Contents

Preface v
About the Authors xiii
1. Introduction **1**
1.1. Audience and Objective 1
1.2. Scope 1
1.3. Structure 2

PART 1 DATA QUALITY: WHAT IT IS, WHY IT IS IMPORTANT, AND HOW TO ACHIEVE IT

2. What Is Data Quality and Why Should We Care? **7**
2.1. When Are Data of High Quality? 7
2.2. Why Care About Data Quality? 10
2.3. How Do You Obtain High-Quality Data? 11
2.4. Practical Tips 13
2.5. Where Are We Now? 13

3. Examples of Entities Using Data to their Advantage/Disadvantage **17**
3.1. Data Quality as a Competitive Advantage 17
3.2. Data Quality Problems and their Consequences 20
3.3. How Many People Really Live to 100 and Beyond? Views from the United States, Canada, and the United Kingdom 25
3.4. Disabled Airplane Pilots – A Successful Application of Record Linkage 26
3.5. Completeness and Accuracy of a Billing Database: Why It Is Important to the Bottom Line 26
3.6. Where Are We Now? 27

4. Properties of Data Quality and Metrics for Measuring It **29**
4.1. Desirable Properties of Databases/Lists 29
4.2. Examples of Merging Two or More Lists and the Issues that May Arise 31
4.3. Metrics Used when Merging Lists 33
4.4. Where Are We Now? 35

5. Basic Data Quality Tools 37

5.1. Data Elements 37
5.2. Requirements Document 38
5.3. A Dictionary of Tests 39
5.4. Deterministic Tests 40
5.5. Probabilistic Tests 44
5.6. Exploratory Data Analysis Techniques 44
5.7. Minimizing Processing Errors 46
5.8. Practical Tips 46
5.9. Where Are We Now? 48

PART 2 SPECIALIZED TOOLS FOR DATABASE IMPROVEMENT

6. Mathematical Preliminaries for Specialized Data Quality Techniques 51

6.1. Conditional Independence 51
6.2. Statistical Paradigms 53
6.3. Capture–Recapture Procedures and Applications 54

7. Automatic Editing and Imputation of Sample Survey Data 61

7.1. Introduction 61
7.2. Early Editing Efforts 63
7.3. Fellegi–Holt Model for Editing 64
7.4. Practical Tips 65
7.5. Imputation 66
7.6. Constructing a Unified Edit/Imputation Model 71
7.7. Implicit Edits – A Key Construct of Editing Software 73
7.8. Editing Software 75
7.9. Is Automatic Editing Taking Up Too Much Time and Money? 78
7.10. Selective Editing 79
7.11. Tips on Automatic Editing and Imputation 79
7.12. Where Are We Now? 80

8. Record Linkage – Methodology 81

8.1. Introduction 81
8.2. Why Did Analysts Begin Linking Records? 82
8.3. Deterministic Record Linkage 82
8.4. Probabilistic Record Linkage – A Frequentist Perspective 83
8.5. Probabilistic Record Linkage – A Bayesian Perspective 91
8.6. Where Are We Now? 92

9. Estimating the Parameters of the Fellegi–Sunter Record Linkage Model 93
9.1. Basic Estimation of Parameters Under Simple Agreement/Disagreement Patterns 93
9.2. Parameter Estimates Obtained via Frequency-Based Matching 94
9.3. Parameter Estimates Obtained Using Data from Current Files 96
9.4. Parameter Estimates Obtained via the EM Algorithm 97
9.5. Advantages and Disadvantages of Using the EM Algorithm to Estimate *m*- and *u*-probabilities 101
9.6. General Parameter Estimation Using the EM Algorithm 103
9.7. Where Are We Now? 106

10. Standardization and Parsing 107
10.1. Obtaining and Understanding Computer Files 109
10.2. Standardization of Terms 110
10.3. Parsing of Fields 111
10.4. Where Are We Now? 114

11. Phonetic Coding Systems for Names 115
11.1. Soundex System of Names 115
11.2. NYSIIS Phonetic Decoder 119
11.3. Where Are We Now? 121

12. Blocking 123
12.1. Independence of Blocking Strategies 124
12.2. Blocking Variables 125
12.3. Using Blocking Strategies to Identify Duplicate List Entries 126
12.4. Using Blocking Strategies to Match Records Between Two Sample Surveys 128
12.5. Estimating the Number of Matches Missed 130
12.6. Where Are We Now? 130

13. String Comparator Metrics for Typographical Error 131
13.1. Jaro String Comparator Metric for Typographical Error 131
13.2. Adjusting the Matching Weight for the Jaro String Comparator 133
13.3. Winkler String Comparator Metric for Typographical Error 133
13.4. Adjusting the Weights for the Winkler Comparator Metric 134
13.5. Where are We Now? 135

PART 3 RECORD LINKAGE CASE STUDIES

14. Duplicate FHA Single-Family Mortgage Records: A Case Study of Data Problems, Consequences, and Corrective Steps 139
14.1. Introduction 139
14.2. FHA Case Numbers on Single-Family Mortgages 141
14.3. Duplicate Mortgage Records 141
14.4. Mortgage Records with an Incorrect Termination Status 145
14.5. Estimating the Number of Duplicate Mortgage Records 148

15. Record Linkage Case Studies in the Medical, Biomedical, and Highway Safety Areas 151
15.1. Biomedical and Genetic Research Studies 151
15.2. Who goes to a Chiropractor? 153
15.3. National Master Patient Index 154
15.4. Provider Access to Immunization Register Securely (PAiRS) System 155
15.5. Studies Required by the Intermodal Surface Transportation Efficiency Act of 1991 156
15.6. Crash Outcome Data Evaluation System 157

16. Constructing List Frames and Administrative Lists 159
16.1. National Address Register of Residences in Canada 160
16.2. USDA List Frame of Farms in the United States 162
16.3. List Frame Development for the US Census of Agriculture..... 165
16.4. Post-enumeration Studies of US Decennial Census 166

17. Social Security and Related Topics 169
17.1. Hidden Multiple Issuance of Social Security Numbers 169
17.2. How Social Security Stops Benefit Payments after Death........ 173
17.3. CPS–IRS–SSA Exact Match File 175
17.4. Record Linkage and Terrorism 177

PART 4 OTHER TOPICS

18. Confidentiality: Maximizing Access to Micro-data while Protecting Privacy 181
18.1. Importance of High Quality of Data in the Original File 182
18.2. Documenting Public-use Files 183
18.3. Checking Re-identifiability 183
18.4. Elementary Masking Methods and Statistical Agencies 186
18.5. Protecting Confidentiality of Medical Data 193
18.6. More-advanced Masking Methods – Synthetic Datasets 195
18.7. Where Are We Now? 198

19. Review of Record Linkage Software 201
19.1. Government 201
19.2. Commercial 202
19.3. Checklist for Evaluating Record Linkage Software 203

20. Summary Chapter 209

Bibliography 211

Index 221