

Contents

Preface	ix
1 The Statistical Matching Problem	1
1.1 Introduction	1
1.2 The Statistical Framework	3
1.3 The Missing Data Mechanism in the Statistical Matching Problem	6
1.4 Accuracy of a Statistical Matching Procedure	8
1.4.1 Model assumptions	8
1.4.2 Accuracy of the estimator	9
1.4.3 Representativeness of the synthetic file	10
1.4.4 Accuracy of estimators applied on the synthetic data set . .	11
1.5 Outline of the Book	11
2 The Conditional Independence Assumption	13
2.1 The Macro Approach in a Parametric Setting	14
2.1.1 Univariate normal distributions case	15
2.1.2 The multinormal case	19
2.1.3 The multinomial case	23
2.2 The Micro (Predictive) Approach in the Parametric Framework . .	25
2.2.1 Conditional mean matching	26
2.2.2 Draws based on conditional predictive distributions	29
2.2.3 Representativeness of the predicted files	30
2.3 Nonparametric Macro Methods	31
2.4 The Nonparametric Micro Approach	34
2.4.1 Random hot deck	37
2.4.2 Rank hot deck	39
2.4.3 Distance hot deck	40
2.4.4 The matching noise	45
2.5 Mixed Methods	47
2.5.1 Continuous variables	47
2.5.2 Categorical variables	50
2.6 Comparison of Some Statistical Matching Procedures under the CIA	51

2.7	The Bayesian Approach	54
2.8	Other Identifiable Models	56
2.8.1	The pairwise independence assumption	57
2.8.2	Finite mixture models	60
3	Auxiliary Information	65
3.1	Different Kinds of Auxiliary Information	65
3.2	Parametric Macro Methods	68
3.2.1	The use of a complete third file	68
3.2.2	The use of an incomplete third file	70
3.2.3	The use of information on inestimable parameters	71
3.2.4	The multinormal case	73
3.2.5	Comparison of different regression parameter estimators through simulation	76
3.2.6	The multinomial case	81
3.3	Parametric Predictive Approaches	82
3.4	Nonparametric Macro Methods	83
3.5	The Nonparametric Micro Approach with Auxiliary Information	84
3.6	Mixed Methods	85
3.6.1	Continuous variables	85
3.6.2	Comparison between some mixed methods	88
3.6.3	Categorical variables	89
3.7	Categorical Constrained Techniques	92
3.7.1	Auxiliary micro information and categorical constraints	93
3.7.2	Auxiliary information in the form of categorical constraints	94
3.8	The Bayesian Approach	95
4	Uncertainty in Statistical Matching	97
4.1	Introduction	97
4.2	A Formal Definition of Uncertainty	100
4.3	Measures of Uncertainty	105
4.3.1	Uncertainty in the normal case	108
4.3.2	Uncertainty in the multinomial case	111
4.4	Estimation of Uncertainty	117
4.4.1	Maximum likelihood estimation of uncertainty in the multi- normal case	120
4.4.2	Maximum likelihood estimation of uncertainty in the multi- nomial case	121
4.5	Reduction of Uncertainty: Use of Parameter Constraints	124
4.5.1	The multinomial case	126
4.6	Further Aspects of Maximum Likelihood Estimation of Uncertainty	132
4.7	An Example with Real Data	136
4.8	Other Approaches to the Assessment of Uncertainty	140

- 4.8.1 The consistent approach 141
- 4.8.2 The multiple imputation approach 141
- 4.8.3 The de Finetti coherence approach 145
- 5 Statistical Matching and Finite Populations 149**
 - 5.1 Matching Two Archives 150
 - 5.1.1 Definition of the CIA 151
 - 5.2 Statistical Matching and Sampling from a Finite Population 153
 - 5.3 Parametric Methods under the CIA 154
 - 5.3.1 The macro approach when the CIA holds 155
 - 5.3.2 The predictive approach 156
 - 5.4 Parametric Methods when Auxiliary Information is Available 156
 - 5.4.1 The macro approach 156
 - 5.4.2 The predictive approach 158
 - 5.5 File Concatenation 158
 - 5.6 Nonparametric Methods 160
- 6 Issues in Preparing for Statistical Matching 163**
 - 6.1 Reconciliation of Concepts and Definitions of Two Sources 163
 - 6.1.1 Reconciliation of biased sources 165
 - 6.1.2 Reconciliation of inconsistent definitions 167
 - 6.2 How to Choose the Matching Variables 167
- 7 Applications 173**
 - 7.1 Introduction 173
 - 7.2 Case Study: The Social Accounting Matrix 175
 - 7.2.1 Harmonization step 176
 - 7.2.2 Modelling the social accounting matrix 179
 - 7.2.3 Choosing the matching variables 182
 - 7.2.4 The SAM under the CIA 196
 - 7.2.5 The SAM and auxiliary information 199
 - 7.2.6 Assessment of uncertainty for the SAM 202
- A Statistical Methods for Partially Observed Data 205**
 - A.1 Maximum Likelihood Estimation with Missing Data 205
 - A.1.1 Missing data mechanisms 205
 - A.1.2 Maximum likelihood and ignorable nonresponse 206
 - A.2 Bayesian Inference with Missing Data 209
- B Loglinear Models 211**
 - B.1 Maximum Likelihood Estimation of the Parameters 212
- C Distance Functions 215**
- D Finite Population Sampling 219**

E R Code	223
E.1 The R Environment	223
E.2 R Code for Nonparametric Methods	223
E.3 R Code for Parametric and Mixed Methods	231
E.4 R Code for the Study of Uncertainty	240
E.5 Other R Functions	243
References	245
Index	253