Advances in statistical analysis from the ISBSG benchmarking database

SBC {Software Benchmarking Committee}

Luca Santillo, Stefania Lombardi, Domenico Natale

Agenda

- Introducing GUFPI-ISMA SBC
- Current Work's Intents
- Demographic Overview
- Variables & Subsamples Selection
- Distribution Analyses
- Correlation Analysis
- 2-variables Distributions
- Summary & Conclusions

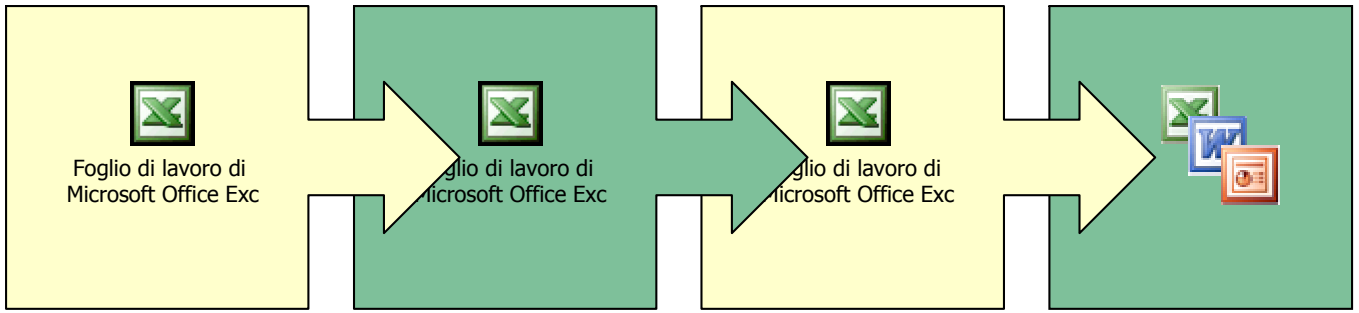
Introducing GUFPI-ISMA SBC

- GUFPI-ISMA is a member of **ISBSG**
  - "To help improve [IT management] through improved project estimation, productivity, risk analysis and benchmarking"
  - Data collection → SW Projects Benchmark
- Committes & SIGs – The **SBC**
  - Active members from SOGEI, DPO, PAT, FINSIEL, IC
  - "[…] the published results are not to be considered as a valid reference for any possible official, commercial or legal utilization. […]"

Current Work's Intents

- Statistical analysis scenarios
  - Development and enhancement projects (Benchmark 8, 2003)
  - On-going, incremental process (voluntary contribution)
- 1st step focus: distributions of DEV / ENH projects
  - Variables: **project size**, **work effort & productivity**, **primary programming language, solar duration**
- Some considerations and hints
  - Identifying categorization criteria and taxonomies +
  - Suggesting standardization topics to ISBSG, to…
  - Improve the data collection and exploitation (not critics!)

ISBSG Data & Analysis Previews



Demographic Overview (*extract*) (2,027 points)

Variable	ISBSG Name	N	%	Type	Range	Multiple	Calc
DQR	Data Quality Rating	2,027	100.0%	Ord.	4		
FP_STD_PRIMARY	FP Standard	1,938	95.6%	Text	25		
WE_TOT	Summary Work Effort	2,025	99.9%	Num.	-		
RL	Resource Level	2,027	100.0%	Ord.	4		
PRJ_TYPE	Development Type	2,027	100.0%	Text	5		
PLATFORM	Development Platform	1,418	70.0%	Text	3		
LANG	Programming Language	1,691	83.4%	Text	122	Yes	
DT	Development Techniques	1,025	50.6%	Text	227	Yes	
PRJ_TIME	Project Elapsed Time	1,639	80.9%	Num.	-		
IMPL_DATE	Implementation Date	1,802	88.9%	Date	-		
PACKAGE	Package Customisation	1,322	65.2%	Y/N	3		
FP_ADD/CHG/DEL	Adds/Changes/Del's (x3)	2,027	100.0%	Num.	-		
WE_NORM	Normalised Work Effort	2,024	99.9%	Num.	-		Yes
UFP	Unadjusted Size (FP)	1,568	77.4%	Num.	-		Part.
UFP_RAT	Size Rating	2,027	100.0%	Ord.	4		Yes
PDR	Project Delivery Rate	1,569	77.4%	Num.	-		Yes

