An Empirical Validation of Function Point Structure and Applicability: A Replication Study

Christian Quesada-López, Marcelo Jenkins

Center for ICT Research, University of Costa Rica, San Pedro, Costa Rica
{cristian.quesadalopez, marcelo.jenkins}@ucr.ac.cr

Abstract. Background: The complexity of providing accurate software size estimation 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 metric 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 ISBSG R12 dataset to evaluate the structure and applicability of function points. The goal of this replication was to aggregate evidence about internal issues of FPA as a metric, and to confirm previous results using a different set of data. First, FPA counting was analyzed in order to determine the extent to which the base functional components (BFC) were independent of each other and thus appropriate for an additive model of size. Second, the correlation between 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. **Methods:** A subset of 72 business application projects from 2008 to 2011 was analyzed. BFCs, UFP, and effort correlation were studied. **Results:** The results aggregated evidence and confirmed that some BFCs of the FPA method are correlated. There is a relationship between BFCs and effort. There are correlations between UFP and inputs, enquiries, and internal files, and between BFCs and effort. Internal files and inputs are found to be correlated always, and external interface files are found to be uncorrelated with the others. A prediction model based on transactions and internal files appear to be as good as a model based on UFP. The use of some contexts attributes may improve effort prediction models. **Limitations:** This is an initial experiment of a research in progress. The limited size and nature of the dataset may influence the results. **Conclusions**: Our results might suggest an improvement in the performance of the measurement process. Simplifying FPA measurement procedure based on counting a subset of BFCs could improve measurement process efficiency and simplify prediction models.

Keywords: Function point Analysis, effort prediction, family of replications, experiment.

1 Introduction

Software estimation process is a key factor for software project success [1]. 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 [2]. Inaccurate estimates are often the main cause of a great number of issues related to low quality and missed deadlines [3]. 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 [3]. Several companies consider formal estimation methods such as function points to be too complex and unpractical for their processes.

Software size measurement is an important part of the software development process [4, 5]. 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 [4, 5, 6], quality assessment, benchmarking, and outsourcing contracts [5]. According to [7], 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 [5]. Software size has proved to be one of the main effort-and-cost drivers [3, 8, 9, 10]. 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 [3, 6, 9, 11, 12].

Base functional components (BFC) inter-correlation is likely to involve two problems. First, from a practical point of view, correlation between BFC 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 [13, 14]. Practitioners use the BFCs relations useful to predict FPA count from single elements without applying the entire method [15].

This paper reports on a family of replications [16] based on [13, 17, 18, 14] and carried out on a subset of the ISBSG R12 dataset to evaluate the structure and applicability of function points. A family of replications is interesting because all studies are related and investigate related questions in different contexts [16]. 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,

after that, use other knew effort drivers in order to try 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 compare results with [13, 17, 18, 19, 14, 20]. 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 [16]. 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. Finally, Section 6 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). Function point analysis (FPA) [8, 9] was the first proposal for a FSM and it is one of the most used FSM methods in the industry [23]. In FPA the user requirements are classified and counted in a set of basic functional size components (BFC). These elementary units are called data and transactional functions. They represent data and operations that are relevant to the users. Data functions (DF) are classified into internal logic files (ILF) and external interface files (EIF). Transactional functions are classified into external inputs (EI), external outputs (EO), and external inquires (EQ). 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. Details about FPA method can be found in FPA manual [21]. FPA is independent from technology based influences [9]. FPA can be used to develop a measure of productivity [4, 22]. FPA have been subject to a number of critiques: the reliability of FPA measurement [4], the BFCs have inter correlations with each other [6, 12, 18], the application and usefulness of the complexity adjustments [22]. FPA is prone to different interpretations by different subjects. It is expected variation in the counts and finally, the counting method is slow and expensive [23]. 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 [14].

