Universitat Politècnica de Catalunya
Departament d'Estadística i Investigació Operativa

Tesi Doctoral

Contribucions a la Microagregació per a la Protecció de Dades Estadístiques

Autor: Àngel Torres Aragó

Directors: Dr. Josep Maria Mateo Sanz
Dr. Josep Domingo Ferrer

Tutor: Dr. Tomàs Aluja i Banet

Abril 2003

Certifiquem que hem llegit aquesta tesi i que, al nostre parer, és plenament adequada, en continguts i qualitat, com a tesi per obtenir el títol de Doctor.

Dr. Josep Maria Mateo Sanz
Dr. Josep Domingo Ferrer
(Directors)

Certifico que, al meu parer, aquesta tesi és plenament adequada, en continguts i qualitat, com a tesi per obtenir el títol de Doctor.

Dr. Tomàs Aluja i Banet
(Tutor)

Aprovada per la Comissió de Doctorat:

Als meus pares i a Maria-Josep.

Agraïments

Al Dr. Josep Domingo i Ferrer, per haver-me animat a començar els estudis de doctorat, oferint-me el seu grup de recerca dintre el camp de la protecció de dades estadístiques confidencials. El seu suport i dedicació han fet créixer dins de mi una forta il·lusió per la recerca que ha orientat tota la tesi.

Al Dr. Josep-Maria Mateo i Sanz, perquè va saber incloure'm en els seus projectes de recerca dintre d'una tècnica més específica com és la microagregació de dades estadístiques. Certament el seu treball i pacient guiatge han fet arribar a bon port aquesta tesi.

Als membres del Departament d'Estadística i Investigació Operativa de la Universitat Politècnica de Catalunya que van acollir de molt bon grat aquesta tesi quan encara era una petita llavor. Especialment m'agradaria expressar el meu agraïment al Dr. Tomàs Aluja i Banet per haver acceptat de ser el tutor d'aquesta tesi, i al Dr. Manuel Martí i Recober, que, com a responsable acadèmic del Programa de Doctorat, sempre m'ha donat empenta.

A Antoni Martínez i al Dr. Francesc Sebé, membres també del nostre grup de treball perquè la seva important col·laboració, mitjançant la implementació de programes informàtics per microagregar i calcular mesures de qualitat, ha contribuït a la culminació dels resultats del present projecte.

A Maria-Josep Maigí, que m'ha encoratjat durant tots aquests anys, especialment en els moments de més desànim, confiant fermament en un projecte que, de vegades, em semblava esvair-se.

Finalment voldria que tot aquest agraïment embolcallés també altres persones que m'han ajudat de molt diverses maneres i que aquí no esmento.

Cal esmentar que el desenvolupament d'aquesta tesi ha tingut lloc en el marc del projecte CASC ("Computational Aspects of Statistical Confidentiality", IST-2000-C02-02), del 5è Programa Marc de la Comissió Europea, alhora que ha rebut l'empar del projecte del Ministeri Espanyol de Ciència i Tecnologia TEL98-0699-C02-02.

Tarragona, febrer del 2003.

Contingut

Agraïments **v**

1 Introducció **1**

1.1 Objectius 1

1.2 Aportacions 2

1.3 Estructura general de la memòria 3

2 Conceptes bàsics **5**

2.1 Classificació dels estadístics 5

2.2 El problema de la revelació 7

2.2.1 Revelació amb macrodades 7

2.2.2 Revelació amb microdades 8

2.3 Criteris d'avaluació 9

2.4 Classificació dels mètodes de control de la revelació per a microdades 11

2.5 Notació i formalització 12

2.5.1 Variables contínues i variables categòriques 13

2.6 Mètodes de reducció de dades 13

2.6.1 Anonimització 13

2.6.2 Mètodes de mostreig 14

2.6.3 Restricció de la mida de la població 14

2.6.4 Reducció de detalls 14

2.6.5 Codificació global 15

2.6.6 Supressió 15

2.7 Mètodes Pertorbatius 16

2.7.1 Pertorbació additiva aleatòria 16

2.7.2 Pertorbació segons una distribució de probabilitat 17

2.7.3 Remostreig 18

2.7.4 Intercanvi de dades 19

2.7.5 PRAM . . . 19
2.7.6 Pèrdua per compressió . . . 19
2.7.7 Microagregació . . . 20

3 Seguretat de la microagregació amb ordenació individual 23
3.1 Conceptes sobre microagregació . . . 23
3.2 Seguretat de la microagregació utilitzant ordenació individual . . . 24
3.3 Estudi analític per a dades contínues uniformement distribuïdes . . . 27
3.4 Estudi de simulació per a la distribució Normal . . . 29
3.5 Estudi de simulació per a dades esbiaixades (Weibull) . . . 31
3.6 Conclusions . . . 32
3.7 Taules de resultats . . . 33

4 Mètodes DM i DMM per a microagregació multivariant 39
4.1 Microagregació Univariant i Multivariant . . . 39
4.2 Mètode DM de la Distància Màxima . . . 40
4.3 Mètode DMM de la Distància Màxima Modificat . . . 41
4.4 Exemple d'aplicació dels mètodes DM i DMM . . . 43
4.5 Implementació i complexitat computacional dels mètodes DM i DMM . . . 44
4.5.1 Implementació dels mètodes DM i DMM . . . 44
4.5.2 Càlcul del número de distàncies i complexitat computacional de l'algorisme DM sense emmagatzemar la matriu de distàncies . . . 45
4.5.3 Càlcul del número de distàncies i complexitat computacional de l'algorisme DMM sense emmagatzemar la matriu de distàncies . . . 48
4.5.4 Conclusió . . . 50

5 Mesures de qualitat per comparar mètodes pertorbatius 51
5.1 Qualitat d'un mètode de control de la revelació pertorbatiu . . . 51
5.2 Caracterització d'un conjunt de microdades contínues . . . 52
5.2.1 Mesures per a la caracterització . . . 53
5.2.2 Format de les matrius . . . 53
5.3 Mesures per a la pèrdua d'informació . . . 55
5.3.1 Mesures per a les discrepàncies $\mathbf{X} - \mathbf{X}'$ . . . 55
5.3.2 Mesures per a les discrepàncies $\overline{\mathbf{X}} - \overline{\mathbf{X}}'$ . . . 56
5.3.3 Mesures per a les discrepàncies $\mathbf{V} - \mathbf{V}'$ . . . 56
5.3.4 Mesures per a les discrepàncies $\mathbf{R} - \mathbf{R}'$ . . . 57
5.3.5 Mesures per a les discrepàncies $\mathbf{RF} - \mathbf{RF}'$ . . . 57
5.3.6 Mesures per a les discrepàncies $\mathbf{C} - \mathbf{C}'$ . . . 57

5.3.7 Mesures per a les discrepàncies $\mathbf{F} - \mathbf{F}'$ . . . 58
5.4 Mesures per a la pèrdua de confidencialitat . . . 58
5.4.1 Escenaris de revelació . . . 60
5.5 Mesura global per comparar mètodes pertorbatius . . . 61

6 Comparació de mètodes pertorbatius 63
6.1 Mètodes i paràmetres usats . . . 63
6.2 Descripció del conjunt de dades . . . 65
6.2.1 Procediment d'extracció de dades . . . 65
6.2.2 Selecció de variables . . . 66
6.2.3 Selecció de registres . . . 66
6.3 Pèrdua d'informació: Mesures utilitzades . . . 67
6.4 Risc de revelació: Mesures utilitzades . . . 67
6.5 Resultats de l'estudi comparatiu i conclusions . . . 69
6.5.1 Mesura global sobre la qualitat d'un mètode pertorbatiu . . . 69
6.5.2 Taules incloses a l'apèndix . . . 69
6.5.3 Estudi dels mètodes . . . 70
6.5.4 Conclusions . . . 81

7 DMM. Particions del conjunt de variables: nombre de variables 83
7.1 Microagregació multivariant i nombre de variables . . . 83
7.2 Particions del conjunt V de variables . . . 84
7.2.1 Nombre de particions per al cas de 13 variables . . . 86
7.3 Estudi sobre el nombre de variables . . . 88
7.3.1 Nombre de variables i pèrdua d'informació . . . 88
7.3.2 Nombre de variables i ERD-pèrdua de confidencialitat . . . 90
7.3.3 Nombre de variables i ICN-pèrdua de confidencialitat . . . 92
7.3.4 Nombre de variables i ICD-pèrdua de confidencialitat . . . 94
7.3.5 Nombre de variables i mesura global **MG** sobre la qualitat . . . 98

8 Estudi de la combinació de variables 99
8.1 Necessitat de mesures prèvies per a cada combinació de variables . . . 99
8.2 Estadístics per al control de la revelació . . . 100
8.2.1 Estadístics Intra-grup . . . 101
8.2.2 Estadístics Inter-grup . . . 104
8.2.3 Notació i Simbologia . . . 108
8.3 Resultats i conclusions de l'estudi sobre la combinació de variables . . . 108

8.3.1 Combinació de variables i pèrdua d'informació 110
8.3.2 Combinació de variables i ERD-pèrdua de confidencialitat 111
8.3.3 Combinació de variables i ICN-pèrdua de confidencialitat 111
8.3.4 Combinació de variables i ICD-pèrdua de confidencialitat 112
8.3.5 Combinació de variables i mesura global MG sobre la qualitat 113

9 Conclusions i recerca futura **115**
9.1 Conclusions 115
9.2 Ampliacions i futura recerca 117

A Taules (comparació de mètodes pertorbatius) **119**

B Correlacions entre estadístics i mesures amb diverses particions de variables **141**

Bibliografia **163**

Llista de Taules

2.1 Exemple de microdades 6

3.1 Mitjana del percentatge de l'amplitud relativa de l'interval (3.4) sobre $[a_{-1}, a_1]$ (dades uniformes) 33

3.2 Percentatge de l'amplitud relativa de l'interval (3.7) sobre $[a_{-1}, a_0]$ (dades uniformes) 34

3.3 Percentatge de l'amplitud relativa de l'interval (3.8) sobre $[a_0, a_1]$ (dades uniformes) . 34

3.4 Amplitud de $[a_{-1}, a_1]$ per a diferents grups i valors n, k (dades normals) 35

3.5 Mitjana del percentatge de l'amplitud relativa de l'interval (3.4) sobre $[a_{-1}, a_1]$ (dades normals) 35

3.6 Percentatge de l'amplitud relativa de l'interval (3.7) sobre $[a_{-1}, a_0]$ (dades normals) . 36

3.7 Percentatge de l'amplitud relativa de l'interval (3.8) sobre $[a_0, a_1]$ (dades normals) . . 36

3.8 Amplitud de $[a_{-1}, a_1]$ per a diferents grups i valors n, k (dades Weibull) 37

3.9 Mitjana del percentatge de l'amplitud relativa de l'interval (3.4) sobre $[a_{-1}, a_1]$ (dades Weibull) 37

3.10 Percentatge de l'amplitud relativa de l'interval (3.7) sobre $[a_{-1}, a_0]$ (dades Weibull) . . 38

3.11 Percentatge de l'amplitud relativa de l'interval (3.8) sobre $[a_0, a_1]$ (dades Weibull) . . 38

7.1 Mesures de pèrdua d'informació obtingudes en aplicar les diverses versions de microagregació multivariant amb $k = 3$. 89

7.2 Mesures de pèrdua d'informació obtingudes en aplicar les diverses versions de microagregació multivariant amb $k = 10$. 89

7.3 Mesures de pèrdua d'informació obtingudes en aplicar les diverses versions de microagregació multivariant amb $k = 20$. 89

7.4 Mesures de pèrdua de confidencialitat per enllaç de registres obtingudes en aplicar les diverses versions de microagregació multivariant amb $k = 3$. 90

7.5 Mesures de pèrdua de confidencialitat per enllaç de registres obtingudes en aplicar les diverses versions de microagregació multivariant amb $k = 10$. 91

7.6 Mesures de pèrdua de confidencialitat per enllaç de registres obtingudes en aplicar les diverses versions de microagregació multivariant amb $k = 20$. 91

7.7 Mesures de pèrdua de confidencialitat per intervals de confiança, basat en rangs, obtingudes en aplicar les diverses versions de microagregació multivariant amb $k = 3$. 92

7.8 Mesures de pèrdua de confidencialitat per intervals de confiança, basats en rangs, obtingudes en aplicar les diverses versions de microagregació multivariant amb $k = 10$. 93

7.9 Mesures de pèrdua de confidencialitat per intervals de confiança, basats en rangs, obtingudes en aplicar les diverses versions de microagregació multivariant amb $k = 20$. 93

7.10 Mesures de pèrdua de confidencialitat per intervals de confiança, basats en desviacions típiques, obtingudes en aplicar les diverses versions de microagregació multivariant amb $k = 3$. 94

7.11 Mesures de pèrdua de confidencialitat per intervals de confiança, basats en desviacions típiques, obtingudes en aplicar les diverses versions de microagregació multivariant amb $k = 10$. 95

7.12 Mesures de pèrdua de confidencialitat per intervals de confiança, basats en desviacions típiques, obtingudes en aplicar les diverses versions de microagregació multivariant amb $k = 20$. 95

7.13 Mesures generals obtingudes en aplicar les diverses versions de microagregació multivariant amb $k = 3$. 96

7.14 Mesures generals obtingudes en aplicar les diverses versions de microagregació multivariant amb $k = 10$. 97

7.15 Mesures generals obtingudes en aplicar les diverses versions de microagregació multivariant amb $k = 20$. 97

A.1 Mesures generals obtingudes en aplicar els diversos mètodes (1/5). 119

A.1 Mesures generals obtingudes en aplicar els diversos mètodes (2/5). 120

A.1 Mesures generals obtingudes en aplicar els diversos mètodes (3/5). 121

A.1 Mesures generals obtingudes en aplicar els diversos mètodes (4/5). 122

A.1 Mesures generals obtingudes en aplicar els diversos mètodes (5/5). 123

A.2 Mesures de pèrdua d'informació obtingudes en aplicar els diversos mètodes (1/5). . . 123

A.2 Mesures de pèrdua d'informació obtingudes en aplicar els diversos mètodes (2/5). . . 124

A.2 Mesures de pèrdua d'informació obtingudes en aplicar els diversos mètodes (3/5). . . 125

A.2 Mesures de pèrdua d'informació obtingudes en aplicar els diversos mètodes (4/5). . . 126

A.2 Mesures de pèrdua d'informació obtingudes en aplicar els diversos mètodes (5/5). . . 127

A.3 Mesures de pèrdua de confidencialitat per enllaç de registres obtingudes en aplicar els diversos mètodes (1/5). 127

A.3 Mesures de pèrdua de confidencialitat per enllaç de registres obtingudes en aplicar els diversos mètodes (2/5). 128

A.3 Mesures de pèrdua de confidencialitat per enllaç de registres obtingudes en aplicar els diversos mètodes (3/5). 129

A.3 Mesures de pèrdua de confidencialitat per enllaç de registres obtingudes en aplicar els diversos mètodes (4/5). 130

A.3 Mesures de pèrdua de confidencialitat per enllaç de registres obtingudes en aplicar els diversos mètodes (5/5). 131

A.4 Mesures de pèrdua de confidencialitat per intervals de confiança, basats en rangs, obtingudes en aplicar diversos mètodes (1/5). 131

A.4 Mesures de pèrdua de confidencialitat per intervals de confiança, basats en rangs, obtingudes en aplicar diversos mètodes (2/5). 132

A.4 Mesures de pèrdua de confidencialitat per intervals de confiança, basats en rangs, obtingudes en aplicar diversos mètodes (3/5). 133

A.4 Mesures de pèrdua de confidencialitat per intervals de confiança, basats en rangs, obtingudes en aplicar diversos mètodes (4/5). 134

A.4 Mesures de pèrdua de confidencialitat per intervals de confiança, basats en rangs, obtingudes en aplicar diversos mètodes (5/5). 135

A.5 Mesures de pèrdua de confidencialitat per intervals de confiança, basats en desviacions típiques, obtingudes en aplicar els diferents mètodes (1/5). 135

A.5 Mesures de pèrdua de confidencialitat per intervals de confiança, basats en desviacions típiques, obtingudes en aplicar els diferents mètodes (2/5). 136

A.5 Mesures de pèrdua de confidencialitat per intervals de confiança, basats en desviacions típiques, obtingudes en aplicar els diferents mètodes (3/5). 137

A.5 Mesures de pèrdua de confidencialitat per intervals de confiança, basats en desviacions típiques, obtingudes en aplicar els diferents mètodes (4/5). 138

A.5 Mesures de pèrdua de confidencialitat per intervals de confiança, basats en desviacions típiques, obtingudes en aplicar els diferents mètodes (5/5). 139

B.1 Suma estandarditzada i correlacions entre els diversos estadístics i les mesures de pèrdua d'informació obtingudes en aplicar les diverses versions de microagregació (1/5). . . . 141

B.1 Suma estandarditzada i correlacions entre els diversos estadístics i les mesures de pèrdua d'informació obtingudes en aplicar les diverses versions de microagregació (2/5). . . . 142

