

Re-evaluation of Estimation Study Results using ISBSG Data -Using a different dataset-

Shinji Kusumoto
Department of Computer Science
Graduate School of Information Science and Technology
Osaka University
kusumoto@ist.osaka-u.ac.jp

2023/2/17



Contents:

- Background
- Research objective
- Research Questions (RQs)
- Results
- Conclusion



Background

In (empirical) software engineering, it is important to conduct replication studies on the findings of existing research studies[1].

- Can findings obtained under certain conditions be replicated under other conditions?
- Would different conditions or environments yield different findings?



Integrate the findings obtained

Increase the relevance of reusing the findings

[1] Fabio Q. B. da Silva et al. "Replication of empirical studies in software engineering research: a systematic mapping study," *Empirical Software Engineering*, 19, 3, pp.501-557 (2014).



Research Objectives

Conducting follow-up (replication) studies on existing studies of estimations where ISBSG repositories are used in the evaluations

- Change the repository used for evaluation



- Different findings:
 - Leads to consideration of causes of different results and additional research
- Similar findings:
 - The external validity of existing research results has been enhanced

the validity of applying the conclusions of a scientific study outside the context of that study



ISBSG dataset

- Dataset collected from IT companies around the world
- Wide array of data available for benchmarking
 - 10,600 projects in the Development and Enhancement Repository and 1177 applications in its Maintenance and Support Repository.
- Example attributes: person-hours, number of developers, size (FP, LOC), development language, platform, development type...



Research Questions (RQs)

- RQ1: How many studies in the past five years (2012-2017) have used ISBSG dataset in the estimation studies?
- RQ2: How many studies exist that can be tested and be easy to apply to actual development projects?
- RQ3: Are there any differences between the findings from ISBSG data set and ones from different data sets?



RQ1:

- The following major international conferences and journals were surveyed:
 - ACM Transactions On Software Engineering and Methodology
 - ASIA-PACIFIC Software Engineering Conference
 - Empirical Software Engineering and Measurement
 - International Conference on Software Engineering
 - International Conference on Software Maintenance
 - IEEE Transactions on Software Engineering
 - Information and Software Technology
 - Journal of Systems and Software
- Twenty-two of the 83 papers about software estimations used ISBSG dataset in the evaluation.



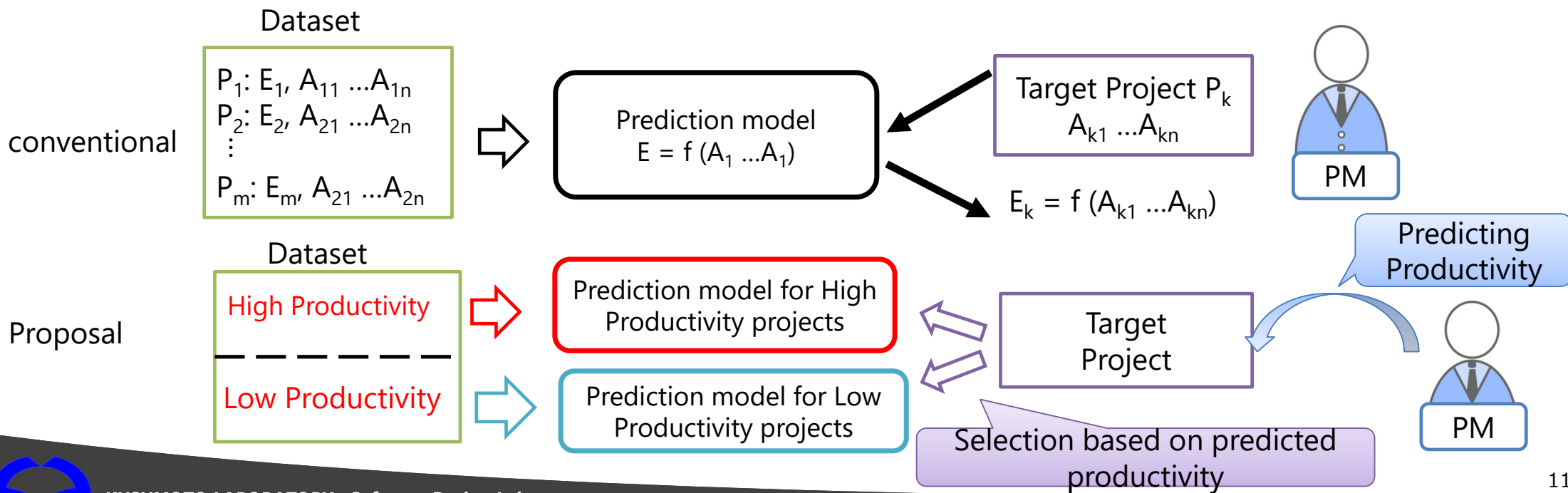
RQ2:

- From the 22 papers, it was necessary to select ones that could be for the follow-up (re-evaluation).
- Conditions for selection
 - Actual project data are used in the evaluation.
 - Only data that are available to us are used.
 - Accurately describes the information necessary for the re-evaluation
 - Easy to apply in the actual projects
 - ...
- The following 4 papers are selected.
 - (P1) M. Tsunoda et al.: Incorporating expert judgment into regression models of software effort estimation. In APSEC, 1, 374–379. IEEE, 2012.
 - (P2) M. Tsunoda et al.: Handling categorical variables in effort estimation. In ESEM, 99-102. 2012.
 - (P3) L. Lavazza et al.: Towards a simplified definition of Function Points. IST, 55, 10, 1796–1809, 2013.
 - (P4) C. López Martín et al.: Neural networks for predicting the duration of new software projects. JSS, 101, 127– 135, 2015.



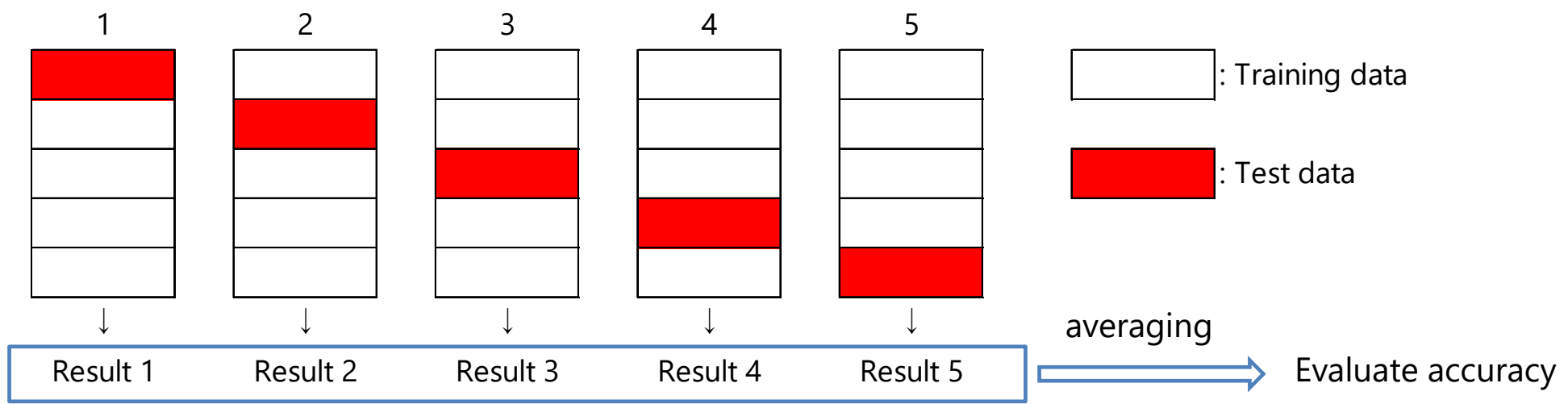
Summary of (P1) (1/3)

- Contents
 - Effort estimation models based on regression analysis
 - It proposes a new method of creating an effort estimation model that takes into account the project manager's prediction of the productivity of the target project.