3 Description of the Original Studies

The original studies 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 [13] analyzed the internal consistency of FPA and the use of FPA to predict effort. Jeffery, Low and Barnes [17] investigated complexity adjustments in FPA and BFCs correlation. Jeffery and Stathis [18] 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 [15] studied correlations between BFCs in FPA and analyzed how factors influenced the balance between BFCs. Quesada-López and Jenkins [20], in a previews study, empirically investigated correlations between BFCs, UFP and effort. Lavazza, Morasca & Robiolo [14] analyzed correlations between BFCs to evaluate the possibility of a simplified definition of function points. The goals and research questions from the original studies and related with the replication are provided in Table 1.

Table 1. Goals and Research Questions

Authors	Goals and Research Questions
Kitchenham & Kansala [13]	(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 [18]	(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 [19]	(1) To describe correlations between the FPA elements according to development type, language type, and program language.
Lavazza, Morasca & Robiolo [14]	(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 [20]	(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.

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.

Table 2. Information about original studies

Authors	Dataset	Dataset Type	Domain
Kitchenham & Kansala [13]	40 projects from 9 software development organizations	Cross company	MIS
Jeffery, Low & Barnes [17]	64 projects from 1 software development organization	Within- company	MIS
Jeffery & Stathis [18]	17 projects from 1 software development organization	Within- company	MIS
Lokan [19]	269 projects from the ISBSG R4 dataset	Cross company	MIS, DSS
Lavazza, Morasca & Robiolo [14]	Over 600 projects from the ISBSG R11 dataset	Cross company	MIS
Quesada-López & Jenkins [20]	14 projects from the ISBSG R4 dataset	Cross company	MIS

Table 3. Independent and dependent variables

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

3.3 Summary of Results

Kitchenham and Kansala [13] 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 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 [17] also found that BFC are not independent. Furthermore, they concluded that processing complexity adjustment had not effect on the accuracy of the effort models. Jeffery and Stathis [18] found statistically 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 a simplified sizing metric may be appropriate. Lokan [19] 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 and Robiolo [14] determine correlations between BFCs and assess encouraging effort prediction models based on a simplified count. Quesada-López and Jenkins [20] 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.

The results showed that BFCs size measures were actually correlated, and this suggests that a simplified form of function point sizing method (i.e. based on data) would be possible across different domains. Some authors expect that simpler counting would reduce the variability of the counting results. Several studies have explored the possibility of a simplified function point method. As an example, Symons [25] based Mark II on the basis of three BFC, Early & Quick Function Points (EQFP) [26] 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 [27] 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. Lavazza et al. [14] proposed a simplified definition of FP using only subsets of BFCs. Many other practical software size approximation and simplified techniques are presented in [24].

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 [16]. In this study, we compare results with [13, 17, 18, 19, 14, 20]. Correlations between the BFCs have been found in previous studies but their findings were different in some respects, but not in others. 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 [15]. 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.

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 [28].

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. Two types of changes were made on purpose: the context and the data and independent variable selection. 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 [24]. The data can be used for effort estimation, trend analysis, comparison of platforms and languages, and productivity benchmarking [29]. The ISBSG repository is a multi-organizational, multi-application, and multi-environment data repository [30]. 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 [24]. The subset of data projects for our study was selected according to the criteria shown in Table 4. For our study, we selected the variables related with FPA functional size components (BFC) and effort of software development. Projects with all BFCs size measures missing were discarded. The list of selected variables is shown in Table 5.

Table 4. Project selection criteria

Criteria	Value	Motivation
Count Approach	IFPUG 4+	Latest FPA standard and counting rules
Data Quality Rating	A	Only data with an high level of quality and integrity
Unadjusted Function Point Rating	A	Counting data with a high level of quality and integrity
Year of project	> 2008	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. ISBSG Dataset Variables used in this study

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	Unadjusted Function Point count (UFP)
Normalized Level 1 Work Effort	Ratio	The development team full life-cycle effort
Normalized Level 1 Productivity Delivery Rate	Ratio	Productivity delivery rate in hours per functional size unit (UFP)
Context Attributes	Nominal	Development Type, Relative Size, Team Size Group, Development Platform, Architecture, Language Type, Program. Language, Development Method

As a result of the selection, a total of seventy two project data were included in our analysis. Twenty nine of them are from 2008, twenty five from 2009, thirteen from

2010, and five from 2011. Table 6 shows details of the groups from projects according different nominal attributes. In each case percentage related to the number of projects and functional size (UFP) by categorical attribute is presented (attributes and categories are the defined in the dataset by the ISBSG).