B.1 Suma estandarditzada i correlacions entre els diversos estadístics i les mesures de pèrdua d'informació obtingudes en aplicar les diverses versions de microagregació (3/5). . . . 143

B.1 Suma estandarditzada i correlacions entre els diversos estadístics i les mesures de pèrdua d'informació obtingudes en aplicar les diverses versions de microagregació (4/5). . . . 144

B.1 Suma estandarditzada i correlacions entre els diversos estadístics i les mesures de pèrdua d'informació obtingudes en aplicar les diverses versions de microagregació (5/5). . . . 145

B.2 Suma estandarditzada i correlacions entre els diversos estadístics i les mesures de pèrdua de confidencialitat per enllaç de registres obtingudes en aplicar les diverses versions de microagregació multivariant amb k=10 (1/5). 145

B.2 Suma estandarditzada i correlacions entre els diversos estadístics i les mesures de pèrdua de confidencialitat per enllaç de registres obtingudes en aplicar les diverses versions de microagregació multivariant amb k=10 (2/5). 146

B.2 Suma estandarditzada i correlacions entre els diversos estadístics i les mesures de pèrdua de confidencialitat per enllaç de registres obtingudes en aplicar les diverses versions de microagregació multivariant amb k=10 (3/5). 147

B.2 Suma estandarditzada i correlacions entre els diversos estadístics i les mesures de pèrdua de confidencialitat per enllaç de registres obtingudes en aplicar les diverses versions de microagregació multivariant amb k=10 (4/5). 148

B.2 Suma estandarditzada i correlacions entre els diversos estadístics i les mesures de pèrdua de confidencialitat per enllaç de registres obtingudes en aplicar les diverses versions de microagregació multivariant amb k=10 (5/5). 149

B.3 Suma estandarditzada i correlacions entre els diversos estadístics i les mesures de pèrdua de confidencialitat per intervals de confiança, basats en rangs, obtingudes en aplicar les diverses versions de microagregació multivariant amb k=10 (1/5). 149

B.3 Suma estandarditzada i correlacions entre els diversos estadístics i les mesures de pèrdua de confidencialitat per intervals de confiança, basats en rangs, obtingudes en aplicar les diverses versions de microagregació multivariant amb k=10 (2/5). 150

B.3 Suma estandarditzada i correlacions entre els diversos estadístics i les mesures de pèrdua de confidencialitat per intervals de confiança, basats en rangs, obtingudes en aplicar les diverses versions de microagregació multivariant amb k=10 (3/5). 151

B.3 Suma estandarditzada i correlacions entre els diversos estadístics i les mesures de pèrdua de confidencialitat per intervals de confiança, basats en rangs, obtingudes en aplicar les diverses versions de microagregació multivariant amb k=10 (4/5). 152

B.3 Suma estandarditzada i correlacions entre els diversos estadístics i les mesures de pèrdua de confidencialitat per intervals de confiança, basats en rangs, obtingudes en aplicar les diverses versions de microagregació multivariant amb k=10 (5/5). 153

B.4 Suma estandarditzada i correlacions entre els diversos estadístics i les mesures de pèrdua de confidencialitat per intervals de confiança, basats en desviacions típiques, obtingudes en aplicar les diverses versions de microagregació multivariant amb k=10 (1/5). . . . 153

B.4 Suma estandarditzada i correlacions entre els diversos estadístics i les mesures de pèrdua de confidencialitat per intervals de confiança, basats en desviacions típiques, obtingudes en aplicar les diverses versions de microagregació multivariant amb k=10 (2/5). . . . 154

B.4 Suma estandarditzada i correlacions entre els diversos estadístics i les mesures de pèrdua de confidencialitat per intervals de confiança, basats en desviacions típiques, obtingudes en aplicar les diverses versions de microagregació multivariant amb k=10 (3/5). . . . 155

B.4 Suma estandarditzada i correlacions entre els diversos estadístics i les mesures de pèrdua de confidencialitat per intervals de confiança, basats en desviacions típiques, obtingudes en aplicar les diverses versions de microagregació multivariant amb k=10 (4/5). . . . 156

B.4 Suma estandarditzada i correlacions entre els diversos estadístics i les mesures de pèrdua de confidencialitat per intervals de confiança, basats en desviacions típiques, obtingudes en aplicar les diverses versions de microagregació multivariant amb k=10 (5/5). . . . 157

B.5 Suma estandarditzada i correlacions entre els diversos estadístics i les mesures generals obtingudes en aplicar les diverses versions de microagregació multivariant amb k=10 (1/4). 158

B.5 Suma estandarditzada i correlacions entre els diversos estadístics i les mesures generals obtingudes en aplicar les diverses versions de microagregació multivariant amb k=10 (2/4). 159

B.5 Suma estandarditzada i correlacions entre els diversos estadístics i les mesures generals obtingudes en aplicar les diverses versions de microagregació multivariant amb k=10 (3/4). 160

B.5 Suma estandarditzada i correlacions entre els diversos estadístics i les mesures generals obtingudes en aplicar les diverses versions de microagregació multivariant amb k=10 (4/4). 161

Llista de Figures

3.1 Grups de mida variable versus grups de mida fixa 24

4.1 Representació dels 9 vectors de dades de l'exemple. 43

6.1 Mesures obtingudes en aplicar la pertorbació additiva aleatòria (Addt). 71
6.2 Mesures obtingudes en aplicar la microagegació amb ordenació individual (MicOI). . . 72
6.3 Mesures obtingudes en aplicar la microagegació per projecció sobre la suma de les z-puntuacions (MicZ). 73
6.4 Mesures obtingudes en aplicar la microagegació per projecte sobre la primera component principal (MicPCP). 74
6.5 Mesures obtingudes en aplicar la microagegació multivariant amb grups de variables de mida 2 (Mic2mul). 75
6.6 Mesures obtingudes en aplicar la microagegació multivariant amb grups de variables de mida 3 (Mic3mul). 76
6.7 Mesures obtingudes en aplicar la microagegació multivariant amb grups de variables de mida 4 (Mic4mul). 77
6.8 Mesures obtingudes en aplicar la microagegació multivariant amb grups de variables de mida 5 (Mic5mul). 77
6.9 Mesures obtingudes en aplicar la microagegació multivariant amb grups de variables de mida 6 (Mic6mul). 78
6.10 Mesures obtingudes en aplicar la microagegació multivariant amb un sol grup de totes les variables (Micmul). 79
6.11 Mesures obtingudes en aplicar la pèrdua per compressió (JPEG). 80
6.12 Mesures obtingudes en aplicar l'intercanvi de dades (Rank). 81

Capítol 1

Introducció

En l'elaboració d'estadístiques oficials es poden diferenciar tres etapes: recollida de la informació, processament de les dades i posterior publicació. Abans de publicar qualsevol tipus d'informació que tingui contingut confidencial, l'oficina d'estadística ha d'assegurar-se que no sigui possible identificar, a partir de les dades publicades, cap dels registres individuals. Tot i això, no és possible eliminar completament el risc de revelació perquè els estadístics publicats han de reflectir d'alguna manera la realitat de la població d'individus d'on s'han tret les dades. Per tant, el terme controlar la revelació és més apropiat que evitar la revelació.

Les oficines d'estadística distribueixen dos tipus de dades a través de les seves bases de dades: les taules de macrodades i els conjunts de microdades (registres individuals). Tot i que existeix una llarga experiència en la publicació de taules de macrodades, recentment ha anat creixent la demanda de registres de microdades, puix que ofereixen molta més flexibilitat a l'hora de processar les dades.

L'objectiu de les tècniques de control de la revelació estadística és justament la *modificació* de les dades que tinguin contingut confidencial respecte a entitats individuals com poden ser persones, famílies, empreses, etc..., perquè, després de la seva publicació, els seus usuaris o el públic en general no puguin extreure'n informació confidencial. Per tal d'aconseguir aquest objectiu, es necessiten criteris per comprovar si unes determinades dades són suficientment segures per ser publicades; de fet, una part molt significativa del treball dintre el camp del control de la revelació ha estat la recerca d'aquests criteris. De manera que si bé tals criteris d'avaluació del risc de revelació orientaran el procés de *modificació* de les dades, un objectiu durant tot aquest procés ha de ser conservar el màxim d'informació possible, és a dir, modificar les dades tan poc com sigui possible. Un dels principals problemes a l'hora de publicar dades estadístiques és precisament aquest: com maximitzar el contingut d'informació lliurat al públic que farà ús de les dades mentre es resguarda la privacitat dels individus. L'objectiu principal serà arribar a un bon equilibri entre la quantitat d'informació continguda en les dades publicades i el nivell de protecció desitjat.

1.1 Objectius

Tot i que podem diferenciar tres etapes en el treball realitzat per les oficines d'estadística: recollida de la informació, processament de les dades i posterior publicació, aquesta memòria de tesi només tractarà temes relacionats amb la protecció de la confidencialitat de les dades en el procés de publicació de les mateixes.

Després de recollir la informació referent a les tècniques més rellevants de control de la revelació

de microdades contínues actualment existents, l'objectiu general de la tesi és l'anàlisi i la millora d'aquestes tècniques de control de la revelació mitjançant mètodes d'estadística matemàtica; millora referida a almenys un dels tres següents aspectes:

Nivell de protecció. Donar un bon grau de protecció a la informació confidencial de les dades que han de ser publicades.

Pèrdua d'informació. Minimitzar la pèrdua d'informació durant el procés de *modificació* de les dades.

Complexitat computacional. Reduir el temps de càlcul i/o computació inherent a l'aplicació de tècniques de control de la revelació.

1.2 Aportacions

L'anàlisi i millora referides als objectius generals d'aquesta tesi han estat aplicades concretament a una tècnica de control de la revelació per a microdades contínues anomenada microagregació que bàsicament ajunta registres individuals del conjunt de microdades per tal de disminuir el risc de revelació.

Podem diferenciar tres tipus d'aportacions de la tesi de la següent manera:

1. Aportacions als mètodes de microagregació univariant, aplicats fonamentalment al tractament de microdades contínues univariants.
2. Aportacions als mètodes de microagregació multivariant, aplicats bàsicament al tractament de microdades contínues multivariants (més d'una variable observada).
3. Mesures comparatives de mètodes pertorbatius.

Microagregació univariant

- S'ha desenvolupat un estudi analític mitjançant estadístics d'ordre sobre la seguretat del mètode de microagregació amb ordenació individual.
- S'ha comparat la qualitat del mètode de microagregació mitjançant ordenació individual amb altres mètodes de control de la revelació per a microdades contínues; qualitat que ha estat mesurada per l'equilibri aconseguit entre la pèrdua d'informació i el risc de revelació.

Microagregació multivariant

- S'ha creat un nou mètode de microagregació multivariant de la "Distància Màxima Modificat" (DMM), modificació d'un altre mètode existent anomenat de la "Distància Màxima" (DM) i s'han comparat les seves complexitats computacionals.
- Hem comparat la qualitat del nou mètode de microagregació de la "Distància Màxima Modificat" (DMM) amb altres mètodes de control de la revelació per a microdades contínues; qualitat que també ha estat mesurada per l'equilibri aconseguit entre la pèrdua d'informació i el risc de revelació.
- Hem desenvolupat un estudi analític per calcular el número de possibles particions d'un conjunt de p variables observades en $h-1$ conjunts de mida s i un únic conjunt de mida $s+r$, on $p = hs + r$.

- S'ha realitzat un estudi sobre el número de variables que han de tenir els conjunts d'una partició sobre la que s'executarà el mètode DMM perquè el conjunt modificat de dades resultant tingui una bona qualitat.
- Finalment, hem fet un estudi sobre la combinació de variables dintre els conjunts que formen una partició que, juntament amb l'anterior estudi sobre el número de variables, proporcionen a l'usuari de la microagregació multivariant una guia per saber quantes i quines variables haurien de formar la partició del conjunt de variables sobre la que s'executarà el mètode DMM perquè el conjunt modificat de dades resultant tingui una millor qualitat.

Mesures comparatives

- Distinció entre les diverses naturaleses que formen part de les mesures emprades per comparar mètodes pertorbatius.
- Ponderació de les diverses mesures tenint en compte les diverses naturaleses trobades en el punt anterior.
- Creació d'una nova mesura de pèrdua de confidencialitat basada en intervals de confiança construïts a partir de desviacions típiques.

1.3 Estructura general de la memòria

Considerant el contingut dels temes tractats en aquesta memòria de tesi, podem distingir cinc parts principals:

- Part 1. La primera part, que conté el capítol 2, introdueix els conceptes bàsics que estan implicats en els mètodes de control de la revelació estadística. També es descriuen breument els mètodes més rellevants de control de la revelació estadística per a microdades.
- Part 2. Aquesta segona part, que conté el capítol 3, està dedicada a l'estudi analític de la seguretat de la microagregació univariant amb ordenació individual mitjançant estadístics d'ordre; mètode molt freqüentment utilitzat donada la seva baixa complexitat computacional.
- Part 3. Aquesta part conté tres capítols (capítols 4, 5 i 6). Al capítol 4 es descriu el nou mètode de la "Distància Màxima Modificat" per microagregar dades multivariants sense projectar i es fa una anàlisi de la seva complexitat computacional. El capítol 5 presenta una caracterització de la qualitat d'un mètode de control de la revelació pertorbatiu a través d'unes determinades mesures, ja que, donada la gran diversitat existent d'aquests mètodes, se'ns presenta el repte de poder comparar la seva qualitat. El capítol 6 presenta tot un estudi comparatiu sobre la qualitat entre diversos mètodes de control de la revelació pertorbatius per a microdades contínues.
- Part 4. La quarta part conté dos capítols (capítols 7 i 8). El capítol 7 comença amb un estudi analític per calcular el número de possibles particions d'un conjunt de p variables observades en $h-1$ conjunts de mida s i un únic conjunt de mida $s+r$, on $p = hs + r$; després segueix amb la discussió sobre els cardinals dels conjunts que formen la partició sobre la que s'executa el mètode DMM perquè el conjunt modificat de dades resultant tingui una bona qualitat. El capítol 8 ofereix un posterior estudi sobre la combinació de les variables dintre els conjunts de la partició sobre la que s'executa el mètode DMM, que complementa els resultats obtinguts al capítol 7.
- Part 5. Aquesta part, que conté el darrer capítol 9, com a conclusió de la memòria de tesi ressalta els principals resultats de tot el treball realitzat i perfila possibles línies de futura recerca.

1. Introducció

Capítol 2

Conceptes bàsics

Aquest capítol introdueix els conceptes que s'usen quan es treballa per tal de protegir les dades estadístiques. També s'ofereix un sumari de les tècniques que existeixen per protegir dades confidencials. Aquestes tècniques agrupen o distorsionen les dades suficientment per reduir el risc d'identificar individus. El repte que es planteja en el problema de la confidencialitat és protegir les dades dels individus sense degradar-ne excessivament la validesa analítica.

No es pot recomanar un mètode, ja que *el millor mètode* depèn del tipus de dades, del tipus de taula que serà publicada, de les anàlisis estadístiques a realitzar i del nivell de protecció desitjat per a les dades.

La secció 2.1 distingeix i defineix els tipus d'estadístics (macrodades i microdades) que generalment podem trobar quan es publiquen les dades recollides dels individus. A la secció 2.2 es fa una descripció dels problemes de confidencialitat que pot provocar la difusió de macrodades o de microdades. Cadascun d'aquests tipus de dades té les seves peculiaritats quant als problemes de revelació que apareixen. A la secció 2.3 es donen els criteris d'avaluació que s'han de tenir en compte quan s'aplica algun mètode per tal de controlar la revelació estadística. A la secció 2.4 es fa una classificació dels diversos mètodes de control de la revelació existents per a microdades. La secció 2.5 introdueix la notació i la formalització que s'usaran en posteriors seccions. A la secció 2.6 es descriuen breument algunes tècniques de protecció de microdades basades en la reducció de les dades. Per acabar aquest capítol, a la secció 2.7, es descriuen els mètodes pertorbatius més rellevants per a la protecció de microdades.

2.1 Classificació dels estadístics

Les variables que tractem poden ser qualitatives o quantitatives. Les variables qualitatives es poden classificar en: qualitatives nominals (o categòriques), com per exemple la variable *sexe* i qualitatives ordinals, com per exemple una classificació segons les expressions "*baix, mitjà, alt*". Les variables quantitatives poden ser mesurades en una escala mètrica (o numèrica), com per exemple l'*edat* o els *ingressos*.

Per agrupar diferents variables en estadístics i ser tractats pels diversos mètodes de control de la revelació existeix una tipologia proposada a Dalenius (1988). Els estadístics es classifiquen segons:

- el seu format: en macrodades i microdades,
- la seva forma: en freqüències i magnituds,

Edat	Sexe	Estat civil	Nombre de fills	Ingressos mensuals (en ptes.)	...
38	femení	solter	0	173500	...
49	masculí	vidu	3	154200	...
23	femení	casat	1	73600	...
...	...	...	...	...	...

