PROMISE
The 2010 International Conference on Predictive Models in Software Engineering

SENSITIVITY OF RESULTS TO DIFFERENT DATA QUALITY META-DATA CRITERIA IN THE SAMPLE SELECTION OF PROJECTS FROM THE ISBSG DATASET

Marta Fernández-Diego

Mónica Martínez-Gómez

José-MaríaTorralba-Martínez

UNIVERSIDAD POLITÉCNICA DE VALENCIA

SPAIN

INTRODUCTION

- **Data quality = fundamental determinant of empirical results in software projects dataset**
  - Which datasets are more reliable?
  - How to choose the specific projects?
- **Data quality = large concept**
  - **Completeness**
  - Accuracy or absence of noise
  - Timeliness
  - Outliers or atypical observations

INVESTIGATION QUESTIONS

I. **Which studies explicitly consider data quality meta-data in the preprocessing of data?**

II. **How data quality meta-data criteria can influence the results?**

I. REVIEW OF DATA QUALITY META-DATA CRITERIA

Which studies, related with prediction models and based on ISBSG (International Software Benchmarking Standards Group**) dataset, explicitly consider data quality meta-data in the preprocessing of data?**

1. **ISBSG data quality variables**
2. **Method and results of the review**

1. ISBSG DATA QUALITY VARIABLES

- **2 variables**
  1. Data Quality Rating
  2. Unadjusted Function Point (UFP) Rating
- **Assessing projects data quality**
- **Rated in 4 categories**
  A. Absolute integrity
  B. Integrity possibly affected
  C. Integrity not assured
  D. Little credibility
- **Completeness of the case**

2. METHOD AND RESULTS OF THE REVIEW

Review method

- **Databases: ACM, IEEE Xplore, ScienceDirect, Web of Knowledge**
- **Search terms: "ISBSG", "ISBSG & quality"**
- **Hand search**
- **Information extracted**
  - Year of publication / ISBSG Release
  - Data Quality Rating / UFP Rating
  - Topic related with prediction models
  - Use of Size/Effort variables
  - Number of projects selected for analysis

2. METHOD AND RESULTS OF THE REVIEW

Results from the 26 papers retrieved

- **Topic related with prediction models**
  - 17 papers related with effort estimation
  - Size, duration, team size, substitution cost, defect estimation and productivity
- **Data Quality Rating**
  - 17 papers using data graded as A or B
- **UFP Rating**
  - 2 articles from the same first author, Gencel
  - Latest paper from 2010

II. EXPERIMENTAL RESULTS

Which is the effect of data quality meta-data criteria in the number of selected projects that can have influence in the behavior of the models obtained?

1. **Strength of the association for general data quality by functional size quality**
2. **Influence of data quality meta-data in the sample size**
3. **The case study: Productivity evolution over time**

SENSITIVITY OF RESULTS TO DIFFERENT DATA QUALITY META-DATA CRITERIA IN THE SAMPLE SELECTION OF PROJECTS FROM THE ISBSG DATASET

1. STRENGTH OF THE ASSOCIATION FOR GENERAL DATA QUALITY BY FUNCTIONAL SIZE QUALITY

- **Data quality meta-data distribution in the ISBSG dataset Release 11**

		Data Quality Rating			
		A | B | C | D	A | B | C	A | B	A
UFP Rating	A | B | C | D | ND	5052	4907	4744	928
	A | B | C | D	4512	4389	4243	840
	A | B | C	4497	4385	4243	840
	A | B	3369	3288	3184	836
	A	2328	2277	2202	676

- **Moderate positive association between both variables**
  - Gamma = 0.525

2. INFLUENCE OF DATA QUALITY META-DATA IN THE SAMPLE SIZE

Initial data preparation

- **Large heterogeneous dataset**
- **Selection criteria for productivity concerns**
  - All projects dated with the implementation date
  - Comparable definition for size
  - Comparable definition for effort along whole life cycle projects
- **Results**
  - 830 remained projects from the initial 5052
  - 4222 removed, representing 83.5% of total

2. INFLUENCE OF DATA QUALITY META-DATA IN THE SAMPLE SIZE

Data quality selection

- **Remained projects applying different data quality ratings**

		Data Quality Rating			
		A | B | C | D	A | B | C	A | B	A
UFP Rating	A | B | D	830	820	802	316
	A | B	828	820	802	316
	A	545	537	521	262

- **Results**
  - Data quality criteria may not add value

3. THE CASE STUDY: PRODUCTIVITY EVOLUTION OVER TIME

- **4 study cases**
  - A sample of 802 projects corresponding to (A | B) & (a | b)
  - A sample of 521 projects corresponding to (A | B) & a
  - A sample of 316 projects corresponding to A & (a | b)
  - A sample of 262 projects corresponding to A & a
- **Evolution of mean productivity over time**
- **Evolution of Byear (**regression coefficients**) over time**
- **Observations derived from the analysis of the 4 selected cases**