Demographic Overview (*preliminary analysis*)

- 66 variables, collected by questionnairies
  - Only 15 variables contain values for more than 90% of the sample
  - "Language Type" and "Implementation Date" contain ambiguous values
  - 8 variables: multiple values (combinations of a smaller set of elementary values)
    - e.g. Primary Programming Language: more than one language
  - Textual variables show typing discrepancies
    - e.g. "COBOL 2", "COBOL II" and "COBOL V2"
    - "IFPUG" (no version) vs. "IFPUG 4" vs. "IFPUG 4.0" & "IFPUG 4.1"
  - "Y/N/Don't Know" variables contain void instances (= "No" or "Don't Know"?)
  - Variables carrying quantitative information expressed in free form
    - e.g. "Degree of Customisation" = "x% (yes)", "y% (no) "less than z%", "some", "partly", "0,1 (?)", …

Data manipulation

- ***Dichotomization*** ("Yes or void value" = binary):
  - "Project Scope" (5 Y/N variables vs. a value range of 26)
  - "FP Standards All" (22 Y/N variables vs. a value range of 67)
  - "Development Techniques" (20 Y/N variables vs. a range of 227)
  - "Organization Type" (37 Y/N variables vs. a range of 92)
  - "Business Area Type" (18 Y/N variables vs. a range of 98)
  - "Application Type" (24 Y/N variables vs. a range of 114)
  - EXCEPTION: Prim. Progr. Language (too many values!!!)

- ***Nomenclature*** (discussion and approval for each case):
  - Expressions interpreted as equivalent – rendered with unique term ("COBOL 2", "COBOL II", "COBOL V2" → "COBOL 2", etc.)

Suggestions to ISBSG

- ***Taxonomies*** (define closed ranges, particularly for textual type variables)
  - Leave the possibility of "other" assignment, with a separate note for comments or distinct values

- ***Completeness*** (insist on the mandatory filling of every and each project attribute)
  - Even if the sample is growing (>3,000 points now?), filtering the sample leads to small, complete-records subsamples

Selected Variables

- Size (UFP)
- Work effort (WE_TOT)
- Project delivery rate (PDR)
- Platform (PLATFORM)
- Primary programming language (LANG), and
- Solar duration (PRJ_TIME)

Generic distributions – not significant

- Mixing different measurement methods, project types, and so on

Subsamples Selection / Filtering Criteria

Step	Filtering Variable	Filtering Criteria	Excluded Records	Residual Records	%
1	PRJ_TYPE	= "New Development" OR "Enhancement"	1 "New Utility", 1 "Purchased Package", 55 "Re-Development"	1,970	97.2%
2	MM	= "IFPUG"	195 not "IFPUG"	1,775	87.6%
3	DQR	= "A" OR "B"	80 "C", 33 "D"	1,662	82.0%
4	FP_STD_PRIMARY	= "IFPUG *"	336 not "IFPUG *"	1,326	65.4%
↳	**Sample A ("soft filter" sub-sample; 1,326 records)**				
5	PACKAGE	≠ "Y"	68 "Y"	1,258	62.1%
6	UFP_RAT	= "A" OR "B"	184 "C", 1 "D"	1073	52.9%
7	FP_STD_PRIMARY	= "IFPUG 4.*"	1 "IFPUG" (1993), 13 "IFPUG 2", 136 "IFPUG 3", 9 "IFPUG 3.4"	914	45.1%
8	RL	= "1" OR "2"	6 "3", 134 "4"	774	38.2%
↳	**Sample B ("severe filter" sub-sample; 774 records)**				

Data categorization

- PRIM_PROG_LANG re-aggregated into LANG_LEV (Language Level)
  - Based on Capers Jones' languages table
- LANG_LEV re-aggregated into LL_CAT (Language Level Category)