Taula 2.1: Exemple de microdades

- el mitjà emprat per publicar: en impremta i bases de dades o altres mitjans.

En aquesta memòria ens basarem en la primera classificació que s'ha fet dels estadístics que es volen publicar, és a dir segons el seu format:

- Les **microdades** es defineixen segons Willenborg i De Waal (1996) com un conjunt de registres sobre dades d'individus, els quals poden ser persones, empreses, companyies, etc... És a dir, les microdades consisteixen en la informació al nivell dels subjectes que responen. Tota la informació d'aquests subjectes ha de ser tractada com a confidencial i ha de ser protegida contra la revelació. Per a cada subjecte j tenim un vector individual de dades V_j (també s'anomena registre de dades), el qual pot tenir variables qualitatives i/o quantitatives. De les variables que formen part d'un conjunt de microdades en podem distingir tres tipus bàsics segons el seu grau de compromís respecte la privacitat dels individus:

  - **Identificadors directes**: Són variables el coneixement de les quals provoca la identificació de manera unívoca de l'individu al qual pertanyen aquests identificadors. Exemples d'aquest tipus de variables, quan es treballa amb persones, serien el nom o el número de DNI.
  - **Identificadors indirectes**: Són variables que poden servir per identificar l'individu al qual pertanyen, però no de manera unívoca. El que sí pot succeir és que hi hagi una combinació estranya o inusual d'identificadors indirectes que puguin provocar la identificació de l'individu en qüestió. Identificadors indirectes poden ser l'edat, el sexe o l'estat civil, quan es treballa amb persones.
  - **Variables sensibles**: Aquestes variables pertanyen al domini privat dels individus i s'ha d'evitar que es pugui relacionar una variable sensible amb l'individu al qual pertany aquesta variable. El criteri per decidir si una variable és sensible o no pot variar segons els països: el que en un lloc és una variable sensible en altres no ho és i a l'inrevés. Exemples de variables sensibles poden ser el passat criminal o les malalties que pateixen o han patit les persones.

  Vegeu la Taula 2.1 per a un exemple de microdades.

- Les **macrodades** són tabulacions de dades individuals. Cada cel·la es defineix ajuntant algunes variables, les quals poden ser qualitatives o quantitatives. Si les variables són quantitatives, s'utilitza un interval de mesura. Depenent del que representin les cel·les trobem dos tipus de taules:

  - Si per a cada cel·la es compta o s'estima el nombre d'elements que hi pertanyen, aleshores l'estadístic s'anomena **taula de contingència**.
  - Si s'agrega una variable quantitativa com *ingressos* o *producció* de tots els elements que pertanyen a una cel·la, aleshores anomenarem la taula resultant **taula de magnituds**.

L'ús d'ordinadors i paquets estadístics permet als usuaris fer avaluacions estadístiques i crear taules fetes a mida en lloc de rebre les anàlisis des d'un institut d'estadística. Aquesta tendència seguirà i, per tant, les oficines d'estadística experimentaran una demanda creixent, per part dels usuaris, de la difusió de fitxers de microdades en lloc de les anàlisis predefinides per les mateixes oficines. Degut a la seva mida, normalment les microdades es publiquen a través de bases de dades o altres fitxers.

Per a les dues classes de macrodades (taules de contingència i de magnituds) hi ha dos camins de publicació: la publicació estadística impresa i les bases de dades o altres fitxers informàtics.

2.2 El problema de la revelació

El problema de la revelació relaciona la possibilitat d'identificar individus a través de la publicació d'informació estadística. Per tenir una visió jurídica dels problemes que planteja el secret estadístic es pot consultar Bacaria-Martrus (1993).

Segons Dalenius (1977), la revelació estadística es produeix si la publicació d'algun estadístic permet a l'usuari extern de les dades obtenir un estimador millor d'alguna dada confidencial del que seria possible sense aquesta publicació. El problema de la revelació apareix si és possible arribar a una estimació massa precisa i, per tant, inacceptable de la informació confidencial d'un individu.

Tota informació confidencial ha de ser protegida contra la revelació emprant algun mètode de control de la revelació. Les tècniques no poden eliminar completament el risc de revelació, però poden donar protecció contra una revelació completa o exacta i contra un cert grau de revelació parcial o aproximada (Adam i Wortmann 1989):

- La **revelació completa o exacta** ocorre si un usuari extern d'un estadístic és capaç de determinar un atribut exacte A_i per a un individu i representat a la base de dades.
- La **revelació parcial o aproximada** es produeix si un usuari pot determinar un estimador $\hat{A}_i$ per a un individu i, la variància del qual satisfà que $\sigma^2(\hat{A}_i) < c_i^2$, on c_i és un paràmetre que ha de ser fixat. En altres paraules, la revelació parcial es produeix si es pot aconseguir una estimació massa precisa de les dades d'un individu.

2.2.1 Revelació amb macrodades

Si la informació està presentada en taules; és a dir, les dades no estan donades a un nivell d'individus, sinó agregades d'alguna manera, ens trobem en una situació diferent a la que tindrem amb microdades. Les cel·les d'una taula que no són publicables degut al risc de revelació estadística seran anomenades **cel·les confidencials**. Hi ha dos tipus de situacions (Schackis 1993)(Duncan i Lambert 1986) on la revelació es pot produir i, per tant, les cel·les seran confidencials:

- **Valors petits**: una cel·la és confidencial si menys de m entitats contribueixen al total d'aquesta cel·la. El valor de m s'anomena *límit* i usualment està donat pels Instituts d'Estadística d'acord amb el grau de protecció de la confidencialitat desitjat: m és almenys 3 ($m = 3$ és el valor més habitual, però de vegades s'agafa fins a $m = 5$). En el cas $m = 3$, una cel·la és confidencial si la dada que es mostra a la cel·la prové d'una o dues entitats i algú té la possibilitat de descobrir quines entitats són. Aquesta regla també s'anomena *regla del límit*.

  Aquesta regla s'aplica generalment si la taula es basa en el cens; la qual cosa significa que les dades provenen de tota la població. En el cas que la taula es basi en una mostra, és possible que l'Institut no apliqui la regla del límit ja que un usuari extern de les dades no sap quins individus formen part de la mostra.

- **Cas predominant**: una cel·la serà confidencial si n individus contribueixen més del k per cent en el total de la cel·la. Aquest cas també es coneix com a *regla*(n, k), *regla de dominància* o *regla de concentració*. Els nombres n i k els donen els Instituts i poden ser molt diferents. Per exemple, si agafem n=2 i k=85, una cel·la serà confidencial si dues unitats aporten més del 85% del total de la cel·la. La regla del cas predominant s'aplica quan treballem amb poblacions i no tant quan ho fem amb mostres.

Encara que totes les cel·les que contenen valors petits o casos predominants hagin estat protegides per mètodes de control de la revelació (és el que s'anomena **protecció primària**), la revelació es pot produir recalculant les cel·les confidencials com a diferència entre el total marginal i la suma de les cel·les no confidencials corresponents al total marginal, o mitjançant la comparació de dues o més taules entre elles. Aquest recàlcul de les cel·les confidencials s'anomena *derivació* i s'ha d'evitar la seva aparició a través d'una protecció suplementària (anomenada **protecció secundària**) adreçada a altres cel·les diferents de les cel·les confidencials.

2.2.2 Revelació amb microdades

Els dos grans grups de variables, identificadores i sensibles, són confidencials ja que contenen informació a nivell dels enquestats. Com que cada conjunt de dades està format pels dos grups de variables, s'ha de protegir contra la revelació abans de la seva difusió. Els mètodes de control de la revelació per a la publicació de microdades han d'assegurar que el risc que algú pugui associar correctament variables sensibles amb un individu, utilitzant les variables identificadores i el seu coneixement a priori, sigui prou baix.

La informació que algú pot tenir sobre les variables identificadores d'un enquestat es defineix com a coneixement a priori. Qualsevol pot tenir coneixement a priori sobre tots o una part dels identificadors directes i/o indirectes. Normalment, aquesta informació a priori està formada per dades que són fàcils d'obtenir pel públic o que ja són conegudes pel públic que té contacte amb la persona objectiu, per exemple els parents, amics, veïns, companys de feina, etc... Per tant, aquestes dades són no confidencials en el sentit que no han de ser protegides pels mètodes de control de la revelació, però aquest coneixement a priori pot permetre a un *intrús* (persona que intenta obtenir informació sobre les variables sensibles d'un individu concret) identificar els enquestats i pot conduir a la revelació de dades sensibles.

La identificació es produeix si és possible una correspondència correcta entre un registre de dades i un individu concret. Si el registre de dades publicat inclou identificadors directes, és possible la identificació d'un enquestat de manera instantània. Però, generalment, els identificadors directes s'esborren del vector de dades abans de la seva difusió. Per tant, l'única possibilitat per a l'investigador d'aconseguir una relació correcta és emprar el seu coneixement a priori sobre els valors dels identificadors indirectes de l'enquestat objectiu. Si tots els identificadors directes s'esborren i l'intrús no té informació a priori sobre l'enquestat objectiu, la identificació és impossible.

La identificació és un requisit previ per a la revelació. La revelació, quan treballem amb microdades, es produeix si hi ha una identificació per part de l'investigador que el portaria a aconseguir informació sobre les variables sensibles, les quals es guarden juntament amb les variables identificadores en el fitxer de microdades. La revelació no és possible sense la identificació. El risc de la identificació i, per tant, de la revelació, depèn de la quantitat i de la naturalesa de la informació a priori de la qual es disposi. Ens podem trobar diversos escenaris que afavoreixin la revelació estadística (Keller i Bethlehem 1992). A Skinner (1992) es fa la distinció entre la revelació per identificació i la revelació per predicció.

Per altra banda, també hi ha una sèrie de barreres que dificulten la identificació dels individus.

A Blien, Wirth i Müller (1992) es mostra, a través d'un exemple amb dades reals, que hi ha més problemes dels previstos per tal d'aconseguir la identificació d'algun enquestat.

Un control de la revelació eficient en el cas de les microdades significa que els microestadístics publicats ho han de ser de manera que la correspondència correcta entre un vector de dades del microestadístic i un individu no sigui factible usant les variables identificadores i el coneixement a priori.

La identificació serà possible per unicitat. Un enquestat s'anomena únic a la població si cap dels altres subjectes de la població té la mateixa combinació de valors dels identificadors indirectes. Així doncs, si s'estudia una mostra de la població, aleshores un individu únic de la població pot ser que estigui o no a la mostra. Si l'individu únic forma part de la mostra, aleshores també serà únic a la mostra. Aquest fet permetrà la revelació ja que si algú coneix a priori que aquesta combinació de valors dels identificadors indirectes és única a la població, identificarà a l'enquestat al fitxer de microdades. Pel contrari, la unicitat a la mostra sense un coneixement a priori que els identificadors indirectes són únics a la població no conduirà necessàriament a la revelació, ja que aquesta combinació d'identificadors indirectes pot ser molt comú a la població. A Bethlehem, Keller i Pannekoek (1990), a Zayatz (1991), a Skinner, Marsh, Openshaw i Wymer (1990), a Mokken, Kooiman, Pannekoek i Willenborg (1992) i a Samuels (1998) es fa un estudi més detallat sobre el problema de la unicitat i la manera d'avaluar el risc de revelació per a un conjunt de microdades.

A continuació s'exposen els passos que se segueixen per arribar a la revelació amb microdades:

- L'intrús té coneixements previs sobre les variables identificadores d'una persona objectiu.
- L'intrús reconeix els identificadors indirectes al microestadístic publicat.
- Els identificadors indirectes són únics a la mostra i a la població.
- És possible la identificació de l'individu objectiu.
- És possible la revelació de variables sensibles de l'individu objectiu.

2.3 Criteris d'avaluació

Algunes característiques han de ser avaluades i comparades per tal de veure els avantatges i desavantatges de cadascun dels mètodes de control de la revelació. No hi ha tècniques que compleixin tots els criteris, ja que hi ha criteris que són contraposats a altres. Cada mètode té els seus punts forts i febles, i algunes vegades es tria una combinació de mètodes per tal de maximitzar els avantatges.

Els criteris que s'utilitzen per avaluar els diversos mètodes els podem trobar a Adam i Wortmann (1989), i són els següents:

1. **Seguretat.** Un mètode de control de la revelació eficient protegeix contra una estimació exacta i contra una estimació massa precisa dels valors de les variables d'un individu; en altres paraules, una tècnica efectiva evita la revelació exacta i/o parcial. Per tant, el nivell de seguretat ha de ser alt. En el cas dels mètodes de control de la revelació per a la publicació de microdades, la protecció és assegurada si la identificació d'un individu no és possible.
2. **Robustesa.** Un mètode de control de la revelació s'anomena robust si un coneixement addicional de l'usuari extern de les dades, a banda dels valors publicats, no permet a aquest usuari descobrir informació confidencial. Un mètode ideal ha de mantenir un nivell de protecció que faci impossible a un usuari trobar valors secrets encara que tingui informació suplementària. De totes maneres, la robustesa d'un mètode és difícil d'avaluar.

3. **Flexibilitat.** La flexibilitat d'un mètode de control de la revelació ha de ser tan alta com sigui possible. Una tècnica s'anomenarà flexible si satisfà tres condicions desitjables:

   (a) Pot ser emprada per a taules de contingència i per a taules de magnituds.

   (b) Pot treballar amb més d'una variable al mateix temps.

   (c) Pot protegir variables qualitatives i quantitatives.

4. **Riquesa d'informació.** Un alt nivell de riquesa d'informació donat per les dades publicades és un altre criteri important per comparar i avaluar diverses tècniques. Diferents aspectes han de ser considerats en aquest context:

   (a) La quantitat d'informació dels estadístics publicats ha de ser el més alta possible donat un nivell de protecció de la revelació. Un mètode ideal només assegura la informació confidencial però dóna a l'usuari la informació no confidencial. Així, la quantitat de dades no confidencials eliminades és un criteri per jutjar el mètode.

   Per ser més precisos, seguint la Teoria de la Informació (Shannon 1948), s'han de distingir les dades segons el criteri d'informació que té l'usuari extern. Les dades publicades poden ser informatives per a l'usuari, per exemple a causa que són noves per a ell, o poden ser no informatives, per exemple a causa que ja són de domini públic abans de la publicació. Per tant, no és la quantitat eliminada d'informació no confidencial sinó la quantitat de dades informatives no confidencials eliminades la que ha de romandre baixa. Aquest és un fet que l'estadístic ha de considerar quan tria un mètode.

   (b) Quan s'usen mètodes de pertorbació, el biaix, la precisió i la consistència són característiques importants que avaluen la qualitat de les dades protegides.

   El biaix és la diferència entre un valor A_i de les dades abans de la pertorbació i el valor esperat de l'estimació $E(\hat{A}_i)$ derivada de les dades pertorbades. Idealment la diferència ha de ser 0.

   La precisió fa referència a la variància de l'estimador obtingut per l'usuari de les estadístiques. La informació donada ha de ser tan precisa com sigui possible; en altres paraules, la variància de l'estimador $\sigma^2(\hat{A}_i)$ s'ha de mantenir el més baix possible. Malgrat tot, per controlar la revelació parcial, la condició $\sigma^2(\hat{A}_i) > c_i^2$, on c_i és un paràmetre que s'ha de donar, s'ha de complir a la majoria dels casos. L'estadístic ha de trobar un conjunt de paràmetres d'un mètode de control de la revelació de manera que hi hagi un balanç entre les dues restriccions.

   La consistència d'un mètode significa que no es produeixen contradiccions ni paradoxes. Poden ocórrer contradiccions, per exemple, si es donen respostes diferents quan es fa la mateixa consulta o si l'estadístic mitjana no concorda amb la mitjana calculada per l'usuari a partir dels estadístics suma i freqüència. Una paradoxa és, per exemple, una resposta negativa a una consulta de freqüència. És desitjable que la consistència d'un mètode sigui el més alta possible.

   (c) L'accés dels usuaris a les dades, és a dir, l'accés a les dades després que han estat tractades per un mètode de control de la revelació i les parts confidencials n'han estat protegides, és un altre aspecte per avaluar una tècnica. Un usuari pot tenir accés al rang complet de les dades i processar la informació en què està interessat (aquesta és la situació ideal), o l'accés a les dades és restringit i només pot tenir accés a cert nombre de taules o potser a taules amb certa estructura. Restringir l'accés és, de vegades, necessari per evitar la derivació entre taules.

   (d) Una alta utilitat de les dades per altres processos, per exemple per crear taules o per fer anàlisis pròpies, és una altra característica desitjable d'un mètode eficient de control de

la revelació. Sovint el desavantatge d'un accés restringit a les dades es combina amb una baixa utilitat per processos posteriors i viceversa.