The normality test indicates that the unadjusted function points (UFP), and productivity data belonged to normal distribution (Kolmogorov-Smirnov test). The Levene test confirmed equality of variances. Table 7 summarizes the normality test results, and the projects UFP, effort, productivity, and BFC data. The smallest project size is 24 UFPs, the average is 240 UFPs, and the largest project is 1,337 UFPs. The average productivity for the dataset is 23.67 hours per UFP, with a range from 3 to 59 hours per UFP.

Table 6. ISBSG Sub Dataset Demographic Summary (72 projects)

Relative Size	**Pry**	**%**	**UFP**	**%**		**CMMI2**	**Pry**	**%**	**UFP**	**%**
6. L (1000-3000)	1	1.4	1,337	7.7		0	17	23.6	3,351	19.4
5. M2 (300-1000)	20	27.8	8,605	49.7		1	7	9.7	1,638	9.5
4. M1 (100-300)	33	45.8	6,084	35.1		2	43	59.7	10,626	61.4
3. S (30-100)	17	23.6	1,260	7.3		5	4	5.6	1,611	9.3
2. XS (10-30)	1	1.4	24	0.1		ND	1	1.4	84	0.5
Team Size	**Pry**	**%**	**UFP**	**%**		**Language**	**Pry**	**%**	**UFP**	**%**
ND	13	18.1	2,624	15.2		Other	8	11.1	1,999	11.5
31-40	2	2.8	1,718	9.9		ABAP	8	11.1	2,273	13.1
21-30	3	4.2	1,354	7.8		ASP.Net	3	4.2	448	2.6
15-20	8	11.1	2,701	15.6		COOL:Gen	9	12.5	1,629	9.4
9-14	21	29.2	5,666	32.7		Java	10	13.9	2,376	13.7
5-8	19	26.4	2,622	15.1		C#	16	22.2	4,521	26.1
3-4	6	8.3	625	3.6		PL/I	18	25.0	4,064	23.5
Dev Type	**Pry**	**%**	**UFP**	**%**		**Architecture**	**Pry**	**%**	**UFP**	**%**
Re-development	1	1.4	112	0.6		Client server	44	61.1	11,761	67.9
New Development	14	19.4	4,650	26.9		Stand-alone	20	27.8	4,410	25.5
Enhancement	57	79.2	12,548	72.5		Multitier & web	8	11.1	1,139	6.6
Devt Platform	**Pry**	**%**	**UFP**	**%**		**Language Type**	**Pry**	**%**	**UFP**	**%**
Multi-Platform	46	63.9	12,322	71.2		ND	1	1.4	372	2.1
Main Frame	19	26.4	4,359	25.2		ApG	9	12.5	1,629	9.4
PC	6	8.3	460	2.7		4GL	16	22.2	3,682	21.3
Mid-Range	1	1.4	169	1.0		3GL	46	63.9	11,627	67.2

5 Comparison and Discussion of Results

5.1 Data Analysis

Scatter plot of actual work against UFP for the dataset shows evidence that there is a positive relationship between effort and UFP (R2 = 0.68). A comparison of these results against previous studies is shown in Table 8. This data shows the sensibility of the results depending of the data selection.