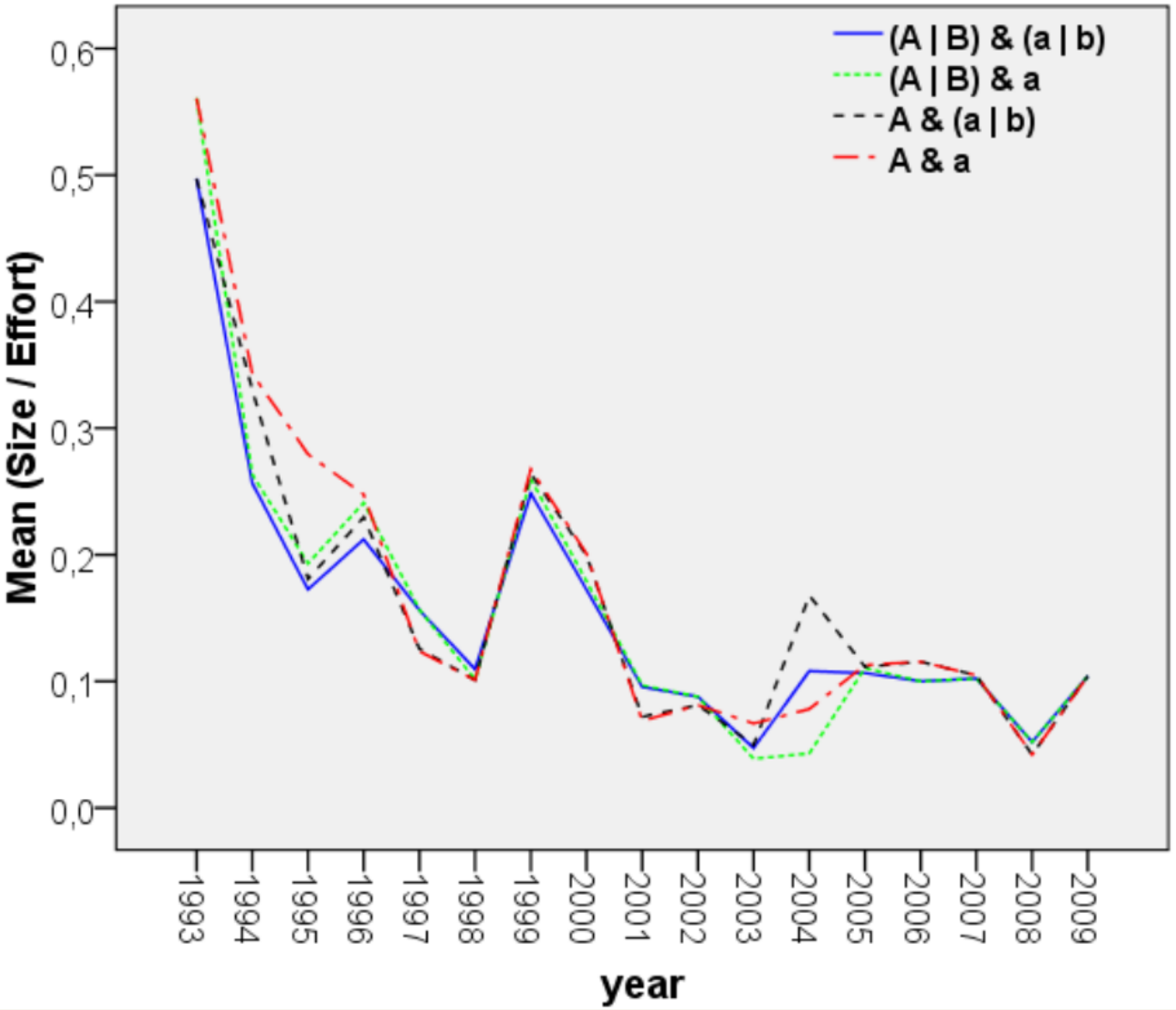
3. THE CASE STUDY: PRODUCTIVITY EVOLUTION OVER TIME

Evolution of mean productivity over time

- **Productivity measured as the ratio Size/Effort**
- **Evolution of mean productivity over time**
  - Figure 1
- **Results**
  - "A & a" plot appears to be the smooth approximation of the other ones

PROMISE
The 2010 International Conference on Predictive Models in Software Engineering

SENSITIVITY OF RESULTS TO DIFFERENT DATA QUALITY META-DATA CRITERIA IN THE SAMPLE SELECTION OF PROJECTS FROM THE ISBSG DATASET



