D. Yeh, D. Wang, S.L. Chu, and C.M. Chen. A Specific Effort Estimation Method Using Function Point. *Journal of Information Science and Engineering*, 27(4):1363–1376, 2011. URL: http://www.iis.sinica.edu.tw/page/jise/2011/201107_11.pdf.

[Kem87] C.F. Kemerer. An empirical validation of software cost estimation models. *Communications of the ACM*, 30(5):416–429, 1987. doi:10.1145/22899.22906.

[Kit95] B. Kitchenham. Using function points for software cost estimation. *Software Quality Assurance and Measurement'(Eds Fenton NE, Whitty RW, Iizuka Y)*, pages 266–280, 1995.

[KK93] B. Kitchenham and K. Kansala. Inter-item correlations among function points. In *Proceedings of 1993 15th International Conference on Software Engineering*, pages 477–480. IEEE, 1993. doi:10.1109/ICSE.1993.346018.

[KM09] B. Kitchenham and E. Mendes. Why comparative effort prediction studies may be invalid. In *Proceedings of the 5th International Conference on Predictor Models in Software Engineering*, pages 1–5. ACM, 2009. doi:10.1145/1540438.1540444.

[KPME02] B. Kitchenham, S.L. Pfleeger, B. McColl, and S. Eagan. An empirical study of maintenance and development estimation accu-

racy. *Journal of Systems and Software*, 64(1):57–77, 2002. doi:10.1016/S0164-1212(02)00021-3.

[KPMS01] B. Kitchenham, L. Pickard, S. MacDonell, and M. Shepperd. What accuracy statistics really measure. In *IEE Proceedings - Software*, volume 148, page 81. IET, 2001. doi:10.1049/ip-sen:20010506.

[LJ90] G. Low and R. Jeffery. Function points in the estimation and evaluation of the software process. *Software Engineering, IEEE Transactions on*, 16(1):64–71, 1990. doi:10.1109/32.44364.

[LM06] C. Lokan and E. Mendes. Cross-company and single-company effort models using the ISBSG database: a further replicated study. In *Proceedings of the 2006 ACM/IEEE international ...*, pages 75–84. ACM, 2006. URL: http://dl.acm.org/citation.cfm?id=1159747, doi:10.1145/1159733.1159747.

[LMR13] L. Lavazza, S. Morasca, and G. Robiolo. Towards a simplified definition of Function Points. *Information and Software Technology*, 55(10):1796–1809, 2013. doi:10.1016/j.infsof.2013.04.003.

[Lok99] C. Lokan. An empirical study of the correlations between function point\nelements [software metrics]. In *Proceedings Sixth International Software Metrics Symposium (Cat. No.PR00403)*, pages 200–206. IEEE, 1999. doi:10.1109/METRIC.1999.809741.

[MA08a] N. Mittas and L. Angelis. Combining regression and estimation by analogy in a semi-parametric model for software cost estimation. In *International Symposium on Empirical Software Engineering and Measurement*, pages 70–79. ACM, 2008. doi:10.1145/1414004.1414017.

[MA08b] N. Mittas and L. Angelis. Comparing cost prediction models by resampling techniques. *Journal of Systems and Software*, 81(5):616–632, 2008. doi:10.1016/j.jss.2007.07.039.

[MJ03] K. Molokken and M. Jorgensen. A review of software surveys on software effort estimation. In *Empirical Software Engineering, 2003. ISESE 2003. Proceedings. 2003 International Symposium on*, pages 223–230. IEEE, 2003. doi:10.1109/ISESE.2003.1237981.

[ML08] E. Mendes and C. Lokan. Replicating studies on cross- vs single-company effort models using the ISBSG Database. *Empirical Software Engineering*, 13(1):3–37, 2008. doi:10.1007/s10664-007-9045-5.

[MLHT05] E. Mendes, C. Lokan, R. Harrison, and C. Triggs. A replicated comparison of cross-company and within-company effort estimation models using the isbsg database. In *Software Metrics, 2005. 11th IEEE International Symposium*, pages 10—-pp. IEEE, 2005. doi:10.1109/METRICS.2005.4.

[MTON94] Y. Miyazaki, M. Terakado, K. Ozaki, and H. Nozaki. Robust regression for developing software estimation models. *Journal of Systems and Software*, 27(1):3–16, 1994. doi:10.1016/0164-1212(94)90110-4.

[Obj14] Object Management Group. Automated Function Points (AFP) Version 1.0, OMG Document Number: formal/2014-01-03. Standard, Object Management Group, http://www.omg.org/spec/AFP, 2014.

[PAP10] C.E.L. Peixoto, J.L.N. Audy, and R. Prikladnicki. The importance of the use of an estimation process. In *Proceedings - International Conference on Software Engineering*, pages 13–17. ACM, 2010. doi:10.1145/1808981.1808983.

[QLJ14] C. Quesada-López and M. Jenkins. Function Point Structure and Applicability Validation Using the ISBSG Dataset: A Replicated Study. In *Proceedings of the 8th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, pages 66:1—-66:1. ACM, 2014. doi:10.1145/2652524.2652595.