Table 7. Data Summary and Kolmogorov-Smirnov Test

	N	Mean	Std. Dev.	Min/Max	p
Size UFP	72	240.42	202.302	24-1,337	<.051
Effort	72	6,134.29	9,135.852	167-71,729	(n.s.)
Productivity	72	23.6778	12.40049	3.00-59.00	<.217
EI	72	88.78	95.263	0-551	<.009
EO	72	46.40	59.703	0-287	<.002
EQ	72	58.72	57.005	0-275	<.073
ILF	72	39.71	48.181	0-252	<.001
EIF	72	6.81	12.617	0-54	(n.s.)

Table 8. Previous studies comparison – UFP against effort

		UFP versus Effort	
Study	**Projects**	**R squared**	**(p)**
Albrecht, Gaffney [9]	24	0.90	<0.001
Kemerer [12]	15	0.54	<0.001
Kitchenham, Kansala [13]	40	0.41	<0.010
Jeffery, Low & Barnes [17]	64	0.36	<0.001
Jeffery & Stathis [18]	17	0.95	<0.001
Jeffery & Stathis [18]	14	0.58	<0.001
Quesada-López. Jenkins [20]	14	0.94	<0.000
Quesada-López. Jenkins [20]	12	0.62	<0.003
This Study	72	0.68	<0.000

5.2 Internal Consistency of Function Points

Table 9 shows the Kendalls's Tau correlation coefficients between all pairs of function point BFCs using the entire dataset (72 projects). Previous study results are also presented in Table 9 for comparison. Outliers were removed from datasets in [18, 19, 20]. The results showed that BFCs are not independent. Jeffery & Stathis [18] reports some differences in results with [13, 19, 20]. These studies found correlations in EO and EI, EO and EQ, EO and EIF, and EO and ILF not presented in [18]. Jeffery & Stathis [18] reports agreement with [13] in EI and EQ, EI and ILF, and EQ and ILF. These correlations are presented also in [19, 20] and the current study. The results in the current study agree with all the studies in correlations between EI and EQ, EI and ILF, and EQ and ILF as is presented in Table 9. We agreed with the authors regarding to 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.

Table 9. Kendall Tau correlation coefficients comparison between BFCs

Study	BFC	UFP	EI	EO	EQ	EIF
[13]	EI	**0.67** p<0.001				
[18]		**0.54** p<0.01				
[19]		(n.r.)				
[14]		**0.658** (n.r.)				
[20]		**0.74** p<0.00				
This Study		**0.64** p<0.00				
[13]	EO	0.53 p<0.001	**0.47** p<0.001			
[18]		0.27 (n.s.)	0.03 (n.s.)			
[19]		(n.r.)	**0.37** p<0.001			
[14]		**0.597** (n.r.)	**0.438** (n.r.)			
[20]		**0.45** p<0.04	**0.55** p<0.01			
This Study		0.34 p<0.00	0.19 p<0.19			
[13]	EQ	**0.47** p<0.001	**0.47** p<0.001	**0.32** p<0.01		
[18]		**0.68** p<0.001	**0.72** p<0.001	-0.06 (n.s.)		
[19]		(n.r.)	**0.48** p<0.001	**0.29** p<0.001		
[14]		**0.528** (n.r.)	**0.448** (n.r.)	**0.288** (n.r.)		
[20]		**0.80** p<0.00	**0.61** p<0.00	**0.25** p<0.27		
This Study		**0.54** p<0.00	**0.38** p<0.00	0.03 p<0.66		
[13]	EIF	0.32 p<0.01	0.14 (n.s.)	**0.31** p<0.01	0.60 (n.s.)	
[18]		-0.37 (n.s.)	-0.56 p<0.05	0.03 (n.s.)	-0.53 p<0.05	
[19]		(n.r.)	-0.02 (n.s.)	0.10 (n.s.)	0.00 (n.s.)	
[14]		0.264 (n.r.)	0.072 (n.r.)	0.194 (n.r.)	0.097 (n.r.)	
[20]		0.42 p<0.07	0.16 p<0.50	0.00 p<1.00	0.41 p<0.08	
This Study		-0.04 p<0.69	-0.15 p<0.11	-0.27 p<0.77	-0.02 p<0.80	
[13]	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.)
[18]		**0.73** p<0.001	**0.44** p<0.05	0.11 (n.s.)	**0.65** p<0.001	-0.39 (n.s.)
[19]		(n.r.)	**0.48** p<0.001	**0.33** p<0.001	**0.41** p<0.001	0.08 p<0.02
[14]		**0.619** (n.r.)	**0.449** (n.r.)	**0.417** (n.r.)	**0.327** (n.r.)	0.195 (n.r.)
[20]		**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
This Study		**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