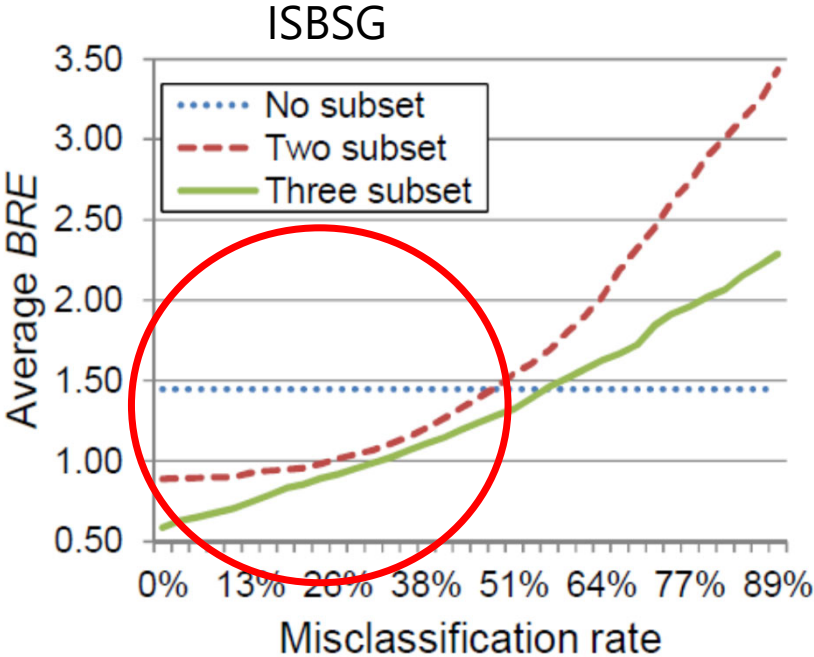
Summary of (P1) (2/3)

- Evaluation
 - Cross validation



Summary of (P1) (3/3)

Accuracy changes with increasing prediction error rate of PM



Selection criteria

FP Standard	IFPUG4
Normalised Work Effort	Equal to effort recorded
Data Quality Rating	A or B
Resource Level	1
Functional Size	Not missing.
Language Type	Not missing.
Development Type	Not missing.
Development Platform	Not missing.

593 data from ISBSG, release 9

The proposed method is more effective when the PM is unlikely to make a mistake in predicting the productivity of the target project



Summary of (P2)

- Contents:
 - How to handle categorical variables effort estimation
 - Categorical variable: Gender, occupation, and other variables that generally cannot be measured in numbers or quantities.
 - In order to use categorical variables as explanatory variables in a regression model, such as dummy variable conversion, etc., is used.
 - Comparison of four different estimation models with different ways of dealing with categorical variables (Stratification, Dummy variables, Interaction, Hierarchical linear model) using four datasets (ISBSG, Desharnais, NASA FP, NASA)
 - Findings
 - The most effective method is different for dataset but the prediction accuracy will be higher if categorical variables are taken into account in the model.



Summary of (P3)

- Contents:
 - Measuring function points (FPs)
 - A simpler FP estimation model is proposed based on ISBSG dataset.
 - Created an FP estimation model that uses as explanatory variables only those factors that have the strongest correlation with FP among the five base functional components (EI, EO, EQ, ILF, EIF) that constitute FP.
- Findings:
 - The factor most strongly correlated with FP was external input (EI), showing a simple model with EI as the explanatory variable
 - $FP = 79 + 1.9 \times EI$

Selection criteria

Data Quality Rating	A or B
UFP rating	A or B
External Input	Not missing.
External Output	Not missing.
Internal Logical File	Not missing.

More than 600 data from ISBSG, release 11



2023/2/17

KUSUMOTO LABORATORY - Software Design Laboratory

Department of Computer Science, Graduate School of Information Science and Technology, Osaka University. <http://sdl.ist.osaka-u.ac.jp/>

Summary of (P4)

- Contents:
 - Software development duration estimation models
 - Accuracy of two different development duration estimation models using neural networks (NN) is investigated
- Findings:
 - Both of the two estimation models using NN improved accuracy over the multiple regression model.

Selection criteria

Data Quality Rating	A or B
UFP rating	A or B
FP Standard	IFPUG4/NESMA
Max Team Size	2 persons >
Project Elapsed Time	2 months >
Development Platform	Mainframe
Language Type	3GL
Development Type	New

49 data from ISBSG, release 11

Model	MAR	MdAR
Multiple linear regression	0.29	0.25
MLP(NN1)	0.18	0.15
RBFNN(NN2)	0.21	0.19

MAR: average of absolute residuals of duration
MdAR: Median of the absolute residuals



Summary of (P1)~(P4)

	Findings using ISBSG dataset
(P1)	The proposed method is more accurate than the conventional estimation method, if PM's prediction is accurate.
(P2)	Models using categorical variables are more accurate . However, the accuracy of the four models is almost the same.
(P3)	The base functional component that correlates most strongly with FP is EI .
(P4)	Both of the two NN models improved accuracy over the regression model.



RQ3:

- Are there any differences between the findings from ISBSG data set and ones from different data sets?