LL_CAT	LANG_LEV Range	Examples
LL_CAT 1	1-3	ASSEMBLER (1); C (2.5); COBOL, COBOL 2, MVS COBOL (3); FORTRAN (3); PASCAL (3.5).
LL_CAT 2	4-8	PL/I (4); Generic 3rd Generation Language (4); LISP (5); C++, JAVA (6); ADA (6.5); CICS (7); ORACLE (8); MS ACCESS (8.5).
LL_CAT 3	9-15	VISUAL BASIC, VBSCRIPT (9); VISUAL C++ (9.5); DELPHI (11); PRO C (12); COOL.GEN (14); LOTUS NOTES, SMALLTALK, UNIX SHELL SCRIPT (15).
LL_CAT 4	16-23	ADS/O, NOTES SCRIPT, Generic 4th Generation Language (16); CLIPPER (17); POWERBUILDER, TELON, ABAP, SAP ABAP (20); HTML (22); ASP (23).
LL_CAT 5	24-55	SQL, EASYTRIEVE (25); PL/SQL, SQL WINDOWS (27); Spreadsheet (50).
LL_CAT 6	>55	Generic 5th Generation Language (70)

Data Transformation & Final Subsets

- Very few relevant data transformation
  - UFP size = (adj.d) FP value, where only "FP", no "VAF" & no "function breakdown", was provided
  - PDR re-calculated, including previously void values

Subsample B ("severe filter")

Variable	N	%	N_{DEV}	$\%_{DEV}$	N_{ENH}	$\%_{ENH}$	Value Type	Value Range	Calc
UFP	774	[100%]	299	[100%]	475	[100%]	Num.	-	Part.
PLATFORM	386	49.9	168	56.2%	217	45.7%	Text	3	
LANG_LEV	587	75.8	214	71.6%	373	78.5%	***Ord.***	***29***	Aggr.
LANG_CAT	587	75.8	214	71.6%	373	78.5%	***Ord.***	***6***	Aggr.
IMPL_PERIOD	719	***92.9***	270	90.3%	449	94.5%	***Ord.***	***6***	Aggr.
WE_TOT	774	***100.0%***	299	100.0%	475	100.0%	Num.	-	
PDR	774	***100.0%***	299	100.0%	475	100.0%	Num.	-	Yes

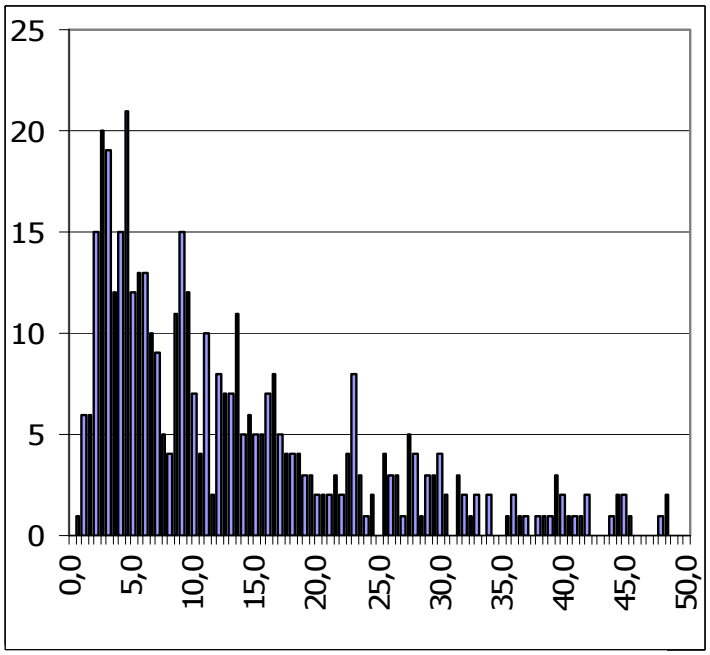
Distribution Analyses / Size (subs. B)

