Contents

Preface xi
Acknowledgements xv

1 Introduction 1
1.1 Concepts and definitions 2
1.1.1 Disclosure 2
1.1.2 Statistical disclosure control 3
1.1.3 Tabular data 3
1.1.4 Microdata 3
1.1.5 Risk and utility 4
1.2 An approach to Statistical Disclosure Control 7
1.2.1 Why is confidentiality protection needed? 7
1.2.2 What are the key characteristics and uses of the data? 8
1.2.3 What disclosure risks need to be protected against? 8
1.2.4 Disclosure control methods 8
1.2.5 Implementation 9
1.3 The chapters of the handbook 9

2 Ethics, principles, guidelines and regulations – a general background 10
2.1 Introduction 10
2.2 Ethical codes and the new ISI code 11
2.2.1 ISI Declaration on Professional Ethics 11
2.2.2 New ISI Declaration on Professional Ethics 12
2.2.3 European Statistics Code of Practice 15
2.3 UNECE principles and guidelines 16
2.3.1 UNECE Principles and Guidelines on Confidentiality Aspects of Data Integration 18
2.3.2 Future activities on the UNECE principles and guidelines 19
2.4 Laws 19
2.4.1 Committee on Statistical Confidentiality 20
2.4.2 European Statistical System Committee 20
3 Microdata

3.1 Introduction

3.2 Microdata concepts

3.2.1 Stage 1: Assess need for confidentiality protection

3.2.2 Stage 2: Key characteristics and use of microdata

3.2.3 Stage 3: Disclosure risk

3.2.4 Stage 4: Disclosure control methods

3.2.5 Stage 5: Implementation

3.3 Definitions of disclosure

3.3.1 Definitions of disclosure scenarios

3.4 Definitions of disclosure risk

3.4.1 Disclosure risk for categorical quasi-identifiers

3.4.2 Notation and assumptions

3.4.3 Disclosure risk for continuous quasi-identifiers

3.5 Estimating re-identification risk

3.5.1 Individual risk based on the sample: Threshold rule

3.5.2 Estimating individual risk using sampling weights

3.5.3 Estimating individual risk by Poisson model

3.5.4 Further models that borrow information from other sources

3.5.5 Estimating per record risk via heuristics

3.5.6 Assessing risk via record linkage

3.6 Non-perturbative microdata masking

3.6.1 Sampling

3.6.2 Global recoding

3.6.3 Top and bottom coding

3.6.4 Local suppression

3.7 Perturbative microdata masking

3.7.1 Additive noise masking

3.7.2 Multiplicative noise masking

3.7.3 Microaggregation

3.7.4 Data swapping and rank swapping

3.7.5 Data shuffling

3.7.6 Rounding

3.7.7 Re-sampling

3.7.8 PRAM

3.7.9 MASSC

3.8 Synthetic and hybrid data

3.8.1 Fully synthetic data

3.8.2 Partially synthetic data

3.8.3 Hybrid data

3.8.4 Pros and cons of synthetic and hybrid data

3.9 Information loss in microdata

3.9.1 Information loss measures for continuous data

3.9.2 Information loss measures for categorical data
CONTENTS vii

3.10 Release of multiple files from the same microdata set 110
3.11 Software 111
  3.11.1 μ-ARGUS 111
  3.11.2 sdcMicro 113
  3.11.3 IVEware 115
3.12 Case studies 116
  3.12.1 Microdata files at Statistics Netherlands 116
  3.12.2 The European Labour Force Survey microdata for research purposes 118
  3.12.3 The European Structure of Earnings Survey microdata for research purposes 121
  3.12.4 NHIS-linked mortality data public use file, USA 128
  3.12.5 Other real case instances 130