5. **Cost.** El cost s'ha de mantenir el més baix possible. Respecte del cost, es consideren tres components:
   (a) El cost per a la implementació caracteritza l'esforç necessari per implementar el mètode i fixar els paràmetres necessaris.
   (b) La quantitat d'entrenament per tal que un usuari entengui les idees bàsiques de la tècnica per facilitar l'ús de les dades publicades.
   (c) El processament diari dels estadístics incloent el temps per fer-lo. Aquest component és important per a l'avaluació dels diferents mètodes, però molt difícil d'estimar sense tenir l'oportunitat de provar i comparar tots els mètodes.

2.4 Classificació dels mètodes de control de la revelació per a microdades

Actualment existeixen molt diversos mètodes de control de la revelació per a microdades. Tots aquests mètodes es poden classificar des de dos punts de vista diferents.

1. Considerant els seus principis operacionals, podem distingir dos tipus de mètodes (Schackis 1993):
   (a) **Mètodes de reducció de dades** Aquests mètodes disminueixen la informació de les dades suprimint generalment parts dels vectors de dades. L'avantatge d'aquests mètodes és que no canvien l'estructura de les dades originals. Però un inconvenient rau en què un d'aquests mètodes per si sol no aporta la protecció suficient, puix que, després del tractament de les dades individuals per algun d'aquests mètodes, la informació continguda en les microdades és encara bastant alta per permetre la identificació dels individus i la revelació de dades confidencials. Per la qual cosa, els mètodes per reducció de dades han de combinar-se normalment amb altres mètodes de control de la revelació.
   (b) **Mètodes de modificació de dades** (**Pertorbatius**) La seva característica bàsica és la modificació de les dades originals abans de la seva publicació. La qual cosa significa que aquests mètodes pertorben les dades originals per reduir el risc de revelació, però intentant mantenir la riquesa d'informació el més alta possible. Així doncs, els estadístics calculats sobre les dades pertorbades no han de discrepar significativament respecte dels mateixos estadístics calculats a partir de les dades sense modificar.

2. Considerant la naturalesa de les dades sobre les quals han de ser aplicats, podem distingir també dos tipus de mètodes:
   (a) **Mètodes orientats a variables categòriques** Aquests mètodes han estat pensats per adaptar-se molt bé al control de la revelació de variables categòriques. Una variable es considera categòrica quan agafa valors sobre un conjunt finit i no té sentit aplicar-li les operacions aritmètiques estàndards. Per exemple, els dies de la setmana, el color dels ulls, etc....
   (b) **Mètodes orientats a variables contínues** Aquests altres mètodes han estat desenvolupats per ajustar-se al control de la revelació de variables contínues. Recordem que una variable és contínua quan pot agafar com a valor qualsevol nombre d'un interval de la recta real. Per exemple, el pes, l'estatura, els ingressos, etc....

Per descriure diversos mètodes de control de la revelació per a microdades, seguirem el següent esquema, tot i comentar la seva millor o pitjor adaptació a dades categòriques i/o contínues:

- Mètodes de reducció de dades
  - Anonimització
  - Mostreig
  - Restricció de la mida de la població
  - Reducció de detalls
  - Codificació global
  - Supressió
- Mètodes de modificació de dades (pertorbatius)
  - Pertorbació additiva aleatòria
  - Pertorbació segons una distribució de probabilitat
  - Remostreig
  - Intercanvi de dades
  - PRAM
  - Pèrdua per compressió
  - Microagregació

Certament, la tria dels mètodes per a un cas concret de publicació de microdades depèn del tipus i de la quantitat de les microdades, del nivell desitjat de protecció, de la precisió de la informació que ha de ser publicada i de les anàlisis que es volen fer.

Cas que, després del tractament de les dades originals per mètodes de reducció de dades, encara no s'hagi arribat al nivell de protecció desitjat, les microdades són aleshores tractades addicionalment per algun mètode pertorbatiu de modificació de les dades. L'avantatge dels mètodes pertorbatius és que aconsegueixen un nivell alt de control del risc de revelació, però tenen el perill que, de vegades, les dades modificades puguin ser massa diferents de les dades originals i no permetin fer anàlisis estadístiques prou vàlides. En aquest sentit, la tasca dels estadístics consisteix a trobar una bona combinació dels mètodes de control de la revelació per reducció de les dades i per modificació de les dades.

2.5 Notació i formalització

Suposarem que la informació d'un fitxer de microdades està representada per una taula bidimensional, on una dimensió correspon al conjunt d'objectes (elements, individus, persones, etc...), i l'altra dimensió és el conjunt d'atributs, és a dir, les variables.

El fitxer de microdades conté un valor per a cada parella objecte-variable, de manera que podem modelitzar-lo com una funció

$$\mathbf{V} : \mathbf{O} \to D(V_1) \times D(V_2) \times \cdots \times D(V_m)$$

on $\mathbf{O}$ és el conjunt d'objectes; $V_1, V_2, \ldots, V_m$ són les variables i $D(V_i)$ és el domini de la variable V_i.

Així doncs, la funció m-dimensional V pot ser representada de la següent manera:

$$\mathbf{V}(O) = (V_1(O), V_2(O), \ldots, V_m(O))$$

on $V_i(\cdot) : \mathbf{O} \rightarrow D(V_i)$ és una funció unidimensional que assigna a cada objecte el valor de la variable V_i.

2.5.1 Variables contínues i variables categòriques

La variable V_i, segons la naturalesa del seu domini $D(V_i)$, es pot classificar com a contínua o com a categòrica:

Contínua: El domini de la variable V_i és un interval de la recta real, és a dir, $D(V_i) = [a, b]$.

Categòrica: El domini de la variable V_i està definit a través d'un conjunt de categories. Així doncs,

$$D(V_i) = \{l_0^i, \ldots, l_{n_i}^i\}$$

A la literatura, aquestes variables també s'anomenen variables lingüístiques, i les seves categories es referencien en termes lingüístics. Els dominis de les variables categòriques, segons la seva estructura d'ordre, es poden classificar també de dues maneres:

Categòric ordinal: En un domini categòric ordinal $D(V_i)$ existeix una relació d'ordre total $\leq_{V_i}$. Per simplificar la notació, suposarem que per a $l_r^i, l_s^i \in D(V_i)$, $l_r^i \leq_{V_i} l_s^i$ quan $r \leq s$. Per exemple, la variable V_i "Nivell de formació" té un domini categòric ordinal.

Categòric nominal: En aquest cas, no existeix cap relació d'ordre definida sobre el domini $D(V_i)$. Per exemple, quan V_i és la variable "Color dels ulls".

Nota. Per a variables digitalment representades, els dominis continus realment no existeixen a la pràctica perquè amb un nombre finit de decimals només podem representar una forma "discretitzada" d'una variable contínua. Tot i això, nosaltres considerarem contínues variables tals com l'estatura, els ingressos, el pes; fins i tot la variable edat.

2.6 Mètodes de reducció de dades

En aquesta secció es fa una breu descripció dels mètodes de control de la revelació per a microdades basats en la reducció de les dades que hem esmentat a la secció 2.4. S'ha dedicat una subsecció per a cada mètode.

2.6.1 Anonimització

Potser el mètode de control de la revelació basat en la reducció de les dades que primerament apliquem i, a la vegada, el més senzill és l'anonimització dels registres de dades (Schackis 1993). La idea és esborrar de cada vector de dades els identificadors directes abans de publicar el microestadístic.

Després de l'anonimització, un enquestat només es pot identificar a partir dels seus identificadors indirectes. Per tant, quan un intrús que intenta obtenir informació sobre les variables sensibles d'algun individu concret comenci a treballar, necessita un coneixement previ sobre tots o almenys una part dels identificadors indirectes d'aquest individu objectiu. I encara que l'intrús tingui un coneixement

previ sobre tots els identificadors indirectes de l'individu objectiu, la identificació d'aquest individu, després de l'anonimització, només serà possible si la combinació de valors dels identificadors indirectes és única en el fitxer microestadístic, ja que només en aquest cas és possible una relació segura entre un vector de dades i un individu.

2.6.2 Mètodes de mostreig

La característica bàsica dels mètodes de mostreig (Schackis 1993) és la publicació solament d'una mostra del conjunt original de dades. D'aquesta manera, una combinació única de valors dels identificadors indirectes en la mostra publicada, no necessariament identifica un individu que forma part de la població perquè la unicitat en la mostra no implica la unicitat en la població.

Així doncs, aquests mètodes, en lloc de publicar el fitxer original de dades

$$\mathbf{V} : \mathbf{O} \rightarrow D(V_1) \times D(V_2) \times \cdots \times D(V_m)$$

el que publiquen és

$$\mathbf{V}' : \mathbf{S} \rightarrow D(V_1) \times D(V_2) \times \cdots \times D(V_m)$$

on $S \subset O$ representa una mostra del conjunt original d'objectes i V' és la restricció de la funció V a S.

Perquè les dades tractades amb aquests mètodes de mostreig tinguin una bona riquesa d'informació, els estadístics calculats sobre la mostra publicada haurien de ser no esbiaixats respecte dels calculats sobre tota la població.

Generalment, els mètodes de mostreig no són adequats per protegir microdades que contenen variables contínues perquè aquests mètodes deixen sense modificar cada variable contínua $V_i(\cdot)$ en tots els objectes de la mostra S. Així doncs, si una variable $V_i(\cdot)$ està en algun fitxer extern publicat de dades, llavors serà molt probable trobar valors únics d'aquesta variable a la mostra, puix que seria molt estrany, en una variable contínua, que $V_i(o_1) = V_i(o_2)$ si $o_1 \neq o_2$.

L'únic escenari de revelació per al que podria tenir sentit utilitzar mètodes de mostreig seria el descrit a Willenborg i De Waal (1996), on se suposa que un intrús no coneix exactament els valors d'una variable contínua identificadora, sinó que solament coneix una aproximació. Però, fins i tot aquests models són molt perillosos quan els fitxers externs publicats tenen una bona qualitat.

2.6.3 Restricció de la mida de la població

La tècnica del mostreig moltes vegades es combina amb una restricció de la mida N de la població (Schackis 1993) a partir de la qual es recullen els vectors de dades per crear el microestadístic. La restricció pot ser, per exemple, $N > 100000$; això significa que la població en la qual es basa el microestadístic ha de contenir almenys 100000 individus i no es publicaran microestadístics de poblacions més petites. La raó per fer una restricció de N és que la fracció de mostreig es fa més petita a mesura que N és més gran. Per tant, el risc d'identificació disminueix perquè la possibilitat de bessons estadístics a la població augmenta. Així doncs, la probabilitat que certa combinació de valors dels identificadors indirectes sigui única a la població és més baixa a mesura que N és més gran.

2.6.4 Reducció de detalls

La reducció de detalls (Schackis 1993) també es coneix amb el nom d'aclariment en el sentit de fer-se menys espès. La idea bàsica d'aquest mètode és reduir la quantitat d'informació continguda en el

microestadístic publicat i per tant la seva informació. S'han de distingir dues versions de la reducció de detalls.

La primera versió és el *down-scaling*, que vol dir reduir la granularitat que s'utilitza per mesurar el valor d'una variable. El *down-scaling* redueix la informació de les dades. Les possibilitats per fer-ho són canviar una escala numèrica per una ordinal o nominal; o canviar una escala ordinal per una nominal.

La segona versió de la reducció de detalls és la reducció del nombre de categories donades per a cada identificador indirecte i/o per a cada variable sensible. Així, la informació del microestadístic difós disminueix.

2.6.5 Codificació global

Mitjançant la Codificació global (Willenborg i De Waal 2001), diverses categories d'una variable categòrica V_i es combinen per formar noves categories (menys específiques), de manera que resulta una nova variable V_i' amb $|D(V_i')| < |D(V_i)|$ (on $|D(\cdot)|$ indica el cardinal del conjunt $D(\cdot)$).

Aplicar la codificació global a una variable contínua V_i significa substituir V_i per una altra variable V_i' que és una versió discretitzada de V_i; és a dir, un domini $D(V_i)$ potencialment infinit, generalment, es transforma en un domini $D(V_i')$ finit. Aquesta tècnica ha estat utilitzada al programari μ-Argus SDC (Hundepool i Willenborg 1999).

La codificació superior i la codificació inferior són casos especials de codificació global, que es poden utilitzar en variables els valors de les quals admeten una ordenació, és a dir, variables contínues o categòriques ordinals. La idea bàsica d'aquestes dues tècniques és agrupar els "valors extrems" (bé siguin els més grans o bé els més petits, respectivament) per formar una nova categoria, puix que precisament en els valors més extrems d'una variable és on més probablement apareixen casos d'unicitat pel que fa als valors dels identificadors indirectes.

En principi, aquestes tècniques de codificació són més apropiades per controlar la revelació de microdades categòriques, perquè ajuden a disfressar vectors de dades que tenen combinacions estranyes en variables categòriques. Per exemple, cas que hi hagi un vector de dades amb "Estat civil = vídua" i "Edat = 17", es podrien utilitzar les tècniques de codificació per crear una categoria més àmplia "Vidu/a o divorciat/da", de manera que disminuís la probabilitat de trobar vectors de dades amb combinacions de valors estranyes.

La codificació global també es pot utilitzar en variables contínues, però, la inherent discretització que això suposa, podria conduir a una alta pèrdua d'informació. A més a més, les operacions aritmètiques que es podien aplicar directament sobre la variable original V_i, potser no seria fàcil aplicar-les sobre la versió discretitzada V_i'.

2.6.6 Supressió

Mitjançant aquest mètode se suprimeixen alguns valors individuals de variables amb l'objectiu d'augmentar el nombre de vectors de dades que coincideixen en valors clau d'aquestes variables. Així es fan desaparèixer valors únics de variables.

El mètode de supressió (Willenborg i De Waal 2001) s'empra generalment quan un "valor extrem" o una "combinació extrema de valors" es troben en algun vector de dades. Aquests valors extrems se suprimeixen abans de la publicació de les microdades, ja que aquests valors o aquestes combinacions de valors, donada la seva molt probable unicitat, poden simplificar massa la identificació d'algun individu.

Podem diferenciar dues versions del mètode de supressió. La primera consisteix a suprimir tots els valors extrems o combinacions extremes de valors que hi ha en un microestadístic i canviar-los per *missings*, abans de la seva difusió. L'usuari del microestadístic coneixerà que els valors suprimits són d'alguna manera valors extrems, però no podrà fer deduccions perquè no coneixerà el grau ni la direcció del valor extrem, és a dir, no sabrà si el valor real és alt o baix ni com d'alt o com de baix és.

La segona versió de la supressió consisteix a esborrar el vector complet. Aquesta possibilitat s'empra si un fitxer de dades conté un valor o combinació de valors molt estranya per a les variables, especialment si dades conegudes dels individus estan contingudes a l'estudi.

De vegades, un sol mètode de reducció de dades no aconsegueix protecció suficient contra la revelació, però sí la combinació de dos o més mètodes. Amb aquesta finalitat es poden combinar el mètode de Supressió i el mètode de Codificació global. A DeWaal i Willenborg (1995) apareixen maneres de combinar aquests dos mètodes, que estan implementades en el programari μ-Argus SDC (Hundepool i Willenborg 1999).

Cas que una variable contínua V_i formi part d'un conjunt de microdades, llavors cada combinació dels seus valors és probablement única. Com que no tindria sentit suprimir sistemàticament els valors de la variable V_i, concluïm que el mètode de Supressió està més bé orientat a variables categòriques.

2.7 Mètodes Pertorbatius

Com ha estat comentat a la secció 2.4, els mètodes de control de la revelació pertorbatius permeten de publicar el conjunt de microdades complet, però amb els seus valors modificats. No tots els mètodes pertorbatius han estat dissenyats per a variables contínues; per això farem referència a aquesta distinció en cada mètode particular.

La majoria dels mètodes pertorbatius que descriurem breument són casos especials d'emmascarament de matrius. La qual cosa significa que si el conjunt original de microdades és V, llavors el conjunt modificat de microdades V' es calcula com

$$V' = AVB + C$$

on la matriu A és una màscara transformadora d'objectes, la matriu B és una màscara transformadora de variables i la matriu C és una matriu pertorbadora additiva.

2.7.1 Pertorbació additiva aleatòria

El mètode de la Pertorbació additiva aleatòria (Kim 1986, Sullivan i Fuller 1989, Sullivan i Fuller 1990, Little 1993) consisteix a afegir a les dades originals una pertorbació aleatòria i independent, de manera que es conservi l'estructura de correlació de les dades originals.

Sigui $v_{ij} = V_i(o_j)$ el valor original de la variable V_i per a l'individu o_j. Sigui $e_{ij} = E_i(o_j)$ la pertorbació aleatòria afegida a v_{ij}. Així doncs, v_{ij} es substitueix per $v'_{ij} = v_{ij} + e_{ij}$.

Designarem per $V = \{v_{ij}\}$ la matriu que conté els v_{ij} com a elements, i per $E = \{e_{ij}\}$ i $V' = \{v'_{ij}\}$ les matrius que contenen com a elements e_{ij} i v'_{ij} respectivament. Suposarem el valor esperat $E(E) = 0$ i la variància $Var(E) = cVar(V)$, per a una constant fixada c. D'aquesta manera, la variància del conjunt modificat de dades és $Var(V') = (1 + c)Var(V)$, i podem recuperar la variància de les dades originals a partir de la variància de les dades pertorbades: $Var(V) = Var(V')/(1 + c)$.