Kitchenham & Kansala [13], Jeffery & Stathis [18], Lokan [19], Lavazza, Morasca &Robiolo [14], Quesada-López & Jenkins [20]. (n.s.) not significant. (n.r.) not reported.

5.3 Using UFP and BFCs to predict effort

Table 8 shows evidence of correlations between UFP and effort and Table 9 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 10 shows that some BFCs are significantly correlated with effort. For the dataset in this study, 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. The results provide

additional evidence to suggest that some subset of FPA UFP base functional components could offer an effort prediction models at least as good as the sum of all the BFCs. For example, Kitchenham & Kansala [13] found that a combination of EI and EO offers better correlation with effort than UPF. Lavazza, Morasca and Robiolo [14] reported that a prediction model based on EI, EO and transactional function (TF) were as good as a model based on UFP.

Table 10. Correlation coefficients between UFP, BFCs and effort

Study	BFC	Pearson	Kendall Tau	Spearman
[13]	**UFP**	0.65 p<0.001	(n.r.)	(n.r.)
[18]		0.58 p<0.01	(n.r.)	(n.r.)
[20]		0.785 p<0.003	(n.r.)	(n.r.)
This Study		0.825 <0.000	0.607 p<0.000	0.793 p<0.000
[13]	**EI**	0.60 p<0.001	(n.r.)	(n.r.)
[18]		0.37 p<0.001	(n.r.)	(n.r.)
[20]		0.531 p<0.076	(n.r.)	(n.r.)
This Study		0.720 p<0.000	0.484 p<0.000	0.667 p<0.000
[13]	**ILF**	0.44 p<0.01	(n.r.)	(n.r.)
[18]		0.73 p<0.001	(n.r.)	(n.r.)
[20]		0.588 p<0.05	(n.r.)	(n.r.)
This Study		0.622 p<0.000	0.456 p<0.000	0.613 p<0.000
[13]	**EQ**	0.28 (n.s)	(n.r.)	(n.r.)
[18]		0.63 p<0.001	(n.r.)	(n.r.)
[20]		0.861 p<0.001	(n.r.)	(n.r.)
This Study		0.596 p<0.000	0.416 p<0.000	0.561 p<0.000
[13]	**EO**	0.66 p<0.001	(n.r.)	(n.r.)
[18]		0.03 (n.s)	(n.r.)	(n.r.)
[20]		0.277 p<0.383	(n.r.)	(n.r.)
This Study		0.525 p<0.000	0.320 p<0.000	0.431 p<0.000
[13]	**EIF**	0.31 (n.s)	(n.r.)	(n.r.)
[18]		0.005 (n.s)	(n.r.)	(n.r.)
[20]		0.857 p<0.00	(n.r.)	(n.r.)
This Study		0.233 p<0.049	0.040 p<0.659	0.057 p<0.632
Kitchenham & Kansala [13], Jeffery & Stathis [18], Quesada-López & Jenkins [20]. (n.s.) not significant. (n.r.) not reported.				

Table 11 shows the correlation coefficient results 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 [31]. Preliminary results shows that the use of these context attributes in prediction models may improve the results, but further research is still needed.