4 Magnitude tabular data 131
  4.1 Introduction 131
    4.1.1 Magnitude tabular data: Basic terminology 131
    4.1.2 Complex tabular data structures: Hierarchical and linked tables 132
    4.1.3 Risk concepts 134
    4.1.4 Protection concepts 137
    4.1.5 Information loss concepts 137
    4.1.6 Implementation: Software, guidelines and case study 138
  4.2 Disclosure risk assessment I: Primary sensitive cells 138
    4.2.1 Intruder scenarios 138
    4.2.2 Sensitivity rules 140
  4.3 Disclosure risk assessment II: Secondary risk assessment 152
    4.3.1 Feasibility interval 152
    4.3.2 Protection level 154
    4.3.3 Singleton and multi cell disclosure 155
    4.3.4 Risk models for hierarchical and linked tables 155
  4.4 Non-perturbative protection methods 157
    4.4.1 Global recoding 157
    4.4.2 The concept of cell suppression 157
    4.4.3 Algorithms for secondary cell suppression 158
    4.4.4 Secondary cell suppression in hierarchical and linked tables 161
  4.5 Perturbative protection methods 163
    4.5.1 A pre-tabular method: Multiplicative noise 165
    4.5.2 A post-tabular method: Controlled tabular adjustment 165
  4.6 Information loss measures for tabular data 166
    4.6.1 Cell costs for cell suppression 166
    4.6.2 Cell costs for CTA 167
    4.6.3 Information loss measures to evaluate the outcome of table protection 167
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.7</td>
<td>Software for tabular data protection</td>
<td>168</td>
</tr>
<tr>
<td>4.7.1</td>
<td>Empirical comparison of cell suppression algorithms</td>
<td>169</td>
</tr>
<tr>
<td>4.8</td>
<td>Guidelines: Setting up an efficient table model systematically</td>
<td>173</td>
</tr>
<tr>
<td>4.8.1</td>
<td>Defining spanning variables</td>
<td>174</td>
</tr>
<tr>
<td>4.8.2</td>
<td>Response variables and mapping rules</td>
<td>175</td>
</tr>
<tr>
<td>4.9</td>
<td>Case studies</td>
<td>178</td>
</tr>
<tr>
<td>4.9.1</td>
<td>Response variables and mapping rules of the case study</td>
<td>178</td>
</tr>
<tr>
<td>4.9.2</td>
<td>Spanning variables of the case study</td>
<td>179</td>
</tr>
<tr>
<td>4.9.3</td>
<td>Analysing the tables of the case study</td>
<td>179</td>
</tr>
<tr>
<td>4.9.4</td>
<td>Software issues of the case study</td>
<td>181</td>
</tr>
<tr>
<td>5</td>
<td>Frequency tables</td>
<td>183</td>
</tr>
<tr>
<td>5.1</td>
<td>Introduction</td>
<td>183</td>
</tr>
<tr>
<td>5.2</td>
<td>Disclosure risks</td>
<td>184</td>
</tr>
<tr>
<td>5.2.1</td>
<td>Individual attribute disclosure</td>
<td>185</td>
</tr>
<tr>
<td>5.2.2</td>
<td>Group attribute disclosure</td>
<td>186</td>
</tr>
<tr>
<td>5.2.3</td>
<td>Disclosure by differencing</td>
<td>187</td>
</tr>
<tr>
<td>5.2.4</td>
<td>Perception of disclosure risk</td>
<td>190</td>
</tr>
<tr>
<td>5.3</td>
<td>Methods</td>
<td>191</td>
</tr>
<tr>
<td>5.3.1</td>
<td>Pre-tabular</td>
<td>191</td>
</tr>
<tr>
<td>5.3.2</td>
<td>Table re-design</td>
<td>192</td>
</tr>
<tr>
<td>5.3.3</td>
<td>Post-tabular</td>
<td>193</td>
</tr>
<tr>
<td>5.4</td>
<td>Post-tabular methods</td>
<td>193</td>
</tr>
<tr>
<td>5.4.1</td>
<td>Cell suppression</td>
<td>193</td>
</tr>
<tr>
<td>5.4.2</td>
<td>ABS cell perturbation</td>
<td>193</td>
</tr>
<tr>
<td>5.4.3</td>
<td>Rounding</td>
<td>194</td>
</tr>
<tr>
<td>5.5</td>
<td>Information loss</td>
<td>199</td>
</tr>
<tr>
<td>5.6</td>
<td>Software</td>
<td>201</td>
</tr>
<tr>
<td>5.6.1</td>
<td>Introduction</td>
<td>201</td>
</tr>
<tr>
<td>5.6.2</td>
<td>Optimal, first feasible and RAPID solutions</td>
<td>202</td>
</tr>
<tr>
<td>5.6.3</td>
<td>Protection provided by controlled rounding</td>
<td>203</td>
</tr>
<tr>
<td>5.7</td>
<td>Case studies</td>
<td>204</td>
</tr>
<tr>
<td>5.7.1</td>
<td>UK Census</td>
<td>204</td>
</tr>
<tr>
<td>5.7.2</td>
<td>Australian and New Zealand Censuses</td>
<td>205</td>
</tr>
<tr>
<td>6</td>
<td>Data access issues</td>
<td>208</td>
</tr>
<tr>
<td>6.1</td>
<td>Introduction</td>
<td>208</td>
</tr>
<tr>
<td>6.2</td>
<td>Research data centres</td>
<td>209</td>
</tr>
<tr>
<td>6.3</td>
<td>Remote execution</td>
<td>209</td>
</tr>
<tr>
<td>6.4</td>
<td>Remote access</td>
<td>210</td>
</tr>
<tr>
<td>6.5</td>
<td>Licensing</td>
<td>211</td>
</tr>
<tr>
<td>6.6</td>
<td>Guidelines on output checking</td>
<td>211</td>
</tr>
<tr>
<td>6.6.1</td>
<td>Introduction</td>
<td>211</td>
</tr>
<tr>
<td>6.6.2</td>
<td>General approach</td>
<td>212</td>
</tr>
<tr>
<td>6.6.3</td>
<td>Rules for output checking</td>
<td>215</td>
</tr>
</tbody>
</table>
CONTENTS

6.6.4 Organisational/procedural aspects of output checking 224
6.6.5 Researcher training 233

6.7 Additional issues concerning data access 236
6.7.1 Examples of disclaimers 236
6.7.2 Output description 236

6.8 Case studies 237
6.8.1 The US Census Bureau Microdata Analysis System 237
6.8.2 Remote access at Statistics Netherlands 239

Glossary 243

References 261

Author index 279

Subject index 282