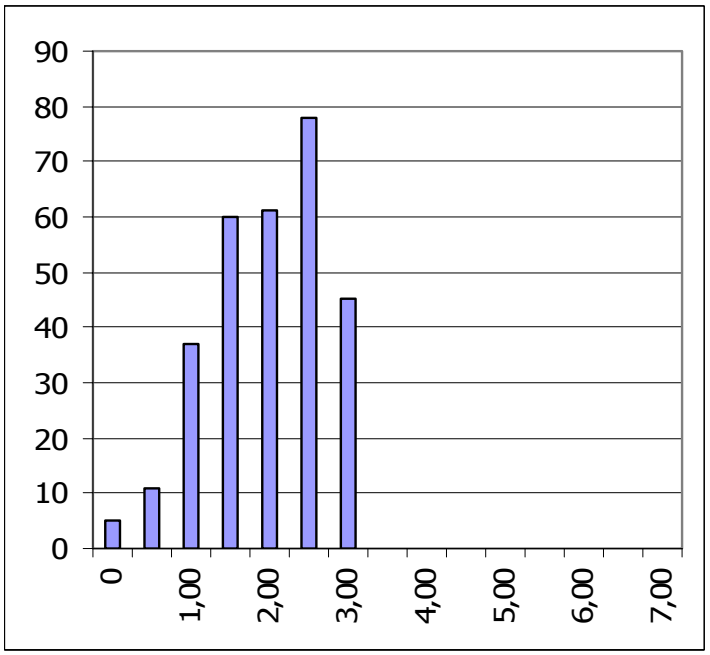
Code	SIZE_CLASS	Size Range (UFP)
DEV_{XS}	Very Small	0-150
DEV_{S}	Small	150-300
DEV_{M}	Medium	300-600
DEV_{L}	Large	600-1,200
DEV_{XL}	Very Large	*1,200-5,000*
DEV_{XXL}	Extremely Large	> 5,000

Distribution Analyses / Size (subs. B)

Code	SIZE_CLASS	Size Range (UFP)
ENH_{XS}	Very Small	0-60
ENH_{S}	Small	60-120
ENH_{M}	Medium	120-240
ENH_{L}	Large	240-480
ENH_{XL}	Very Large	*480-2,000*
ENH_{XXL}	Extremely Large	> 2,000

Distribution Analyses / PDR (subs. B)





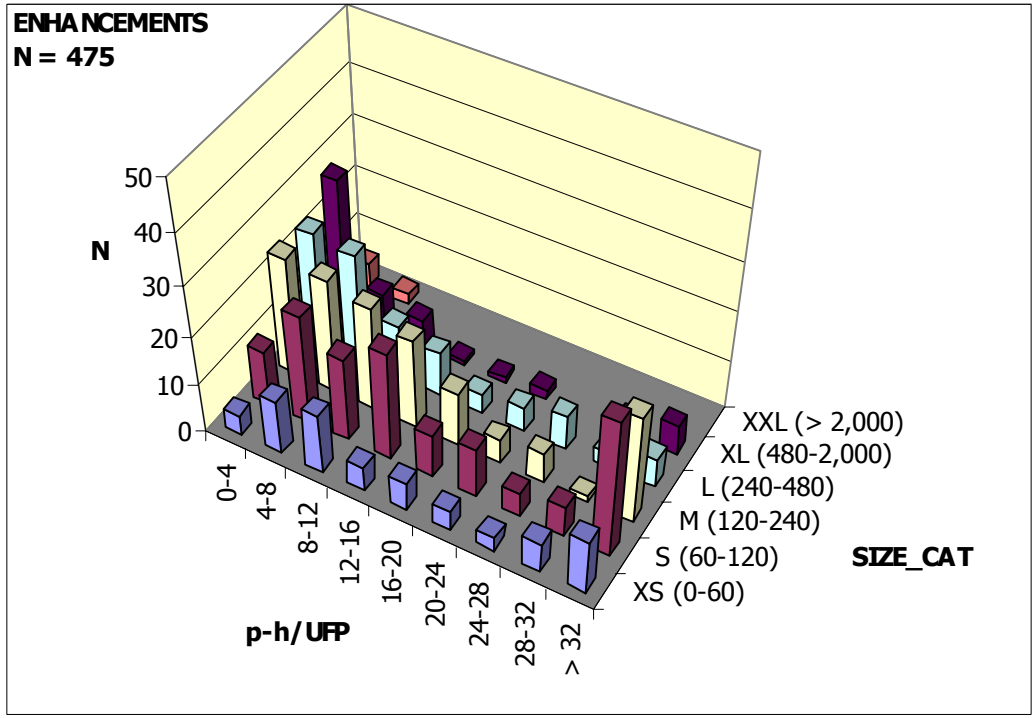
ENH PDR (ph/UFP); N = 475

Correlation Analysis ($r_{Pearson}$)