Table 11. Effort models based on UFP and BFCs

Study	Based on	R2	Model
[13]	UFP	0.42	Stepwise regression
[13]	EI and EO	0.50	Stepwise regression
[18]	UFP	0.58	Stepwise regression
[18]	EI and EO	(n.s)	Stepwise regression
[14]	TF	0.74	LMS Regression. Log transformation
[14]	EI	0.41	LMS Regression. Log transformation
This Study	UFP	0.68	Stepwise regression
	EI and EO	0.56	Stepwise regression
	EI	0.52	Stepwise regression
	TF	0.63	Stepwise regression
	EI, EO and ILF	0.65	Stepwise regression
	UFP, DevType, Language, Architecture and TeamSize	0.87	Stepwise regression. Dummy coding for nominal attributes
	EI, EO, ILF, LangType, Language, Platform, Architecture and TeamSize	0.89	Stepwise regression. Dummy coding for nominal attributes
Kitchenham & Kansala [13], Jeffery & Stathis [18], Lavazza, Morasca &Robiolo [14]. (n.s.) not significant. LMS: Least Median of Squares. TF: (EI+EO+EQ)			

6 Threats to Validity

This section analyses the threats to the validity for this study and the actions undertaken to mitigate them.

- Threats to internal validity: 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 dataset 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.
- Threats to external validity: 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 models for current projects in the industry.
- Threats to 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.

7 Conclusions and Future Work

This paper reports an empirical study of a family of replications applying the guidelines proposed by Carver [16]. The study evaluates the structure and applicability of function points in a project dataset from the ISBSG repository. The results presented above support some of the findings of the original studies. First, most of the BFCs appear to be

correlated with UFP as shown in Table 9. The results showed that BFCs are not independent because there are correlations between EI and EQ, EI and ILF, and EQ and ILF. Table 10 shows that 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. Besides, ILF and EI are found to be correlated always, and EIF is found to be uncorrelated with the others. The results provide additional evidence to suggest that some subset of FPA UFP base functional components could offer an effort prediction models at least as good as the sum of all the BFCs. Table 11 shows the correlation coefficient results between UFP and effort, and BFCs and effort. The results from this study support the findings of the previous studies. Preliminary results in this study shows that the use of some context attributes in prediction models may improve the results. Further research is needed.

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 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. Further research 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 [14] 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 asses some simplified effort predictions models based on the preliminary results using BFCs, and context nominal attributes. To evaluate the models we would like to use the MMRE and PRED() metrics as accuracy indicators. Besides, in order to examine differences with related studies, an analysis of correlations between the FPA BFC according to development type, industry sector, organization type, application type, language type, and program language will be conducted. Based on these results, future work could investigate the correlation between FPA, FFP and NESMA and their BFCs.

8 Acknowledgments

This research was supported by the Costa Rican 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.

9 References

1. Peixoto, C. E. L., Audy, J. L. N., & Prikladnicki, R. (2010, May). The importance of the use of an estimation process. In *Proceedings of the 2010 ICSE Workshop on Software Development Governance* (pp. 13-17). ACM.