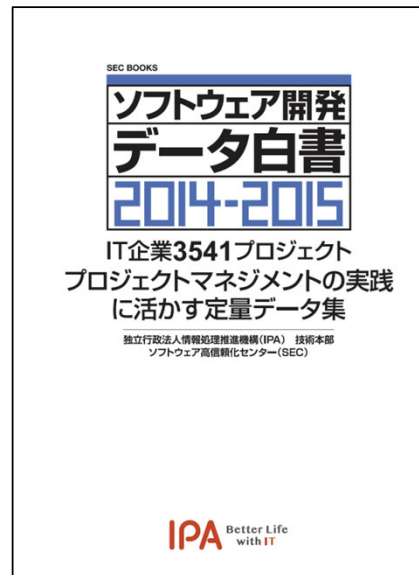


Dataset used in follow-up studies

IPA: INFORMATION-TECHNOLOGY PROMOTION AGENCY, JAPAN

SEC: Software Engineering Center (IT Knowledge Center on emerging tech trends)

- IPA/SEC dataset (P1 ~ P4)
 - Includes 3541 projects, each of which has 194 attributes
 - Example attributes : person-hours, number of developers, size (FP, LOC), development language, platform, development type...
 - It published "Software Development Data White Paper".
- A certain company dataset (P3)
 - Includes 98 projects, each of which has 144 attributes
 - From Insurance and Banking software projects



2023/2/17



KUSUMOTO LABORATORY - Software Design Laboratory

Department of Computer Science, Graduate School of Information Science and Technology, Osaka University. <http://sdl.ist.osaka-u.ac.jp/>

RQ3: Summary of (P1)~(P4) re-evaluation

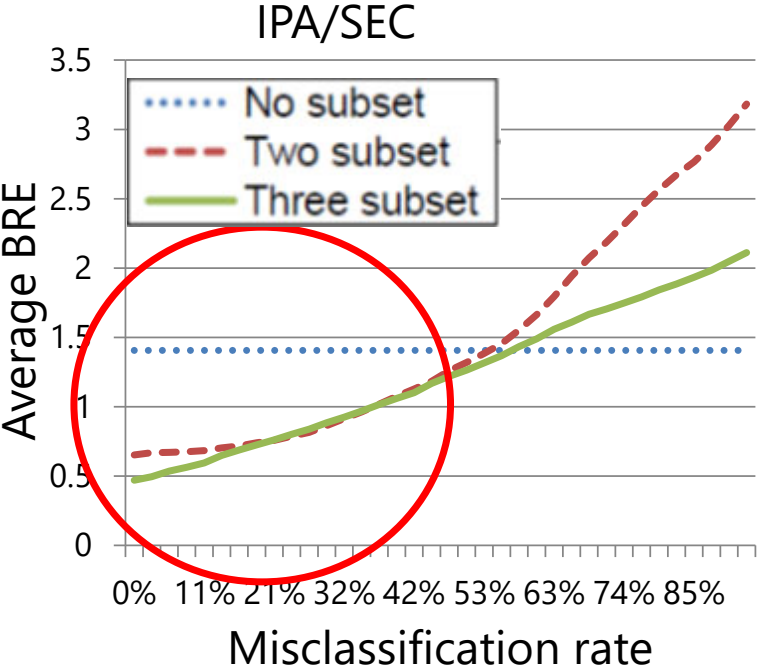
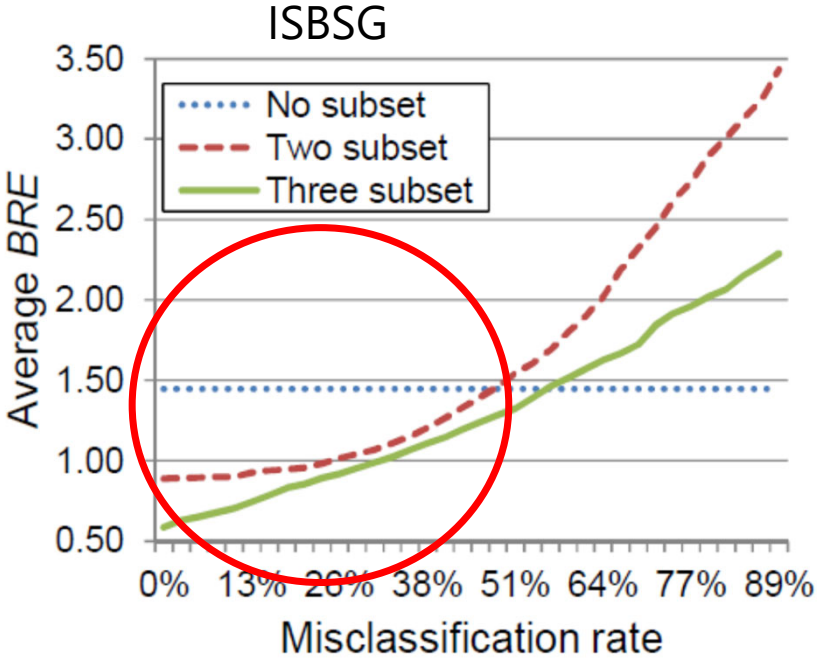
	Findings using ISBSG dataset	Findings using IPA/SEC dataset
(P1)	The proposed method is more accurate than the conventional estimation method, if PM's prediction is accurate.	Same as left.
(P2)	Models using categorical variables are more accurate . However, the accuracy of the four models is almost the same.	Same as left.
(P3)	The base functional component that correlates most strongly with FP is EI .	The base functional component that correlates most strongly with FP is EI or EQ (IPA/SEC dataset) and EO (A certain company dataset)
(P4)	Both of the two NN models improved accuracy over the regression model.	One of the two NN models improved accuracy over the regression model.