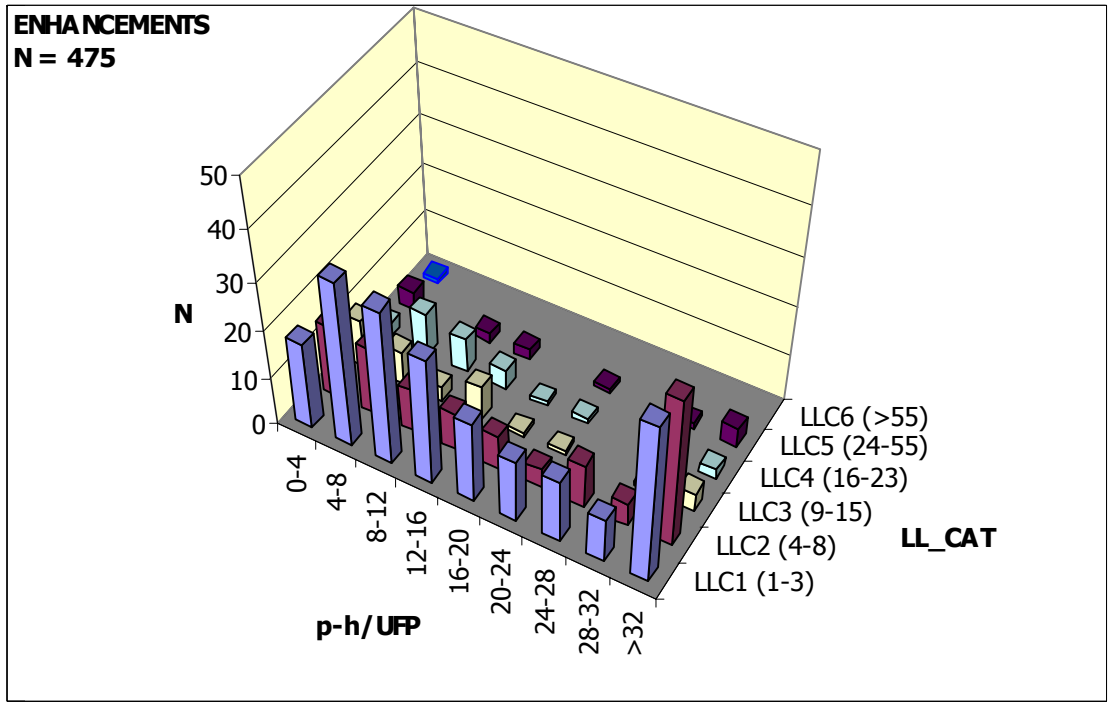
	DEV (N = 264)				*ENH (N = 309)*			
	UFP	**WE_TOT**	**PDR**	**PRJ_TIME**	**UFP**	**WE_TOT**	**PDR**	**PRJ_TIME**
UFP	1,00	**0,71**	**-0,07**	**0,36**	1,00	**0,40**	**-0,14**	**0,21**
WE_TOT	0,71	1,00	**0,32**	**0,65**	0,40	1,00	**0,30**	**0,21**
PDR	-0,07	0,32	1,00	**0,28**	-0,14	0,30	1,00	**0,10**
PRJ_TIME	0,36	0,65	0,28	1,00	0,21	0,21	0,10	1,00

- Positive correlation between size & effort (DEV), but…
  - …sparse and/or not linear (r^2 = 0.5)
  - …not true for enhancements?!?
    - Likely, because of IFPUG formula for Enhancements (ADD + 100% CHG + 100% DEL)