3. THE CASE STUDY: PRODUCTIVITY EVOLUTION OVER TIME

Evolution of Byear over time

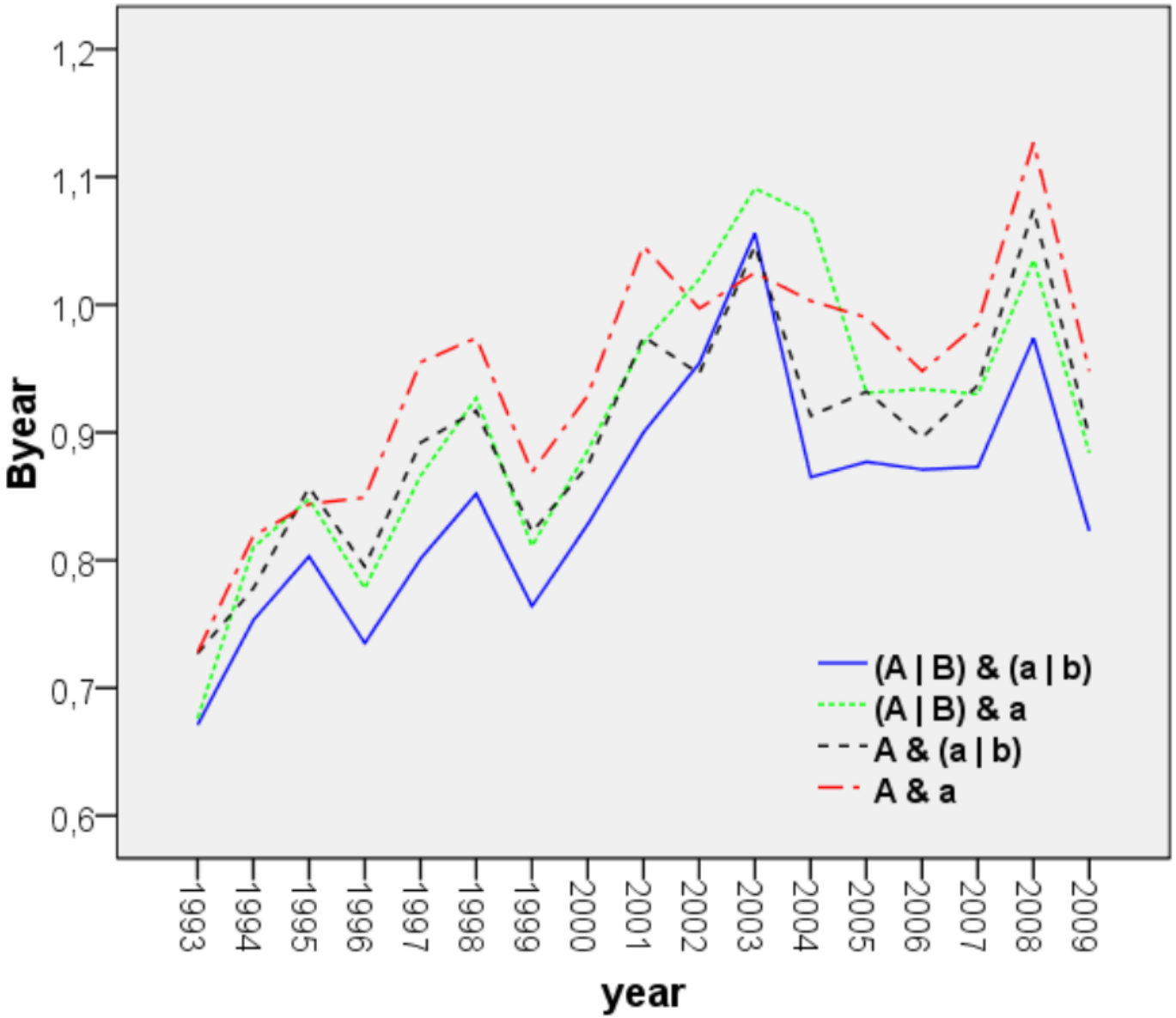
- **Linear regression model relating effort with size**

$$\ln(Effort) = C + B_{93}\ln(Size_{93}) + B_{94}\ln(Size_{94}) + \cdots + B_{09}\ln(Size_{09})$$

- **Evolution of Byear over time**
  - Figure 2
- **Results**
  - Trend towards a decline in productivity, even if variable

PROMISE
The 2010 International Conference on Predictive Models in Software Engineering

SENSITIVITY OF RESULTS TO DIFFERENT DATA QUALITY META-DATA CRITERIA IN THE SAMPLE SELECTION OF PROJECTS FROM THE ISBSG DATASET



3. THE CASE STUDY: PRODUCTIVITY EVOLUTION OVER TIME

Observations derived from the analysis of the 4 selected cases

- **Distribution of projects over time depending on data quality requirements**
  - Major percentage decrease occurs in the interval between 2001 and 2005
- **Mean and variance of productivity ratio depending on data quality requirements**
  - Mean: Both factors statistically significant, but not the interaction between them
  - Variance: only Data Quality Rating

CONCLUSIONS

- **Criteria adopted by other researchers with respect to data quality meta-data**
- **Redundancy between the initial data preparation and data quality meta-data in terms of completeness**
- **Some variability pointed out in the behavior of the models obtained by varying these criteria**
  - Reduction in the sample size
  - Different proportion of removed projects over time
  - Variation in average productivity
- **Interaction and simple effects of both variables in the average and variance of Size/Effort variables?**

PROMISE
The 2010 International Conference on Predictive Models in Software Engineering

SENSITIVITY OF RESULTS TO DIFFERENT DATA QUALITY META-DATA CRITERIA IN THE SAMPLE SELECTION OF PROJECTS FROM THE ISBSG DATASET

Marta Fernández-Diego

Mónica Martínez-Gómez

José-MaríaTorralba-Martínez

UNIVERSIDAD POLITÉCNICA DE VALENCIA

SPAIN