

Contents

1	Software	1
1.1	Prerequisites	1
1.1.1	Installation and Updates	2
1.1.2	Install sdcMicro and Its Browser-Based Point-and-Click App	3
1.1.3	Updating the SDC Tools	3
1.1.4	Help	3
1.1.5	The R Workspace and the Working Directory	5
1.1.6	Data Types	5
1.1.7	Generic Functions, Methods and Classes	11
1.2	Brief Overview on SDC Software Tools	14
1.3	Differences Between SDC Tools	15
1.4	Working with sdcMicro	17
1.4.1	General Information About sdcMicro	18
1.4.2	S4 Class Structure of the sdcMicro Package	18
1.4.3	Utility Functions	23
1.4.4	Reporting Facilities	25
1.5	The Point-and-Click App sdcApp	26
1.6	The simPop package	31
	References	33
2	Basic Concepts	35
2.1	Types of Variables	35
2.1.1	Non-confidential Variables	35
2.1.2	Identifying Variables	36
2.1.3	Sensitive Variables	36
2.1.4	Linked Variables	37
2.1.5	Sampling Weights	37
2.1.6	Hierarchies, Clusters and Strata	38
2.1.7	Categorical Versus Continuous Variables	38

2.2	Types of Disclosure	38
2.2.1	Identity Disclosure	39
2.2.2	Attribute Disclosure	39
2.2.3	Inferential Disclosure	40
2.3	Disclosure Risk Versus Information Loss and Data Utility	42
2.4	Release Types	45
2.4.1	Public Use Files (PUF)	45
2.4.2	Scientific Use Files (SUF)	46
2.4.3	Controlled Research Data Center	46
2.4.4	Remote Execution	47
2.4.5	Remote Access	47
	References	48
3	Disclosure Risk	49
3.1	Introduction	49
3.2	Frequency Counts	50
3.2.1	The Number of Cells of Equal Size	51
3.2.2	Frequency Counts with Missing Values	53
3.2.3	Sample Frequencies in sdcMicro	54
3.3	Principles of k -anonymity and l -diversity	58
3.3.1	Simplified Estimation of Population Frequency Counts	60
3.4	Special Uniques Detection Algorithm (SUDA)	67
3.4.1	Minimal Sample Uniqueness	68
3.4.2	SUDA Scores	68
3.4.3	SUDA DIS Scores	69
3.4.4	SUDA in sdcMicro	69
3.5	The Individual Risk Approach	72
3.5.1	The Benedetti-Franconi Model for Risk Estimation	73
3.6	Disclosure Risks for Hierarchical Data	75
3.7	Measuring Global Risks	77
3.7.1	Measuring the Global Risk Using Log-Linear Models:	79
3.7.2	Standard Log-Linear Model	79
3.7.3	Clogg and Eliason Method	79
3.7.4	Pseudo Maximum Likelihood Method	80
3.7.5	Weighted Log-Linear Model	80
3.8	Application of the Log-Linear Models	80
3.9	Global Risk Measures	85
3.10	Quality of the Risk Measures Under Different Sampling Designs	90
3.11	Disclosure Risk for Continuous Variables	91

3.12	Special Treatment of Outliers When Calculating Disclosure Risks	93
	References	96
4	Methods for Data Perturbation	99
4.1	Kind of Methods	99
4.2	Methods for Categorical Key Variables	100
4.2.1	Recoding	100
4.2.2	Local Suppression	103
4.2.3	Post-randomization Method (PRAM)	116
4.3	Methods for Continuous Key Variables	119
4.3.1	Microaggregation	119
4.3.2	Noise Addition	125
4.3.3	Shuffling	130
	References	132
5	Data Utility and Information Loss	133
5.1	Element-Wise Comparisons	133
5.1.1	Comparing Missing Values	133
5.1.2	Comparing Aggregated Information	134
5.2	Element-Wise Measures for Continuous Variables	139
5.2.1	Element-Wise Comparisons of Mixed Scaled Variables	143
5.3	Entropy	144
5.4	Propensity Score Methods	145
5.5	Quality Indicators	148
5.5.1	General Procedure	148
5.5.2	Differences in Point Estimates	149
5.5.3	Differences in Variances and MSE	150
5.5.4	Overlap in Confidence Intervals	151
5.5.5	Differences in Model Estimates	153
	References	155
6	Synthetic Data	157
6.1	Introduction	157
6.2	Model-Based Generation of Synthetic Data	159
6.2.1	Setup of the Structure	161
6.2.2	Simulation of Categorical Variables	162
6.2.3	Simulation of Continuous Variables	164
6.2.4	Splitting Continuous Variables into Components	168
6.3	Disclosure Risk of Synthetic Data	169
6.3.1	Confidentiality of Synthetic Population Data	171
6.3.2	Disclosure Scenarios for Synthetic Population Data	172
6.4	Data Utility of Synthetic Data	176
	References	178