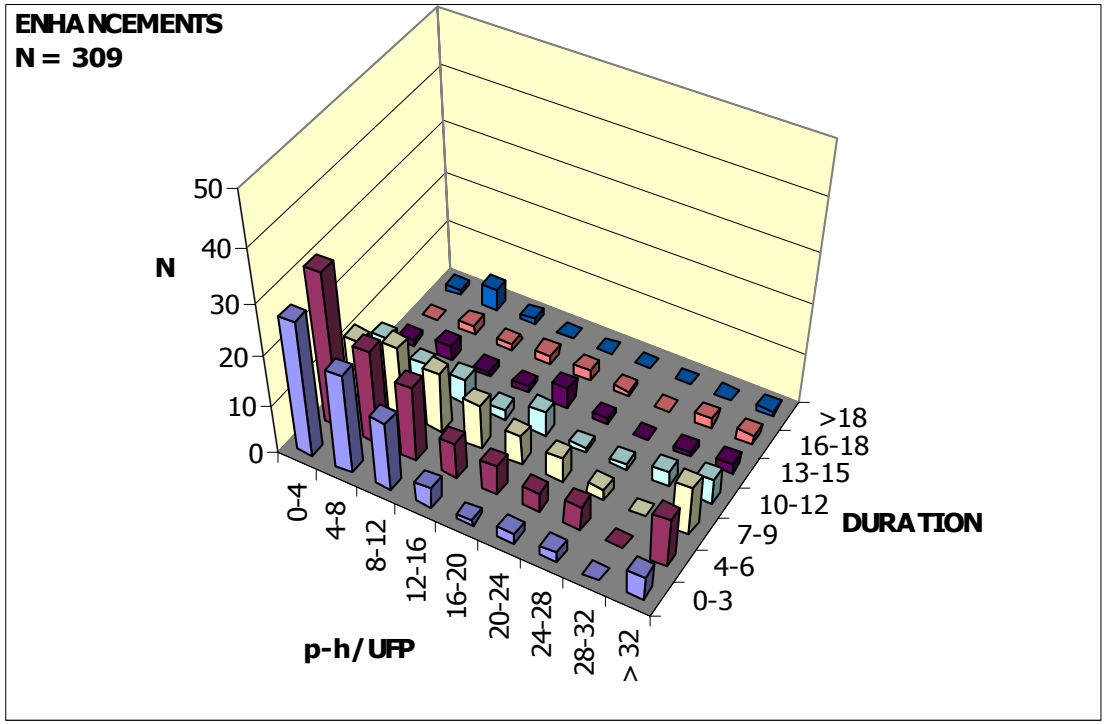
2-variables Distributions / PDR vs SIZE_CLASS



2-variables Distributions / PDR vs LL_CAT



2-variables Distributions / PDR vs PRJ_TIME



Summary

- 1st step analysis: several hints highlighted
  - Dichotomization (to avoid multiple values per record)
  - Nomenclature (to avoid distinct values for identical instances)
  - Taxonomy (to avoid open ranges)
  - Mandatory completeness (to avoid high filtering reduction)
- More suggestions (for better data interpretation)
  - New "comment" and "reliability" field per each variable
- Size classes suggested for standard use
- Upcoming suggestions for project estimation models

Conclusions

- Obviously, some old ISBSG data got obsolete by filtering
- Counterpart:
  - ⇑ Usability, ⇓ Analysis effort, ⇑ Analysis significance
- Further research
  - Move to Benchmark 9 (over 3,000 points)
  - Extension to COSMIC measures
  - Analysis of normalized PDR
  - Extension to more variables
  - Further two-variables and N-variables analyses

Co-coordinators: luca.santillo@gufpi.org, domenico.natale@gufpi.org

The truth…

- Say you were standing with one foot in the oven and one foot in an ice bucket. According to the percentage people, you should be **perfectly comfortable**.
  - Bobby Bragan

- Do not put your faith in what statistics say until you have carefully considered what they do **not** say.
  - William W. Watt

- Statistics can be made to prove anything… **even** the truth.
  - Unknown