Com que la pertorbació aleatòria ha estat generada independentment del conjunt original de variables, la covariància entre una variable modificada i una variable original és la mateixa que la

covariància entre les dues variables sense modificar: $Cov(V_i', V_j) = Cov(V_i, V_j)$. Per tant, l'estructura de correlació de V' és la mateixa que l'estructura de correlació de V.

La pertorbació E pot ser qualsevol variable aleatòria la distribució de la qual compleixi les condicions anteriorment referides; tot i que, generalment, la pertorbació aleatòria afegida serà gaussiana. Si la variable V_i a la qual s'afegeix la pertorbació és contínua amb una funció de densitat estrictament decreixent, com, per exemple, la densitat exponencial, i la pertorbació afegida és simètrica al voltant de 0, es mostra a Matloff (1986) que

$$E(V_i|V_i' = w) < w$$

on V_i' és la versió pertorbada de V_i.

Així, la constant c és l'únic paràmetre que s'ha d'ajustar, tot considerant que com més gran sigui l'amplada de la pertorbació aleatòria, o, en altres paraules, com més alta sigui la variància $Var(E)$, més gran serà també la pèrdua d'informació en canviar V per V'. Ens trobem novament davant d'una relació inversa entre protecció de les dades originals per una banda, i utilitat de les dades pertorbades per l'altra; un dels objectius de qualsevol mètode de control de la revelació és trobar un bon equilibri entre totes dues.

Aquest mètode és apropiat per protegir variables contínues per les següents raons:

- No suposa cap restricció per al domini de la variable V_i (que podria ser infinit).
- La pertorbació additiva aleatòria generalment és contínua i amb mitjana zero; per la qual cosa s'adapta molt bé a microdades originals contínues.
- No és pràcticament possible aparellar exactament dades entre un fitxer extern publicat i el conjunt original. Segons la quantitat de pertorbació afegida, solament seria possible aparellar dades de manera aproximada a través d'intervals.

2.7.2 Pertorbació segons una distribució de probabilitat

El mètode de Pertorbació segons una distribució de probabilitat (K. Liew, Choi i J. Liew 1985) es desenvolupa a través de tres etapes:

A. Identificació de la funció de densitat subjacent a cada una de les variables confidencials del conjunt original de dades

$$\mathbf{V} : \mathbf{O} \rightarrow D(V_1) \times D(V_2) \times \cdots \times D(V_m)$$

i estimació dels seus paràmetres.

B. Generació d'una sèrie de valors modificats V_i' per a cada variable confidencial V_i a partir de la funció de densitat estimada a la primera etapa.

C. Aparellament de valors i substitució de la sèrie original de valors confidencials per la sèrie de valors modificats.

Identificació i estimació

La primera etapa consisteix a trobar, entre un conjunt predeterminat de funcions de densitat, quina és la que millor s'ajusta a cada variable original. Un cas, per exemple, de funcions de densitat predeterminades podria ser: Poisson, exponencial, normal, gamma, Weibull, log-normal, χ^2 de Pearson. Es pot utilitzar l'estadístic de Kolmogorov-Smirnov per avaluar la bondat d'ajustament de les possibles funcions de densitat (donat un determinat nivell de significació, se selecciona la funció de densitat que produeix el valor més petit de l'estadístic Kolmogorov-Smirnov).

Generació del conjunt modificat de dades

Una vegada ha estat seleccionada la funció de densitat més ajustada al conjunt original de dades, a partir d'ella i dels seus paràmetres estimats es genera el conjunt modificat de dades.

Aparellament i substitució de valors

Cas que el conjunt modificat de valors s'utilitzés només per fer anàlisis estadístiques sobre una sola variable independentment de les altres, no seria necessari un criteri d'aparellament ordenat entre dades originals i dades modificades (és a dir, ordenar la sèrie original i la sèrie pertorbada de dades segons un mateix criteri i substituir cada element de la sèrie original pel corresponent valor pertorbat d'acord amb l'ordenació feta).

Però, en molts casos, els valors modificats d'una determinada variable s'utilitzen conjuntament amb valors modificats d'altres variables per realitzar anàlisis estadístiques multivariants. En aquests casos, s'han de trobar criteris adequats d'aparellament i substitució de valors entre dades originals i modificades perquè les referides anàlisis multivariants sobre el conjunt modificat de dades tinguin sentit.

La Pertorbació segons una distribució de probabilitat és un mètode molt apropiat per protegir variables contínues ja que únicament requereix que es pugui ajustar una funció de densitat al conjunt original de dades.

2.7.3 Remostreig

El mètode de Remostreig (Heer 1993, Domingo-Ferrer i Mateo-Sanz 1999) ha estat inicialment proposat com un mètode per protegir taules de contingència, però també es pot usar per protegir microdades.

Tot seguit expliquem molt breument la proposta original d'aquest mètode (Heer 1993) per a taules de contingència. Sigui V_i una variable categòrica. Siguin $v_1, \ldots, v_n$ els valors que agafa la variable V_i en un fitxer de microdades. Construïm una mostra $v'_1, \ldots, v'_n$ mitjançant n extraccions amb reemplaçament de les microdades originals. Es pot comprovar que les freqüències esperades de la mostra es corresponen amb les freqüències del conjunt original de microdades, de manera que la taula de contingència construïda a partir de $v'_1, \ldots, v'_n$ serà similar a la taula que obtindríem a partir de les dades originals $v_1, \ldots, v_n$.

A continuació, detallem una proposta d'utilització del Remostreig per a la protecció de microdades:

1. Agafem m mostres independents $S_1, \ldots, S_m$ del conjunt original d'objectes O. Totes les mostres han de tenir la mateixa mida n, essent n el cardinal del conjunt O.
2. Classifiquem cada una de les m mostres mitjançant un mateix criteri d'ordenació.
3. Construïm el conjunt modificat de microdades que s'ha de publicar com $\bar{v}_1, \ldots, \bar{v}_n$, on $\bar{v}_i$ és la mitjana dels i-èsims valors ordenats de les mostres $S_1, \ldots, S_m$.

Aquest procediment es pot executar per a cada variable independentment de les altres, o bé es pot executar basant-se en el vector de dades complet:

- Cas que aquest procediment s'executi de manera independent per a cada variable i posteriorment s'hagin de realitzar anàlisis multivariants entre conjunts de variables, llavors es fa necessari trobar criteris adequats d'aparellament i substitució de valors entre dades originals i modificades perquè les anàlisis multivariants sobre el conjunt modificat de dades tinguin sentit.

- Cas que aquest procediment s'executi basant-se en el vector de dades complet, llavors, tot i que no caldrà un aparellament ordenat entre dades originals i dades modificades perquè les correlacions entre variables es conservaran, sí serà necessari un criteri d'ordenació multivariant per classificar els vectors de dades de cadascuna de les m mostres independents amb la finalitat de calcular la mitjana dels corresponents vectors de dades.

L'avantatge que té el mètode de Remostreig respecte dels mètodes de mostreig anteriorment descrits és que en el Remostreig no es publiquen els valors exactes que prenen les variables per als individus de la mostra. D'aquesta manera, el mètode de Remostreig es pot utilitzar per protegir variables contínues.

2.7.4 Intercanvi de dades

La idea bàsica dels mètodes d'intercanvi de dades és transformar el conjunt original de dades en un altre conjunt modificat que conservi aproximadament l'estructura estadística del conjunt original (Dalenius i Reiss 1982)(Reiss 1984). Concretament, a Reiss (1984) aquest mètode s'utilitza per a variables multicategorials.

El conjunt original de dades es substitueix per un altre conjunt de dades generat de forma aleatòria que aproximadament té els mateixos t-estadístics d'ordre que el conjunt original. Aquests mètodes tenen prou complexitat computacional; i, a més a més, requereixen finalment tècniques d'aparellament i substitució de valors entre el conjunt original i el conjunt de dades generat aleatòriament (Adam i Wortmann 1989). Com trobar la millor tècnica d'aparellament i substitució de valors és una qüestió encara oberta, tot i que l'enllaç de registres (Winkler 1998) és una clara alternativa.

Encara que l'intercanvi de dades hagi estat originàriament descrit per a variables categòriques, existeixen versions que es poden utilitzar en qualsevol variable numèrica.

2.7.5 PRAM

El "Post-RAndomization Method" (PRAM) (Kooiman, Willenborg i Gouweleeuw 1997)(De Wolf, Gouweleeuw, Kooiman i Willenborg 1999) és un mètode pertorbatiu probabilístic per al control de la revelació de variables categòriques en fitxers de microdades. Mitjançant aquest mètode, els valors d'algunes variables categòriques de determinats vectors de dades del conjunt original es canvien, en el fitxer modificat de dades, a diferents valors, a través d'un mecanisme probabilístic que inclou una matriu de Markov.

El mètode PRAM és molt general perquè, mitjançant l'aproximació de Markov, combina la pertorbació de dades amb la supressió i la recodificació de dades. La pèrdua d'informació i el risc de revelació d'aquest mètode depenen en gran manera de l'elecció de la matriu de Markov i són encara temes oberts de recerca.

La matriu de dades del mètode PRAM conté una fila per cada possible valor de cadascuna de les variables que s'han de protegir; la qual cosa, pràcticament impossibilita la utilització d'aquest mètode en dades contínues.

2.7.6 Pèrdua per compressió

Aquest mètode és molt recent i encara és objecte de recerca pels seus autors. Se'l pot considerar com una manera "ingeniosa" d'afegir una pertorbació a les dades originals.

Sigui $v_{ij} = V_i(o_j)$ el valor original de la variable V_i per a l'individu o_j. Sigui $V = \{v_{ij}\}$ la matriu de microdades, on suposem que s'han estandarditzat a un mateix rang tots els valors v_{ij}. Així, la matriu V es pot considerar com una "imatge", on la columna i-èsima està formada per $V_i(o_j)$, per a tots els objectes o_j; i la fila j-èsima està formada per $V_i(o_j)$, per a totes les variables V_i.

La pèrdua per compressió (Joint Photographic Experts Group `http://www.jpeg.org`), o més específicament l'algorisme JPEG, s'aplica sobre V com si fos una vertadera imatge. La pèrdua per compressió és més alta en les "regions" de la imatge on els valors canvien més ràpidament. L'avantatge d'aquest mètode és que la pèrdua es distribueix de forma que minimitzi el dany infringit a la semàntica de la imatge.

En aquest mètode, la pèrdua d'informació i el risc de revelació òbviament depenen del nivell de compressió, el qual pot variar des del 0% fins al 100%. A més compressió, tindrem més pèrdua d'informació i menys risc de revelació. Cal observar que l'ordenació de les variables V_i i dels individus o_j abans d'aplicar la compressió també influeixen sobre la pèrdua d'informació i sobre el risc de revelació.

El mètode de Pèrdua per compressió ha estat dissenyat per protegir microdades contínues. De fet, per poder aplicar aquest mètode, les variables han de ser numèriques i contínues.

2.7.7 Microagregació

La microagregació (Defays i Anwar 1995) és una família de tècniques del control de la revelació basades en la modificació de les dades i especialment dissenyades per protegir microdades contínues. El principi subjacent a la microagregació és que les regles de confidencialitat en ús permeten la publicació de conjunts de microdades si els vectors de dades corresponen a grups de k o més individus, on cap individu no domina (és a dir, no contribueix massa) el grup. L'aplicació d'aquestes regles de confidencialitat condueix a la substitució de valors individuals per valors calculats sobre petits agregats (microagregats) com a pas previ a la publicació de les dades.

Així doncs, els registres s'agrupen en microagregats o grups de mida, almenys, k. En lloc de publicar el valor original d'una variable X_i per a un determinat registre, es publica la mitjana dels valors X_i que pren la variable dintre el grup al qual pertany el registre (Defays i Nanopoulos 1993)(Anwar 1993). Per tal de minimitzar la pèrdua d'informació, els grups haurien de ser el més homogenis possible.

El problema respecte a la partició de la població d'individus és diferent del clàssic problema de *clustering* on la finalitat és dividir la població en un nombre fixat de conjunts disjunts (Hartigan 1975), independentment de la mida dels grups. En la partició resultat d'un procés de microagregació, cap grup pot tenir una mida més petita que k; anomenarem aquestes particions, k-particions. D'altra banda, per resoldre el problema de la k-partició òptima (Ward 1963) serà necessària una mesura sobre la similaritat entre vectors de dades. Si considerem cada vector de dades com un punt i el conjunt de totes les microdades com un conjunt multidimensional de punts on la dimensió és el número de variables observades, la similaritat entre vectors de dades es pot mesurar a través d'una distància.

Per ser més específics, considerem un conjunt de microdades amb p variables contínues i n vectors de dades (és a dir, el resultat d'observar p variables sobre n individus). Un determinat vector de dades es pot representar, per exemple, com $\mathbf{X}' = (X_1, \ldots, X_p)$, on X_i són els valors de les variables. Tots aquests individus es distribueixen en g grups, de manera que el i-èsim grup tingui n_i individus ($n_i \geq k$ i $n = \sum_{i=1}^{g} n_i$). Representarem com $\mathbf{x}_{ij}$ el j-èsim vector de dades del i-èsim grup; representarem com $\bar{\mathbf{x}}_i$ la mitjana corresponent al grup i-èsim, i com $\bar{\mathbf{x}}$ la mitjana calculada sobre tots els n individus.

Definim la suma de quadrats intra-grups SSE de la següent manera:

$$SSE = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)'(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)$$

Definim, d'altra banda, la suma de quadrats inter-grups SSA com:

$$SSA = \sum_{i=1}^{g} n_i (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})'(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})$$

La suma de quadrats total és $SST = SSA + SSE$, on

$$SST = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}})'(\mathbf{x}_{ij} - \bar{\mathbf{x}})$$

Considerant la suma de quadrats intra-grups SSE com una mesura sobre la pèrdua d'informació (Gordon i Henderson 1977), la k-partició òptima serà la que minimitzi SSE (o equivalentment, maximitzi SSA).

Podem definir també una mesura L de pèrdua d'informació estandarditzada entre 0 i 1 de la següent manera:

$$L = \frac{SSE}{SST}$$

Direm que la microagregació és univariant (Domingo-Ferrer i Mateo-Sanz 1998) quan el conjunt original de dades a microagregar s'emmarca dintre d'una de les tres situacions següents:

1. El conjunt original de dades a microagregar està format per una sola variable ($p = 1$).

2. El conjunt original de dades és multivariant, però s'agafen de manera independent les diverses variables a l'hora de fer microagregació, és a dir, els grups d'individus que s'obtindran per a una variable poden ser diferents dels obtinguts per a una altra variable. En aquest cas, es fa microagregació variable a variable (ordenació individual).

3. Tot i ser també multivariant el conjunt original de dades, es consideren les projeccions d'aquestes dades originals sobre un eix; i es treballa, de manera exclusiva, amb aquestes projeccions per tal de fer microagregació. Generalment s'estudien tres tipus de projecció del conjunt original de dades sobre un eix:

   Projecció sobre una determinada variable. Tot i que la variable sobre la qual es projecten els vectors de dades sigui molt representativa dels registres individuals, per a les altres variables, els agrupaments de k vectors obtinguts no necessàriament seran d'individus semblants (això dependrà de la correlació entre les variables). Per la qual cosa, les variables sobre les quals no s'han projectat els vectors de dades tindran generalment més pèrdua d'informació que la variable emprada per projectar.

   Projecció sobre la primera component principal. És raonable esperar que la projecció sobre la primera component principal produeixi grups més homogenis que la projecció sobre una determinada variable, ja que les variables es combinen de manera que la primera component principal es troba altament correlacionada amb la majoria de variables originals.

Projecció sobre la suma de les z-puntuacions. Aquest mètode, juntament amb l'anteriorment comentat sobre la primera component principal, considera totes les variables dels vectors de dades per obtenir els punts projecció. La diferència rau en què la suma de les z-puntuacions dóna la mateixa importància a totes les variables, mentre que en la primera component principal la importància de cada variable depèn de l'estructura de correlació. S'aconsella estandarditzar totes les variables abans de calcular la suma de components de cada vector de dades, puix que pot haver molta diferència pel que fa als valors de cada variable respecte les altres.

Direm que la microagregació és multivariant, quan el conjunt de dades a microagregar és multivariant i es treballa directament sobre els vectors de dades sense projectar.