[QLJ15] C. Quesada-López and M. Jenkins. An empirical validation of function point structure and applicability: A replication study. In *CIBSE 2015 - XVIII Ibero-American Conference on Software Engineering*, pages 418–431, 2015.

[RBLB00] P. Runeson, N. Borgquist, M. Landin, and W. Bolanowski. An evaluation of functional size methods and a bespoke estimation method for real-time systems. In *Product Focused Software Process Improvement*, pages 339–352. Springer, 2000. doi:10.1007/978-3-540-45051-1_30.

[Rid02] O. Ridge. A survey of dimension reduction techniques, 2002. URL: https://e-reports-ext.llnl.gov/pdf/240921.pdf.

[SBJ13] Y.S. Seo, D.H. Bae, and R. Jeffery. AREION: Software Effort Estimation based on Multiple Regressios with Adaptive Recursive Data Partitioning. *Information and Software Technology*, 55(rilis 9):1710–1725, 2013. doi:10.1016/j.infsof.2013.03.007.

[SCM05] L. Santillo, M. Conte, and R. Meli. Early & Quick Function Point: Sizing More with Less. In *Software Metrics, 2005. 11th IEEE International Symposium*, pages 1 –6. IEEE, 2005. doi:10.1109/METRICS.2005.33.

[SCVJ08] F. Shull, J. Carver, S. Vegas, and N. Juristo. The role of replications in Empirical Software Engineering. *Empirical Software Engineering*, 13(2):211–218, 2008. doi:10.1007/s10664-008-9060-1.

[SM98] E. Stensrud and I. Myrtveit. Human performance estimating with analogy and regression models: an empirical validation. In *Proceedings Fifth International Software Metrics Symposium. Metrics (Cat. No.98TB100262)*, pages 205–213. IEEE, 1998. doi:10.1109/METRIC.1998.731247.

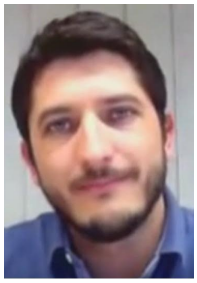
[Sym88] C.R. Symons. Function Point Analysis: Difficulties and Improvements. *IEEE Transactions on Software Engineering*, 14(1):2–11, 1988. doi:10.1109/32.4618.

[Tea05] R Development Core Team. *tR: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2005. URL: http://www.r-project.org.

[TOOD08] O. Turetken, O. Ozcan, B. Ozkan, and O. Demirors. The Impact of Individual Assumptions on Functional Size Measurement. In *Software Process and Product Measurement*, volume NA, pages 155–169. Springer, 2008. doi:10.1007/978-3-540-89403-2_14.

[WRH+12] C. Wohlin, P. Runeson, M. Host, M. Ohlsson, B. Regnell, and A. Wesslen. *Experimentation in Software Engineering*. Springer Science & Business Media, 2012. arXiv:arXiv:1011.1669v3, doi:10.1016/S0065-2458(08)60338-1.

About the authors



Christian Quesada-López is a doctoral candidate in the Department of Computer Science at the University of Costa Rica, where he also holds a teaching and research position. He has been continuously involved in empirical software engineering research since his graduate master's thesis, in the Center for ICT Research (CITIC) at University of Costa Rica. Prior to his doctoral work he had a 10 year career in software product development, software project management, and software quality assurance. His main research interests are empirical software engineering, measurement, and software quality assurance. Contact him at cristian.quesadalopez@ucr.ac.cr. http://www.citic.ucr.ac.cr/perfil/cristian-quesada-lópez.

Marcelo Jenkins obtained a B.S. degree in Computer and Information Sciences at the University of Costa Rica in 1986 and a M.Sc. and Ph.D. degrees from the University of Delaware, USA, in 1988 and 1992 respectively. Since 1986 he has been teaching computer science at the University of Costa Rica. From 1993 until 1998 he coordinated the Graduate Committee and from 1998 through 2001 he was the Chairman of the Department of Computer and Information Sciences. His research interests are in empirical software engineering, software quality assurance, project management, and object-oriented programming. He has authored more than 60 technical papers on these subjects. As an independent consultant, he has worked with some of the largest software companies in the Central America region in establishing software quality management systems. In the last 15 years, he has taught several seminars on software quality assurance and software project management in 7 different countries. Dr. Jenkins is an ASQ Certified Software Quality Engineer (CSQE) and a member of the IEEE Computer Society. Contact him at Contact him at marcelo.jenkins@ucr.ac.cr. http://www.citic.ucr.ac.cr/perfil/marcelo-jenkins-coronas.

Acknowledgments This research was supported by University of Costa Rica Project No. 834-B5-A18, and Ministry of Science, Technology and Telecommunications (MICITT). Our thanks to The International Software Benchmarking Standards Group (ISBSG) and the Empirical Software Engineering (ESE) Group at University of Costa Rica.