2023/2/17



Follow-up results (P1)

Accuracy changes with increasing prediction error rate of PM



189 data from IPA/SEC dataset

The proposed method is more effective when the PM is unlikely to make a mistake in predicting the productivity of the target project



Follow-up results (P3) (1/4)

The five base functional components of function point

- EI(External Input), EO(External Output), EQ(External Inquiry), Internal logical files (ILF), External interface files (EIF)

Base functional components most correlated with FP

Dataset (# of data)	Kendall	Spearman
ISBSG(600>)	EI	EI
IPA/SEC(122)	EI	EQ
A certain company(55)	EO	EO

When creating a simple FP estimation model using Base functional components, It is necessary to select the appropriate elements for each dataset



2023/2/17

KUSUMOTO LABORATORY - Software Design Laboratory

Department of Computer Science, Graduate School of Information Science and Technology, Osaka University. <http://sdl.ist.osaka-u.ac.jp/>

Follow-up results (P3) (2/4)

- Dataset Characteristics
 - ISBSG, IPA/SEC: Data from various application domains
 - A certain company: Data from Insurance + Banking application



Re-evaluation each repository, categorizing by application domain



2023/2/17

KUSUMOTO LABORATORY - *Software Design Laboratory*

Department of Computer Science, Graduate School of Information Science and Technology, Osaka University. <http://sdl.ist.osaka-u.ac.jp/>

Follow-up results (P3) (3/4)

Base functional components most correlated with FP (by Application domain)

dataset	Application domain(# of ata)	Kendall	Spearman
ISBSG	Total (619)	EI	EI
	Insurance (131)	EO	EO
	Communications (113)	ILF	ILF
	Government (90)	EI	EI
	Banking (47)	EI	EI
	Manufacturing (39)	EI	ILF
	Insurance + Banking (178)	EO	EO



2023/2/17

KUSUMOTO LABORATORY - Software Design Laboratory

Department of Computer Science, Graduate School of Information Science and Technology, Osaka University. <http://sdl.ist.osaka-u.ac.jp/>

Follow-up results (P3) (4/4)

Base functional components most correlated with FP (by Application domain)

dataset	Application domain(# of ata)	Kendall	Spearman
ISBSG	Total (619)	EI	EI
	Insurance (131)	EO	EO
	Communications (113)	ILF	ILF
	Government (90)	EI	EI
	Banking (47)	EI	EI
	Manufacturing (39)	EI	ILF
	Insurance + Banking (178)	EO	EO
IPA/SEC	Total (122)	EI	EQ
A certain company	Total (Insurance + Banking) (55)	EO	EO

It is better to make some categorization of the data before conducting the analysis.



Conclusions

- Conducting follow-up studies (re-evaluation) on existing studies of estimations where ISBSG repositories are used in the evaluations by Changing the repository used for evaluation
 - RQ1: There were 22 papers about software estimations used ISBSG dataset in the evaluation from 2012-2017.
 - RQ2: 4 papers were selected for re-evaluation.
 - RQ3: 2 papers had findings that differed from existing studies, while the other 2 papers had nearly identical findings.
- Future works:
 - Conduct follow-up studies on other existing software estimation.
 - A follow-up studies with a different type of data set.