Capítol 3

Seguretat de la microagregació amb ordenació individual

3.1 Conceptes sobre microagregació

La microagregació clàssica (Defays i Nanopoulos 1993, Defays i Anwar 1995, Anwar 1993, Hundepool, Willenborg, Wessels, Gemerden, Tiourine i Hurkens 1998) demana que tots els grups siguin de mida fixa k, excepte un que seria de mida $\geq k$. Tot i això, parlarem de *microagregació orientada a les dades* (Mateo-Sanz i Domingo-Ferrer 1999, Domingo-Ferrer i Mateo-Sanz 2002, Sande 2001), quan tots els grups poden ser de mida $\geq k$, depenent de l'estructura del conjunt de dades originals. En la figura 3.1 podem observar els avantatges de fer grups de mida variable per a un conjunt de dades bidimensional; puix que, utilitzant la microagregació clàssica de mida fixa amb $k = 3$, obtenim una partició de les dades en tres grups, que sembla molt poc adequada per a la distribució de dades donada. Tanmateix, fent grups de mida variable, es poden ajuntar les cinc dades situades més a l'esquerra de la figura 3.1 en un mateix grup, i les quatre dades restants en un altre grup. D'aquesta manera, obtenim grups molt més homogenis amb una menor pèrdua d'informació.

Els mètodes de microagregació poden ser classificats en mètodes univariants o bé multivariants, tot depenent de si tracten una sola variable o bé diverses variables simultàniament:

- Els **mètodes univariants** poden tractar dades multivariants de dues diferents maneres:
  1. Mitjançant projecció de les dades multivariants sobre un eix, utilitzant la primera component principal, o bé la suma de les "z-puntuacions" o inclús una determinada variable (Defays i Nanopoulos 1993, Defays i Anwar 1995). D'aquesta manera, les dades projectades poden ser considerades com un conjunt de dades univariants i ser tractades a través de microagregació univariant.
  2. Microagregant una sola variable cada vegada, és a dir, les variables es microagreguen seqüencialment i independentment unes de les altres. Aquesta aproximació a la microagregació es coneguda com *ordenació individual* (Defays i Nanopoulos 1993), i és, precisament, l'objecte del present estudi.

  Tot i que l'*ordenació individual* produeix poca pèrdua d'informació, mostrarem que el seu risc de revelació és inacceptablement alt. Els nostres resultats analítics obtinguts en aquest estudi confirmen els resultats empírics obtinguts en Domingo-Ferrer i Torra (2001) utilitzant la tècnica d'enllaç de registres sobre dos conjunts de dades.

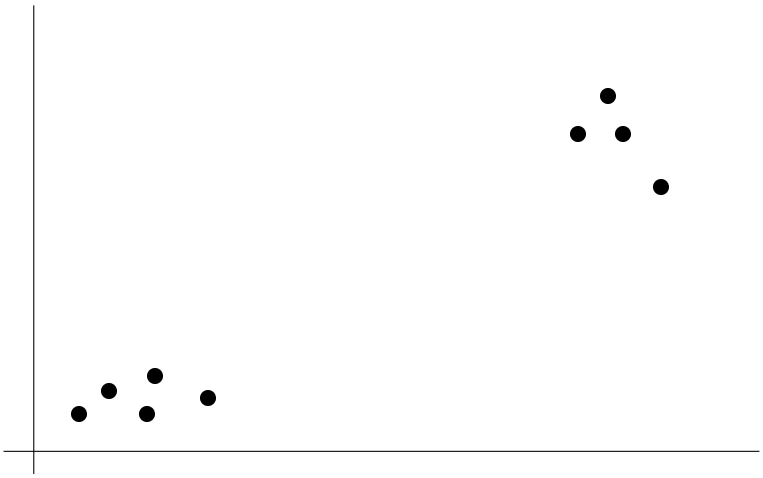


Figura 3.1: Grups de mida variable versus grups de mida fixa

- Els **mètodes multivariants** treballen directament sobre les dades multivariants sense projectar (Mateo-Sanz i Domingo-Ferrer 1999, Domingo-Ferrer i Mateo-Sanz 2002, Sande 2001). En aquest cas, es poden microagregar conjuntament totes les variables del conjunt de dades, o, independentment, microagregar grups de dos variables a la vegada, tres variables a la vegada, etc.

Resoldre exactament el problema de la microagregació en el cas multivariant sense projecció, és a dir, trobar la manera de microagregar perquè els grups tinguin la màxima homegeneitat i siguin de mida almenys k en un espai Euclidi de dimensió més gran o igual que 2, ha estat provat que és NP-hard (Oganian i Domingo-Ferrer 2001). D'altra banda, ha estat recentment provat que el problema de la microagregació univariant que apareix en ambdues aproximacions: ordenació individual i dades projectades, és polinomialment resoluble (Hansen i Mukherjee 2002).

Però, pràcticament, la microagregació univariant, tot i la seva més baixa complexitat, no és molt atractiva perquè o bé, de vegades, conté alt risc de revelació (això, justament, provarem en aquest estudi per al cas d'ordenació individual), o bé, contràriament, conté alta pèrdua d'informació (provocada per la projecció en el cas de dades projectades).

Els resultats empírics de l'estudi Domingo-Ferrer i Torra (2001) suggereixen que la microagregació multivariant sobre dades no projectades ofereix un millor equilibri entre la pèrdua d'informació i el risc de revelació, especialment quan es microagreguen grups de tres o quatre variables a la vegada (millor que totes simultàniament).

3.2 Seguretat de la microagregació utilitzant ordenació individual

Perquè un mètode de microagregació sigui *segur*, no ha de ser possible estimar, amb gran precisió, qualsevol valor del conjunt de dades originals, a partir del conjunt de dades microagregades. L'ordenació individual, certament, és bastant utilitzada per a microagregar un conjunt de dades

multivariants (Hundepool, Willenborg, Wessels, Gemerden, Tiourine i Hurkens 1998)(Baeyens i Defays 1999). Amb *ordenació individual*, cada variable es considera independentment de les altres. Els vectors de dades s'ordenen per la seva primera variable; després, es formen grups de k valors successius de la primera variable i, dintre de cada grup, els seus valors es substitueixen per la mitjana del grup. A continuació, es repeteix aquest mateix procediment per a la resta de variables.

L'*ordenació individual* no utilitza una única partició dels n vectors de dades per a microagregar, sinó que construeix una partició diferent per a cada variable que es microagrega.

De vegades utilitzarem el terme "element" en lloc de "vector de dades", per reflectir el fet que l'ordenació individual no tracta tot el vector de dades a la vegada, sinó que senzillament tracta elements univariants (els valors d'una variable cada cop).

La popularitat de l'*ordenació individual* és deguda a la seva simplicitat computacional i al fet que, generalment, conserva més informació que la projecció sobre un eix (Mateo-Sanz i Domingo-Ferrer 1998, Domingo-Ferrer i Torra 2001). Tot i això, fa més fàcil que un intrús estimi dades originals a partir de les dades microagregades. Certament, amb l'*ordenació individual*, qualsevol intrús sap que el valor original d'un element del grup i-èsim està situat entre la mitjana del grup $(i-1)$-èsim i la mitjana del grup $(i+1)$-èsim; si aquestes dues mitjanes estan molt properes una a l'altra, llavors l'intrús ha aconseguit trobar un interval molt estret (precís) per a la dada original que, suposadament, està buscant.

A més a més, l'*ordenació individual* és vulnerable a un atac, potser menys obvi, basat sobre els estadístics d'ordre de les variables ordenades. Suposem que el conjunt de dades original conté almenys una variable aleatòria X que segueix una distribució contínua amb funció de distribució $F(x)$ i funció de densitat $f(x)$.

Sigui $X_1, \ldots, X_n$ una mostra aleatòria treta d'X.

Siguin

$$X_{1:n} \leq X_{2:n} \leq \ldots \leq X_{n:n}$$

els successius estadístics d'ordre que resulten d'ordenar l'anterior mostra en ordre ascendent. En altres paraules: el succés

$$x < X_{i:n} \leq x + \Delta x$$

per a un suficientment petit Δx, és equivalent a dir que $i-1$ valors de la mostra són més petits o iguals que x, exactament un valor de la mostra es troba dintre l'interval $(x, x+\Delta x]$, i la resta de $n-i$ valors són més grans que $x + \Delta x$.

En general, per a qualsevol Δx, és ben conegut (Arnold, Balakrishnan i Nagaraja 1993) que

$$P(x < X_{i:n} \leq x + \Delta x) = \tag{3.1}$$

$$\frac{n!}{(i-1)!(n-i)!} \cdot [F(x)]^{i-1} \cdot [1 - F(x+\Delta x)]^{n-i} \cdot [F(x+\Delta x) - F(x)] + o(\Delta x) \tag{3.2}$$

on $o(\Delta x)$ és la probabilitat del succés "més d'un valor de la mostra es troba dintre l'interval $(x, x+\Delta x]$". Així doncs, derivant l'equació (3.1), es pot obtenir la funció de densitat de l'i-èsim estadístic d'ordre $X_{i:n}$:

$$f_{i:n}(x) = \lim_{\Delta x \to 0} \frac{P(x < X_{i:n} \leq x + \Delta x)}{\Delta x} =$$

$$= \frac{n!}{(i-1)!(n-i)!} \cdot [F(x)]^{i-1} \cdot [1-F(x)]^{n-i} \cdot f(x) \tag{3.3}$$

on $-\infty < x < +\infty$.

D'aquesta manera, hem aconseguit trobar dues propietats importants de l'i-èsim estadístic d'ordre $X_{i:n}$:

- La seva funció de densitat $f_{i:n}(x)$. En sentit Bayesià (Box i Tiao 1992), aquesta es pot considerar com la funció de densitat de la probabilitat anterior de $X_{i:n}$.
- L'interval on es troba, ja que $X_{i:n}$ es troba entre la mitjana a_{-1} del grup immediatament anterior al seu propi grup en el conjunt ordenat de les dades i la mitjana a_1 del grup immediatament posterior al seu propi grup en el conjunt ordenat de les dades.

Combinant les dues propietats anteriors, es pot obtenir la funció de densitat de la probabilitat posterior de $X_{i:n}$ condicionada a l'interval $[a_{-1}, a_1]$, que és $f_{i:n}(x \mid a_{-1} \leq X_{i:n} \leq a_1)$.

Així, donat un coeficient de confiança α, es pot calcular l'interval de probabilitat posterior

$$[x_{\alpha/2}, x_{1-\alpha/2}] \tag{3.4}$$

per a $X_{i:n}$, a partir de la seva funció de densitat posterior, puix que

$$\alpha/2 = \int_{a_{-1}}^{x_{\alpha/2}} f_{i:n}(y \mid a_{-1} \leq X_{i:n} \leq a_1)\, dy \tag{3.5}$$

$$1-\alpha/2 = \int_{a_{-1}}^{x_{1-\alpha/2}} f_{i:n}(y \mid a_{-1} \leq X_{i:n} \leq a_1)\, dy \tag{3.6}$$

Per resoldre l'equació (3.5) per a $x_{\alpha/2}$, es pot utilitzar el mètode de la bisecció, combinat possiblement amb integració numèrica. El mateix procediment es pot utilitzar per resoldre l'equació (3.6) per a $x_{1-\alpha/2}$. L'interval $[x_{\alpha/2}, x_{1-\alpha/2}]$ sempre serà més estret (precís) que l'interval obvi $[a_{-1}, a_1]$, i pot ser substancialment més estret per a i proper a 1 o a n (és a dir, per als valors extrems del conjunt de dades original; veure Seccions 3.3, 3.4 i 3.5).

L'atac anterior pot ser encara més feridor si es coneix que $X_{i:n}$ correspon al valor més petit o bé més gran de la dades originals d'un grup microagregat. Si $X_{i:n}$ és el valor més petit de les dades originals d'un grup la mitjana del qual és a_0, llavors sabem que $X_{i:n}$ es troba entre la mitjana a_{-1} del grup immediatament anterior al seu grup i la mitjana a_0 del seu propi grup. De manera semblant, si $X_{i:n}$ és el valor més gran de les dades originals d'un grup la mitjana del qual és a_0, llavors sabem que $X_{i:n}$ es troba entre la mitjana a_0 del seu propi grup i la mitjana a_1 del grup immediatament posterior al seu grup. Observem que $[a_{-1}, a_0] \subseteq [a_{-1}, a_1]$ i $[a_0, a_1] \subseteq [a_{-1}, a_1]$, de manera que en ambdós casos obtenim intervals més estrets (més precisos) per a $X_{i:n}$. Per tant:

- Quan es coneix que $X_{i:n}$ és el valor més petit de les dades originals d'un grup, llavors, a partir de la funció de densitat posterior $f_{i:n}(x \mid a_{-1} \leq X_{i:n} \leq a_0)$, es pot calcular un interval de probabilitat posterior

  $$[x'_{\alpha/2}, x'_{1-\alpha/2}] \tag{3.7}$$

  que és més estret que l'interval de probabilitat posterior (3.4) que resulta de les equacions (3.5) i (3.6).

- Quan es coneix que $X_{i:n}$ és el valor més gran de les dades originals d'un grup, llavors, a partir de la funció de densitat posterior $f_{i:n}(x \mid a_0 \leq X_{i:n} \leq a_1)$, es pot calcular un interval de probabilitat posterior

$$[x''_{\alpha/2}, x''_{1-\alpha/2}] \tag{3.8}$$

que és més estret que l'interval de probabilitat posterior (3.4) que resulta de les equacions (3.5) i (3.6).

A la pràctica els atacs descrits en aquesta secció es desenvoluparien de la següent manera:

1. Per a cada variable s'estimaria la distribució de les dades originals a partir de la distribució empírica de les dades microagregades (únicament es fan públiques les dades microagregades). Quan s'utilitza *l'ordenació individual*, les variables es microagreguen independentment unes de les altres; la qual cosa significa que la distribució subjacent de les dades originals està ben reflectida per les dades microagregades publicades.
2. A partir de la distribució estimada de les dades originals, es deduïrien les expressions per a les funcions de densitat posteriors
$$f_{i:n}(x \mid a_{-1} \leq X_{i:n} \leq a_1)$$
$$f_{i:n}(x \mid a_{-1} \leq X_{i:n} \leq a_0)$$
$$f_{i:n}(x \mid a_0 \leq X_{i:n} \leq a_1)$$
i els seus corresponents intervals de probabilitat posterior.

3.3 Estudi analític per a dades contínues uniformement distribuïdes

Si X segueix una distribució uniforme contínua restringida a l'interval $[0, 1]$ (simbolitzada per $U[0, 1]$), la funció de densitat de $X_{i:n}$ és

$$f_{i:n}(x) = \frac{n!}{(i-1)!(n-i)!} \cdot x^{i-1} \cdot (1-x)^{n-i} \tag{3.9}$$

$$0 \leq x \leq 1$$

Cal observar que la funció de densitat de l'equació (3.9) és precisament la funció de densitat d'una distribució Beta amb paràmetres $(i, n-i+1)$.

Puix que coneixem l'interval $[a_{-1}, a_1] \subset [0, 1]$ on es troba $X_{i:n}$, podem deduir la funció de densitat de $X_{i:n}$ restringida a aquest interval.

Per a $x \in [a_{-1}, a_1]$

$$f_{i:n}(x \mid a_{-1} \leq X_{i:n} \leq a_1) =$$
$$= \frac{\frac{n!}{(i-1)!(n-i)!} \cdot x^{i-1} \cdot (1-x)^{n-i}}{\int_{a_{-1}}^{a_1} \frac{n!}{(i-1)!(n-i)!} \cdot x^{i-1} \cdot (1-x)^{n-i}\, dx} = \frac{x^{i-1} \cdot (1-x)^{n-i}}{\int_{a_{-1}}^{a_1} x^{i-1} \cdot (1-x)^{n-i}\, dx} \tag{3.10}$$

Per a $x \notin [a_{-1}, a_1]$

$$f_{i:n}(x \mid a_{-1} \leq X_{i:n} \leq a_1) = 0$$

Quan es coneix que $X_{i:n}$ és el valor més petit o bé el més gran de les dades d'un grup, es poden obtenir expressions anàlogues a (3.10), senzillament substituint l'interval $[a_{-1}, a_1]$ per l'interval $[a_{-1}, a_0]$, per al valor més petit; i per l'interval $[a_0, a_1]$, per al valor més gran.