7 Practical Guidelines	181
7.1 The Workflow	181
7.2 How to Determine the Key Variables	182
7.3 The Level of Disclosure Risk Versus Information Loss	183
7.4 Which SDC Methods Should Be Used	183
References	186
8 Case Studies	187
8.1 Practical Issues	187
8.2 Anonymization of the FIES Data	188
8.2.1 FIES Data Description	188
8.2.2 Pre-processing Steps	189
8.2.3 Frequency Counts and Disclosure Risk	190
8.2.4 Recoding	191
8.2.5 Local Suppression	192
8.2.6 Perturbing the Continuous Key Variables	193
8.2.7 PRAM	194
8.2.8 Remark	195
8.3 Application to the Structural Earnings Statistics (SES) Survey	195
8.3.1 General Information About SES	195
8.3.2 Details on Some Variables	196
8.3.3 Applications and Statistics Based on SES	198
8.3.4 The Synthetic SES Data	199
8.3.5 Key Variables for Re-identification	199
8.3.6 Pre-processing Steps	200
8.3.7 Risk Estimation	201
8.3.8 Perturbing the Continuous Scaled Variables	203
8.3.9 Measuring the Data Utility	204
8.4 I2D2	213
8.4.1 About I2D2 Data	213
8.4.2 Disclosure Scenario/Key Variables	213
8.4.3 Anonymization of One Example Country	214
8.4.4 Results for All Other Countries	219
8.4.5 Data Utility	221
8.5 Anonymization of P4 Data	222
8.5.1 Key Variables	222
8.5.2 Key Variables on Individual Level	223
8.5.3 sdcMicro Code for One Example Country	223
8.5.4 Results for All Other Countries	226
8.6 Anonymization of the SHIP Data	227
8.6.1 Key Variables	228
8.6.2 sdcMicro Code for One Example Country	229
8.6.3 Results for All Other Countries	233

8.7 A Synthetic Socio-economic Population and Sample	236
8.7.1 Data Preprocessing	237
8.7.2 Simulation of the Population	243
8.7.3 Optionally: Draw a Sample from the Population	250
8.7.4 Exploration of the Final Synthetic Population and Sample	251
References	258
Software Versions Used in the Book	261
Solutions	263
Index	285

CRAN	Comprehensive R Archive Network
CSS	Complex Survey Samples
EU-SILC	European Statistics on Income and Living Conditions
FIES	Family Income and Expenditure Survey
GUI	Point-and-click graphical user interface
I2D2	International Income Distribution Database
IHSN	International Household Survey Network
MCD	Minimum Covariance Determinant
MDAV	Maximum Distance to Average Vector
NSI	National Statistical Institutes
PP	Purchase Power Parity Data
PRAM	Post Randomization Method
RMD	Robust Mahalanobis Distance
ROMM	Random Orthogonal Matrix Masking
SDC	Statistical Disclosure Control
SES	Structural Earning Statistics
SHIP	Survey-Based Harmonized Indicator Program
SDA	Special Uniqueness Detection Algorithm

The book does not contain a glossary on Statistical Disclosure Control, because the important keywords are explained in Chap. 2. Moreover, glossaries on this topic can be found on the Internet, e.g., at <http://econ.vb.chs.at/cso/Glossary.htm>.