2. Molokken, K., & Jorgensen, M. (2003, October). A review of software surveys on software effort estimation. In *Empirical Software Engineering*, 2003. ISESE 2003, (pp. 223-230). IEEE.
3. Boehm, B. W. (1981). *Software engineering economics*.
4. Low, G. C., & Jeffery, D. R. (1990). Function points in the estimation and evaluation of the software process. *Software Engineering, IEEE Transactions on*, *16*(1), 64-71.
5. Garmus, D., & Herron, D. (2001). *Function point analysis: measurement practices for successful software projects*. Addison-Wesley Longman Publishing Co., Inc.
6. Kitchenham, B. (1993) Using Function Points for Software Cost Estimation – Some Empirical Results. *10th Annual Conference of Software Metrics and Quality Assurance in Industry, Amsterdam*, Netherlands.
7. ISO. (2007). *ISO/IEC 14143-1- Information Technology - Software measurement - Functional Size Measurement. Part 1: Definition of Concepts*.
8. Albrecht, A. J. (1979, October). Measuring application development productivity. In *Proceedings of the Joint SHARE/GUIDE/IBM Application Development Symposium* (Vol. 10, pp. 83-92). Monterey, CA: SHARE Inc. and GUIDE International Corp.
9. Albrecht, A. J., & Gaffney, J. E. (1983). Software function, source lines of code, and development effort prediction: a software science validation. *Software Engineering, IEEE Transactions on*, (6), 639-648.
10. Jeng, B., Yeh, D., Wang, D., Chu, S. L., & Chen, C. M. (2011). A Specific Effort Estimation Method Using Function Point. *Journal of Information Science and Engineering*, *27*(4), 1363-1376.
11. Jones, T. C. (1998). *Estimating software costs*. McGraw-Hill, Inc.
12. Kemerer, C. F. (1987). An empirical validation of software cost estimation models. *Communications of the ACM*, *30*(5), 416-429.
13. Kitchenham, B., & Kansala, K. (1993, May). Inter-item correlations among function points. In *Software Engineering, 1993. Proceedings, 15th International Conference on* (pp. 477-480). IEEE.
14. Lavazza, L., Morasca, S., & Robiolo, G. (2013). Towards a simplified definition of Function Points. *Information and Software Technology*, *55*(10), 1796-1809.
15. Lokan, C. J. (1999). An empirical study of the correlations between function point elements.
16. Carver, J. C. (2010, May). Towards reporting guidelines for experimental replications: A proposal. *In International Workshop on Replication in Empirical Software Engineering Research*, Cape Town, South Africa.
17. Jeffery, D. R., Low, G. C., & Barnes, M. (1993). A comparison of function point counting techniques. *Software Engineering, IEEE Transactions on*, *19*(5), 529-532.
18. Jeffery, R., & Stathis, J. (1996). Function point sizing: structure, validity and applicability. *Empirical Software Engineering*, *1*(1), 11-30.
19. Lokan, C. J. (1999). An empirical study of the correlations between function point elements.
20. Quesada-López, C., & Jenkins, M. (2014). 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 (ESEM '14)*. Torino, Italy.
21. ISO. (2009). *ISO/IEC 20926, Software and systems engineering - Software measurement – IFPUG functional size measurement method*.
22. Jones, T. C. (1998). *Estimating software costs*. McGraw-Hill, Inc.
23. Jones, C. (2013). Function points as a universal software metric. *ACM Software Engineering Notes*, *38*(4), 1-27.
24. Hill, P. (2010). *Practical Software Project Estimation*. Tata McGraw-Hill Education.
25. Symons, C. R. (1988). Function point analysis: difficulties and improvements. *Software Engineering, IEEE Transactions on*, 14(1), 2-11.
26. Conte, M., Iorio, T., Meli, R., & Santillo, L. (2004, January). E&Q: An Early & Quick Approach to Functional Size Measurement Methods. In *Software Measurement European Forum–SMEF*.
27. ISO. (2005). *ISO/IEC 24570 - Software engineering -- NESMA functional size measurement method version 2.1 -- Definitions and counting guidelines for the application of Function Point Analysis*.
28. Shull, F. J., Carver, J. C., Vegas, S., & Juristo, N. (2008). The role of replications in empirical software engineering. *Empirical Software Engineering*,*13*(2), 211-218.
29. Mendes, E., Lokan, C., Harrison, R., & Triggs, C. (2005, September). A replicated comparison of cross-company and within-company effort estimation models using the isbsg database. In *Software Metrics, 2005. 11th IEEE International Symposium* (pp. 10-pp). IEEE.
30. Cheikhi, L., & Abran, A. (2013, October). Promise and ISBSG Software Engineering Data Repositories: A Survey. In *Software Measurement and the 2013 Eighth International Conference on Software Process and Product Measurement (IWSM-MENSURA), 2013 Joint Conference of the 23rd International Workshop on* (pp. 17-24). IEEE.
31. Dejaeger, K., Verbeke, W., Martens, D., & Baesens, B. (2012). Data mining techniques for software effort estimation: A comparative study. Software Engineering, IEEE Transactions on, 38(2), 375-397.