El fet que $X_{i:n}$ segueix una distribució $Beta(i, n-i+1)$, condueix a altres resultats teòrics que són molt útils per valorar analíticament la seguretat de la microagregació amb ordenació individual sobre una distribució $U[0,1]$. Concretament:

- El valor esperat per a l'i-èsim estadístic d'ordre és:

$$E(X_{i:n}) = E[Beta(i, n-i+1)] = \frac{i}{n+1}$$

- El grup j-èsim està format per $X_{kj-(k-1):n}$, $X_{kj-(k-2):n}$, $\cdots$, $X_{kj:n}$. Per tant, el valor esperat de la mitjana a_j del grup j-èsim és:

$$E(a_j) = \frac{\sum_{l=0}^{k-1} \frac{kj-l}{n+1}}{k} = \frac{\sum_{l=0}^{k-1}(kj-l)}{k(n+1)} = \frac{k^2 j - \frac{k(k-1)}{2}}{k(n+1)} = \frac{k(2j-1)+1}{2(n+1)}$$

- La diferència esperada entre a_{j+1} i a_j (mitjanes del grup $(j+1)$-èsim i del grup j-èsim) és:

$$E(a_{j+1} - a_j) = \frac{k[2(j+1)-1]+1}{2(n+1)} - \frac{k(2j-1)+1}{2(n+1)} = \frac{k}{n+1} \tag{3.11}$$

Així doncs, la diferència esperada (3.11) no depén del valor j que es consideri. Per la qual cosa, si a_{-1}, a_0 i a_1 són les mitjanes de tres grups consecutius qualssevol, tindrem que:

$$\begin{aligned} E(a_1 - a_{-1}) &= \frac{2k}{n+1} \\ E(a_1 - a_0) &= E(a_0 - a_{-1}) = \frac{k}{n+1} \end{aligned}$$

Amb els anteriors resultats, es poden determinar analíticament els intervals (3.4), (3.7) i (3.8), i, després d'això, es pot calcular en quina mesura aquests intervals redueixen els intervals obvis $[a_{-1}, a_1]$, $[a_{-1}, a_0]$ i $[a_0, a_1]$ respectivament, on a_{-1} és la mitjana del grup immediatament anterior al grup que es considera, a_0 és la mitjana del grup central que es considera i a_1 és la mitjana del grup immediatament posterior al grup que es considera. Per comparar l'interval (3.4) amb $[a_{-1}, a_1]$, calculem:

$$q = 100 \frac{x_{1-\alpha/2} - x_{\alpha/2}}{a_1 - a_{-1}} \tag{3.12}$$

S'ha calculat aquest valor q per a diferents valors d'α, k, n i per a diferents estadístics d'ordre $X_{i:n}$. Els valors q que apareixen a la Taula 3.1 (veure secció 7) corresponen a grups; per a cada grup, s'ha calculat el valor mitjà dels k estadístics d'ordre (elements) que el formen. Calcular aquest valor mitjà dels k estadístics d'ordre que formen cada grup és el més adequat en aquest cas, perquè l'intrús no coneix l'ordre dels vectors de dades dintre d'un grup (un fitxer de dades microagregades només

conté la mitjana de cada grup, repetida k vegades). La Taula 3.1 conté els q-valors per al grup $G_{(1)}$ amb els k valors més petits, i per als grups ordenats corresponents als percentils 5, 15, 30 i 50 sobre el número total de grups (aquests grups es denoten $G_{5\%}$, $G_{15\%}$, $G_{30\%}$ i $G_{50\%}$, respectivament). Cal observar que, degut a la simetria de la distribució uniforme, els resultats per al grup $G_{P\%}$ són similars als resultats per al grup $G_{(100-P)\%}$.

Quan es coneix que $X_{i:n}$ és el valor més petit dels elements del seu grup, llavors trobarem l'interval de probabilitat posterior més estret (més precís) donat per l'expressió (3.7). Compararem aquest interval amb l'interval obvi $[a_{-1}, a_0]$, mesurant la millora aconseguida a través de:

$$q' = 100\frac{x'_{1-\alpha/2} - x'_{\alpha/2}}{a_0 - a_{-1}} \tag{3.13}$$

La Taula 3.2 conté els valors q' corresponents als intervals per a l'element més petit de cada grup (de diferent manera, els valors q de la Taula 3.1 corresponien a valors mitjans dels k estadístics d'ordre de cada grup).

Quan es coneix que $X_{i:n}$ és el valor més gran dels elements del seu grup, llavors trobarem l'interval de probabilitat posterior donat per l'expressió (3.8). Compararem aquest interval amb l'interval obvi $[a_0, a_1]$, mesurant la millora aconseguida a través de:

$$q'' = 100\frac{x''_{1-\alpha/2} - x''_{\alpha/2}}{a_1 - a_0} \tag{3.14}$$

La Taula 3.3 conté els valors q'' corresponents als intervals per a l'element més gran de cada grup (també diferents als valors q de la Taula 3.1, que corresponien a valors mitjans dels k estadístics d'ordre de cada grup). Per ser més breus, a la Taula 3.3 no apareixen els q''-valors que han estat molt similars als corresponents q'-valors de la Taula 3.2.

3.4 Estudi de simulació per a la distribució Normal

Si X segueix una distribució $N(0,1)$, la funció de densitat d'$X_{i:n}$ és:

$$f_{i:n}(x) =$$

$$= \frac{n!}{(i-1)!(n-i)!} \cdot \left(\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}\, dz\right)^{i-1} \cdot \left(1 - \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}\, dz\right)^{n-i} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{3.15}$$

per a $-\infty < x < +\infty$.

Novament coneixem l'interval $[a_{-1}, a_1]$ on es troba $X_{i:n}$, i podem deduir la funció de densitat de $X_{i:n}$ restringida a aquest interval:

Per a $x \in [a_{-1}, a_1]$

$$f_{i:n}(x \mid a_{-1} \leq X_{i:n} \leq a_1) =$$

$$= \frac{\left(\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}\, dz\right)^{i-1} \cdot \left(1 - \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}\, dz\right)^{n-i} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}}{\int_{a_{-1}}^{a_1} \left(\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}\, dz\right)^{i-1} \cdot \left(1 - \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}\, dz\right)^{n-i} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx} \tag{3.16}$$

Per a $x \notin [a_{-1}, a_1]$

$$f_{i:n}(x \mid a_{-1} \leq X_{i:n} \leq a_1) = 0$$

Quan es coneix que $X_{i:n}$ és el valor més petit o bé el més gran dels elements del seu grup, es poden obtenir expressions anàlogues a (3.16), senzillament substituint l'interval $[a_{-1}, a_1]$ per l'interval $[a_{-1}, a_0]$, per al valor més petit; i per $[a_0, a_1]$, per al valor més gran.

En aquest cas de la distribució normal, i per calcular expressions anàlogues a (3.12), (3.13) i (3.14), serà necessari fer una estimació dels intervals $[a_{-1}, a_1]$, $[a_{-1}, a_0]$ i $[a_0, a_1]$. Tanmateix, a diferència de la distribució uniforme de la secció anterior, no és fàcil estimar aquests intervals sense fer simulació, perquè les seves amplituds depenen de l'ordre dels elements. Tot i això, la simetria de la distribució normal implica algunes simetries per a aquests intervals:

1. La diferència $a_1 - a_{-1}$ per a $X_{i:n}$ i per a $X_{n-i+1:n}$ és semblant.
2. La diferència $a_0 - a_{-1}$ per a $X_{i:n}$ és semblant a la diferència $a_1 - a_0$ per a $X_{n-i+1:n}$.
3. La diferència $a_1 - a_0$ per a $X_{i:n}$ és semblant a la diferència $a_0 - a_{-1}$ per a $X_{n-i+1:n}$.

Procediment 1 (Procediment de simulació) *Hem considerat diferents valors de n i k (concretament $k = 3$, $k = 4$ i $k = 5$). S'ha generat un conjunt de 50 dades $N(0, 1)$ per a cada combinació de n i k; i, per a cada conjunt de dades, s'han estudiat els coeficients de confiança $\alpha = 0.1, 0.05, 0.01$.*

D'aquesta manera, donada una combinació de n, k i α, per a cada estadístic d'ordre $X_{i:n}$, s'han obtingut 50 amplituds d'intervals. Finalment, s'ha calculat la mitjana de les 50 amplituds d'intervals.

La Taula 3.4 mostra com varien les amplituds de l'interval $[a_{-1}, a_1]$ per als diferents valors de n i k, i per als diferents grups. S'han considerat els mateixos grups $G_{(1)}$, $G_{5\%}$, $G_{15\%}$, $G_{30\%}$ i $G_{50\%}$ de la Secció 3.3. Cal remarcar que, degut a la simetria de la distribució normal, els resultats per a $G_{P\%}$ són similars als resultats per a $G_{(100-P)\%}$.

A la Taula 3.4 podem observar els següents fets:

- En augmentar el número n d'elements, l'amplitud dels intervals disminueix.
- En augmentar la mida k dels grups, l'amplitud dels intervals augmenta.
- Els intervals dels grups més extrems (amb els valors dels elements més petits o bé més grans) tenen més amplitud que els intervals dels grups centrals.

Un cop han estat estimats els intervals $[a_{-1}, a_1]$, $[a_{-1}, a_0]$ i $[a_0, a_1]$ mitjançant simulació, calculem els intervals de probabilitat posterior (3.4), (3.7) i (3.8). El següent pas serà comparar i valorar en quina mesura aquests últims intervals redueixen l'amplitud dels primers intervals obvis. Per comparar l'interval (3.4) amb $[a_{-1}, a_1]$, trobem, per al cas normal, el coeficient q tal com ha estat definit a l'expressió (3.12). La Taula 3.5 és la versió de la Taula 3.1, per a dades normals.

Quan es coneix que $X_{i:n}$ és el valor més petit dels elements del seu grup, llavors calculem l'interval (3.7), que compararem amb $[a_{-1}, a_0]$; la millora aconseguida serà mesurada mitjançant el coeficient q' definit a l'expressió (3.13), i calculat per a dades normals. La Taula 3.6 és la versió de la Taula 3.2, per a dades normals.

Quan es coneix que $X_{i:n}$ és el valor més gran dels elements del seu grup, llavors calculem l'interval (3.8), que compararem amb $[a_0, a_1]$; la millora aconseguida serà mesurada mitjançant el coeficient q'' definit a l'expressió (3.14), i calculat per a dades normals.

La Taula 3.7 és la versió de la Taula 3.3 per a dades normals.

3.5 Estudi de simulació per a dades esbiaixades (Weibull)

Per completar el nostre estudi, presentem una tercera anàlisi sobre dades esbiaixades, concretament per a dades que segueixen una distribució Weibull.

Les dades contínues, i molt especialment les dades financeres i mercantils, són, de vegades, prou esbiaixades.

Si X segueix una distribució Weibull amb paràmetres $\alpha > 0$ i $\beta > 0$, la funció de densitat de $X_{i:n}$ és:

$$f_{i:n}(x) = \frac{n!}{(i-1)!(n-i)!} \cdot (1 - e^{-(\frac{x}{\alpha})^\beta})^{i-1} \cdot (e^{-(\frac{x}{\alpha})^\beta})^{n-i} \cdot \frac{\beta x^{\beta-1}}{\alpha^\beta} e^{-(\frac{x}{\alpha})^\beta} \tag{3.17}$$

per a $x > 0$.

La funció de densitat de $X_{i:n}$ restringida a l'interval $[a_{-1}, a_1]$ és:

Per a $x \in [a_{-1}, a_1]$

$$f_{i:n}(x \mid a_{-1} \leq X_{i:n} \leq a_1) =$$

$$= \frac{\frac{n!}{(i-1)!(n-i)!} \cdot (1 - e^{-(\frac{x}{\alpha})^\beta})^{i-1} \cdot (e^{-(\frac{x}{\alpha})^\beta})^{n-i} \cdot \frac{\beta x^{\beta-1}}{\alpha^\beta} e^{-(\frac{x}{\alpha})^\beta}}{\int_{a_{-1}}^{a_1} \frac{n!}{(i-1)!(n-i)!} \cdot (1 - e^{-(\frac{x}{\alpha})^\beta})^{i-1} \cdot (e^{-(\frac{x}{\alpha})^\beta})^{n-i} \cdot \frac{\beta x^{\beta-1}}{\alpha^\beta} e^{-(\frac{x}{\alpha})^\beta} dx} \tag{3.18}$$

per a $x \notin [a_{-1}, a_1]$

$$f_{i:n}(x \mid a_{-1} \leq X_{i:n} \leq a_1) = 0$$

Quan es coneix que $X_{i:n}$ és el valor més petit o bé el més gran dels elements del seu grup, es poden obtenir expressions anàlogues a (3.18).

Per calcular els coeficients (3.12), (3.13) i (3.14) en el cas de dades Weibull, haurem de fer una estimació dels intervals $[a_{-1}, a_1]$, $[a_{-1}, a_0]$ i $[a_0, a_1]$. Com en el cas de dades normals de la secció anterior, també, amb dades Weibull, serà necessària la simulació per fer aquestes estimacions.

Hem desenvolupat un procediment simulador, per a dades Weibull amb $\alpha = 1$ i $\beta = 1.5$, similar a l'Algorisme 1.

La Taula 3.8 és la versió de la Taula 3.4 per a dades Weibull. Donat que aquesta distribució no és simètrica, hem considerat els grups $G_{(1)}$, $G_{10\%}$, $G_{50\%}$, $G_{90\%}$ i $G_{100\%}$.

A la Taula 3.8 podem observar els següents fets:

- En augmentar el número n d'elements, l'amplitud dels intervals disminueix.
- En augmentar la mida k dels grups, l'amplitud dels intervals augmenta.

- A causa del biaix de la distribució Weibull, els intervals per als grups dels elements més grans són molt amplis.

Les Taules 3.9, 3.10 i 3.11 són les versions de les Taules 3.1, 3.2 i 3.3 per a dades Weibull.

3.6 Conclusions

De tot l'anterior estudi, i de l'observació de les taules de la secció següent, podem concloure que l'atac mitjançant estadístics d'ordre contra la microagregació amb ordenació individual, és bastant efectiu per a dades uniformement distribuïdes, per a dades normals i per a dades esbiaixades (Weibull).

Tot i això, en comparar les taules corresponents a les tres distribucions, podem observar que els intervals de probabilitat posterior obtinguts per a dades normals i dades Weibull, produeixen més reducció de l'amplitud sobre els intervals trivials ($[a_{-1}, a_1]$,$[a_{-1}, a_0]$,$[a_0, a_1]$), que els obtinguts per a dades uniformes. Per tant, l'atac és encara més efectiu quan les dades originals són normals, o bé, Weibull; les quals, a la vegada, són distribucions que s'ajusten molt bé a les dades financeres contínues ordinàries.

A més a més, quan les dades són normals o bé Weibull, els intervals trivials per als estadístics d'ordre que ocupen les posicions centrals, són prou més estrets per si mateixos; la qual cosa és un factor d'inseguretat afegit per a aquestes dues distribucions.

A continuació, detallarem alguns trets comuns als resultats de les Seccions 3.3, 3.4 i 3.5. El fet que les següents observacions siguin comunes per a dades uniformes, per a dades normals i per a dades esbiaixades (Weibull), ens empeny a dir que, versemblantment, també es mantindran per a altres distribucions de les dades originals:

- Els intervals de probabilitat posterior obtinguts per als estadístics d'ordre extrems (els valors més petits o bé els valors més grans) són especialment més estrets que els seus corresponents intervals trivials. *Això implica que l'atac és molt més efectiu quan s'aplica per estimar els valors extrems; la qual cosa és especialment preocupant dins del marc de la confidencialitat estadística, on habitualment s'aplica la microagregació amb ordenació individual*: una estimació prou precisa d'un valor extrem, fa que sigui molt fàcil la identificació de l'individu a qui pertany.

  Com a exemple, suposem que la microagregació amb ordenació individual ha estat aplicada per protegir un conjunt de microdades que contenen la variable "Edat": si el nostre atac produeix un interval de probabilitat posterior $[98, 101]$, per a l'edat d'un determinat individu, llavors, és molt fàcil identificar la persona de qui es tracta.

- Els intervals de probabilitat posterior obtinguts per als estadístics d'ordre que ocupen les posicions centrals (els situats entre els percentils 15 i 85), no produeixen molta millora sobre els seus corresponents intervals trivials. Més concretament, l'amplitud dels intervals de probabilitat posterior per als estadístics d'ordre centrals, aproxima l'amplitud dels seus corresponents intervals trivials en un percentatge de $(1 - \alpha)100$.

  Per exemple, observant les Taules 3.1, 3.5 i 3.9, per a $n = 1000$, $\alpha = 0.1$ i $G_{50\%}$, veiem amplituds relatives de 89.9% per a dades uniformes, 89.5% per a dades normals i 89.6% per a dades Weibull; tots aquests valors són similars al $90\% = (1 - \alpha)100\%$.

- Per a un número n d'elements fixat, en augmentar la mida k dels grups des de 3 fins a 5, l'efectivitat relativa de l'atac (quantitat d'amplitud que els intervals de probabilitat posterior redueixen respecte dels intervals trivials) també augmenta. Així doncs, en augmentar la mida k dels grups, com que els intervals trivials tendeixen a fer-se més grans, llavors, l'efectivitat

absoluta de l'atac (inversament proporcional a l'amplitud dels intervals de probabilitat posterior) no creix necessàriament de forma proporcional.

- Versemblantment, quan α disminueix, l'efectivitat de l'atac també disminueix. En altres paraules, en exigir molta més precisió de l'atac, els intervals de probabilitat posterior es fan més amples, disminuint la millora sobre els intervals trivials.

Per totes aquestes raons, sembla prou justificat buscar altres aproximacions a la microagregació diferents a l'ordenació individual. Per a valors de k petits (entre 3 i 5, tal com habitualment utilitzen les oficines d'estadística), hem provat que microagregar independentment variable a variable (ordenació individual) condueix a resultats massa transparents, que no ofereixen suficient seguretat. Per a valors de k més grans, únicament disposem de resultats empírics basats en experiments sobre enllaç de registres (Domingo-Ferrer i Torra 2001); i tals resultats no mostren cap mena de millora.

La microagregació multivariant que considera totes les variables simultàniament, bé sigui projectant-les sobre un eix, bé sigui tractant directament les dades no projectades, ofereix menys risc de revelació, tot i que produeix més pèrdua d'informació que l'ordenació individual. De fet, els resultats empírics de Domingo-Ferrer i Torra (2001) indiquen que la projecció de les dades produeix més pèrdua d'informació; però, certament, algunes versions de la microagregació multivariant aconsegueixen un bon equilibri entre el risc de revelació i la pèrdua d'informació.

3.7 Taules de resultats

Taula 3.1: Mitjana del percentatge de l'amplitud relativa de l'interval (3.4) sobre $[a_{-1}, a_1]$ (dades uniformes)

		$k=3$				$k=4$				$k=5$			
		n				n				n			
		200	400	700	1000	200	400	700	1000	200	400	700	1000
$\alpha = 0.1$	$G_{(1)}$	81.2	81.2	81.3	81.3	76.9	77.1	77.2	77.2	72.7	73.0	73.2	73.2
	$G_{5\%}$	86.8	88.4	89.2	89.4	83.0	86.6	88.4	88.9	80.5	83.9	87.2	88.2
	$G_{15\%}$	89.0	89.5	89.7	89.8	87.9	89.0	89.5	89.6	86.1	88.3	89.1	89.4
	$G_{30\%}$	89.4	89.7	89.8	89.9	88.8	89.4	89.7	89.8	88.0	89.1	89.5	89.6
	$G_{50\%}$	89.5	89.7	89.9	89.9	89.0	89.5	89.7	89.8	88.4	89.2	89.6	89.7
$\alpha = 0.05$	$G_{(1)}$	89.5	89.5	89.6	89.6	86.1	86.2	86.3	86.3	82.4	82.7	82.8	82.8
	$G_{5\%}$	93.2	94.1	94.5	94.7	90.8	93.1	94.1	94.4	89.0	91.4	93.4	94.0
	$G_{15\%}$	94.4	94.7	94.8	94.9	93.8	94.5	94.7	94.8	92.8	94.1	94.5	94.7
	$G_{30\%}$	94.7	94.8	94.9	94.9	94.3	94.7	94.8	94.9	93.9	94.5	94.7	94.8
	$G_{50\%}$	94.7	94.9	94.9	94.9	94.5	94.7	94.9	94.9	94.1	94.6	94.8	94.8
$\alpha = 0.01$	$G_{(1)}$	97.6	97.6	97.6	97.6	96.3	96.4	96.4	96.5	94.5	94.6	94.7	94.7
	$G_{5\%}$	98.6	98.8	98.9	98.9	98.0	98.6	98.8	98.9	97.5	98.2	98.6	98.8
	$G_{15\%}$	98.9	98.9	99.0	99.0	98.7	98.9	98.9	99.0	98.5	98.8	98.9	98.9
	$G_{30\%}$	98.9	99.0	99.0	99.0	98.9	98.9	99.0	99.0	98.8	98.9	98.9	99.0
	$G_{50\%}$	98.9	99.0	99.0	99.0	98.9	98.9	99.0	99.0	98.8	98.9	99.0	99.0

Taula 3.2: Percentatge de l'amplitud relativa de l'interval (3.7) sobre $[a_{-1}, a_0]$ (dades uniformes)

		$k=3$				$k=4$				$k=5$			
		n				n				n			
		200	400	700	1000	200	400	700	1000	200	400	700	1000
$\alpha = 0.1$	$G_{(1)}$	88.0	88.0	88.0	88.0	87.0	87.1	87.1	87.1	86.0	86.1	86.1	86.1
	$G_{5\%}$	89.3	89.6	89.8	89.9	88.5	89.3	89.6	89.8	88.1	88.7	89.4	89.6
	$G_{15\%}$	89.8	89.9	89.9	89.9	89.5	89.8	89.9	89.9	89.2	89.7	89.8	89.9
	$G_{30\%}$	89.8	89.9	90.0	90.0	89.7	89.9	89.9	89.9	89.6	89.8	89.9	89.9
	$G_{50\%}$	89.9	89.9	90.0	90.0	89.8	89.9	89.9	90.0	89.6	89.8	89.9	89.9
$\alpha = 0.05$	$G_{(1)}$	93.9	93.9	93.9	93.9	93.3	93.3	93.3	93.3	92.7	92.7	92.8	92.8
	$G_{5\%}$	94.6	94.8	94.9	94.9	94.2	94.6	94.8	94.9	93.9	94.3	94.7	94.8
	$G_{15\%}$	94.9	94.9	95.0	95.0	94.7	94.9	94.9	95.0	94.6	94.8	94.9	94.9
	$G_{30\%}$	94.9	95.0	95.0	95.0	94.8	94.9	95.0	95.0	94.8	94.9	94.9	95.0
	$G_{50\%}$	94.9	95.0	95.0	95.0	94.9	94.9	95.0	95.0	94.8	94.9	94.9	95.0
$\alpha = 0.01$	$G_{(1)}$	98.8	98.8	98.8	98.8	98.6	98.6	98.6	98.6	98.5	98.5	98.5	98.5
	$G_{5\%}$	98.9	99.0	99.0	99.0	98.8	98.9	99.0	99.0	98.8	98.9	98.9	99.0
	$G_{15\%}$	99.0	99.0	99.0	99.0	98.9	99.0	99.0	99.0	98.9	99.0	99.0	99.0
	$G_{30\%}$	99.0	99.0	99.0	99.0	99.0	99.0	99.0	99.0	98.9	99.0	99.0	99.0
	$G_{50\%}$	99.0	99.0	99.0	99.0	99.0	99.0	99.0	99.0	99.0	99.0	99.0	99.0

Taula 3.3: Percentatge de l'amplitud relativa de l'interval (3.8) sobre $[a_0, a_1]$ (dades uniformes)

		$k=3$				$k=4$				$k=5$			
		n				n				n			
		200	400	700	1000	200	400	700	1000	200	400	700	1000
$\alpha = 0.1$	$G_{(1)}$	88.5	88.5	88.5	88.5	88.1	88.1	88.1	88.1	87.7	87.7	87.7	87.7
	$G_{5\%}$	89.3	89.6	89.8	89.9	88.8	89.3	89.6	89.8	88.5	88.9	89.4	89.6
$\alpha = 0.05$	$G_{(1)}$	94.2	94.2	94.2	94.2	94.0	94.0	94.0	94.0	93.7	93.7	93.7	93.7
	$G_{5\%}$	94.6	94.8	94.9	94.9	94.3	94.6	94.8	94.9	94.2	94.4	94.7	94.8
$\alpha = 0.01$	$G_{(1)}$	98.8	98.8	98.8	98.8	98.8	98.8	98.8	98.8	98.7	98.7	98.7	98.7
	$G_{5\%}$	98.9	99.0	99.0	99.0	98.9	98.9	99.0	99.0	98.8	98.9	98.9	99.0

Taula 3.4: Amplitud de $[a_{-1}, a_1]$ per a diferents grups i valors n, k (dades normals)

		$k=3$				$k=4$				$k=5$			
		n				n				n			
		200	400	700	1000	200	400	700	1000	200	400	700	1000
	$G_{(1)}$	.67	.60	.56	.53	.71	.60	.58	.58	.76	.70	.61	.59
	$G_{5\%}$	.26	.26	.22	.22	.41	.34	.33	.31	.43	.38	.34	.32
	$G_{15\%}$	.12	.11	.09	.09	.18	.15	.13	.12	.23	.18	.18	.16
	$G_{30\%}$	.08	.06	.05	.05	.12	.08	.07	.07	.15	.11	.09	.08
	$G_{50\%}$	.07	.05	.04	.03	.10	.06	.05	.05	.12	.08	.06	.06

Taula 3.5: Mitjana del percentatge de l'amplitud relativa de l'interval (3.4) sobre $[a_{-1}, a_1]$ (dades normals)

		$k=3$				$k=4$				$k=5$			
		n				n				n			
		200	400	700	1000	200	400	700	1000	200	400	700	1000
$\alpha = 0.1$	$G_{(1)}$	67.8	66.1	68.2	65.3	62.3	61.8	61.9	60.2	56.4	56.8	58.0	56.7
	$G_{5\%}$	81.1	83.9	85.7	87.6	75.4	80.3	83.9	87.2	71.0	77.7	82.2	84.0
	$G_{15\%}$	84.4	86.9	88.5	89.3	83.1	87.2	87.5	88.8	78.2	84.5	87.3	88.0
	$G_{30\%}$	86.6	88.5	89.2	89.5	85.0	87.9	88.6	88.9	83.3	86.5	88.8	88.8
	$G_{50\%}$	88.2	89.0	89.4	89.5	85.8	88.4	88.8	89.1	84.1	86.9	88.5	88.5
$\alpha = 0.05$	$G_{(1)}$	77.2	75.0	77.1	74.2	71.4	70.7	70.8	69.0	65.1	65.6	66.7	65.4
	$G_{5\%}$	88.8	91.0	92.3	93.6	84.3	88.3	90.8	93.4	80.2	86.4	89.8	91.2
	$G_{15\%}$	91.4	92.9	94.2	94.6	90.1	93.4	93.5	94.3	86.4	91.3	93.4	93.9
	$G_{30\%}$	92.9	94.2	94.5	94.7	91.8	93.8	94.2	94.4	90.8	92.9	94.3	94.3
	$G_{50\%}$	94.0	94.4	94.7	94.7	92.3	94.1	94.3	94.5	91.3	93.2	94.2	94.1
$\alpha = 0.01$	$G_{(1)}$	90.3	87.7	89.3	86.8	85.0	84.0	84.1	82.4	78.9	79.6	80.4	79.2
	$G_{5\%}$	96.7	97.8	98.3	98.7	94.8	96.9	97.4	98.6	92.4	96.2	97.5	98.0
	$G_{15\%}$	98.0	98.3	98.8	98.9	96.9	98.6	98.7	98.9	95.6	97.8	98.6	98.7
	$G_{30\%}$	98.5	98.8	98.9	98.9	98.1	98.7	98.8	98.9	97.9	98.5	98.9	98.9
	$G_{50\%}$	98.8	98.9	98.9	98.9	98.3	98.8	98.9	98.9	98.0	98.6	98.8	98.8

Taula 3.6: Percentatge de l'amplitud relativa de l'interval (3.7) sobre $[a_{-1}, a_0]$ (dades normals)

		$k=3$				$k=4$				$k=5$			
		n				n				n			
		200	400	700	1000	200	400	700	1000	200	400	700	1000
$\alpha = 0.1$	$G_{(1)}$	78.5	75.1	78.3	75.7	74.0	73.8	73.7	72.4	69.3	70.1	70.9	70.3
	$G_{5\%}$	86.5	88.7	88.0	89.2	85.2	87.7	87.4	89.3	83.0	86.7	88.0	88.3
	$G_{15\%}$	87.5	89.2	89.6	89.7	87.4	89.2	89.2	89.6	86.4	88.1	89.2	89.4
	$G_{30\%}$	89.0	89.7	89.7	89.9	88.6	89.5	89.5	89.6	88.0	89.0	89.5	89.6
	$G_{50\%}$	89.4	89.7	89.7	89.8	88.3	89.6	89.7	89.7	88.5	89.0	89.6	89.6
$\alpha = 0.05$	$G_{(1)}$	87.0	83.6	86.1	83.8	82.6	82.4	82.0	81.4	78.3	79.0	79.7	79.2
	$G_{5\%}$	92.8	94.3	93.8	94.5	92.1	93.7	93.3	94.6	90.5	93.1	93.8	94.0
	$G_{15\%}$	93.5	94.5	94.8	94.9	93.3	94.5	94.5	94.8	92.8	93.8	94.6	94.7
	$G_{30\%}$	94.4	94.8	94.9	94.9	94.2	94.7	94.7	94.8	93.9	94.4	94.8	94.8
	$G_{50\%}$	94.7	94.8	94.9	94.9	94.0	94.8	94.8	94.8	94.1	94.5	94.8	94.8
$\alpha = 0.01$	$G_{(1)}$	96.2	94.2	95.0	93.1	93.5	93.2	92.1	93.1	90.9	90.9	91.3	90.9
	$G_{5\%}$	98.5	98.8	98.7	98.9	98.3	98.7	98.6	98.9	97.8	98.5	98.7	98.8
	$G_{15\%}$	98.6	98.9	98.9	99.0	98.5	98.9	98.9	99.0	98.5	98.7	98.9	98.9
	$G_{30\%}$	98.9	99.0	99.0	99.0	98.8	98.9	98.9	99.0	98.7	98.9	98.9	99.0
	$G_{50\%}$	98.9	99.0	99.0	99.0	98.7	98.9	99.0	99.0	98.8	98.9	99.0	98.9

Taula 3.7: Percentatge de l'amplitud relativa de l'interval (3.8) sobre $[a_0, a_1]$ (dades normals)

		$k=3$				$k=4$				$k=5$			
		n				n				n			
		200	400	700	1000	200	400	700	1000	200	400	700	1000
$\alpha = 0.1$	$G_{(1)}$	81.1	80.6	81.3	79.9	79.0	78.2	77.7	78.2	74.6	74.6	74.9	74.8
	$G_{5\%}$	86.3	88.9	88.3	89.3	85.9	87.7	87.7	89.4	84.4	86.8	88.0	88.4
$\alpha = 0.05$	$G_{(1)}$	89.0	88.6	88.8	87.5	87.2	86.6	85.7	86.8	83.6	83.3	83.6	83.5
	$G_{5\%}$	92.7	94.4	94.0	94.6	92.5	93.6	93.5	94.7	91.6	93.2	93.8	94.1
$\alpha = 0.01$	$G_{(1)}$	97.2	97.1	96.8	95.7	96.3	96.0	94.6	96.3	94.7	93.9	94.1	94.0
	$G_{5\%}$	98.4	98.9	98.8	98.9	98.4	98.7	98.6	98.9	98.1	98.6	98.7	98.8

Taula 3.8: Amplitud de $[a_{-1}, a_1]$ per a diferents grups i valors n, k (dades Weibull)

	$k=3$				$k=4$				$k=5$			
	n				n				n			
	200	400	700	1000	200	400	700	1000	200	400	700	1000
$G_{(1)}$	.07	.05	.03	.02	.10	.06	.04	.03	.11	.07	.05	.03
$G_{10\%}$	.05	.02	.01	.01	.06	.03	.02	.01	.08	.04	.02	.02
$G_{50\%}$	.04	.02	.01	.01	.06	.03	.02	.01	.08	.04	.02	.01
$G_{90\%}$	.13	.07	.04	.03	.19	.09	.06	.04	.23	.12	.08	.05
$G_{100\%}$	.68	.66	.69	.59	.74	.68	.67	.66	.77	.75	.69	.66

Taula 3.9: Mitjana del percentatge de l'amplitud relativa de l'interval (3.4) sobre $[a_{-1}, a_1]$ (dades Weibull)

		$k=3$				$k=4$				$k=5$			
		n				n				n			
		200	400	700	1000	200	400	700	1000	200	400	700	1000
$\alpha = 0.1$	$G_{(1)}$	73.2	70.8	73.5	75.8	68.5	69.1	69.2	70.1	65.8	65.8	64.4	65.0
	$G_{10\%}$	83.7	87.3	88.1	88.2	80.5	85.3	87.9	87.5	77.2	82.8	85.2	86.6
	$G_{50\%}$	87.3	88.5	88.9	89.6	85.0	87.8	89.2	89.2	84.8	86.2	88.4	89.0
	$G_{90\%}$	83.4	86.9	87.9	88.6	78.3	85.6	87.8	87.8	75.8	83.2	84.3	87.3
	$G_{100\%}$	65.0	63.0	61.6	67.1	60.4	61.7	59.8	61.0	55.7	56.0	54.9	55.8
$\alpha = 0.05$	$G_{(1)}$	82.2	79.8	82.4	84.3	74.8	78.2	78.1	79.1	75.1	75.0	73.7	74.3
	$G_{10\%}$	90.7	93.4	93.9	94.0	88.4	92.0	93.8	93.5	85.7	90.4	91.8	92.9
	$G_{50\%}$	93.4	94.2	94.4	94.8	91.9	93.7	94.5	94.6	91.7	92.6	94.1	94.5
	$G_{90\%}$	90.5	93.1	93.8	94.2	86.6	92.2	93.7	93.7	84.4	90.5	91.3	93.4
	$G_{100\%}$	74.3	71.8	70.6	76.3	69.8	71.0	68.6	70.2	64.5	64.9	63.8	64.3
$\alpha = 0.01$	$G_{(1)}$	93.6	91.6	93.4	94.5	87.8	90.6	90.2	91.2	88.4	88.2	87.3	87.7
	$G_{10\%}$	97.7	98.6	98.7	98.8	96.8	98.2	98.7	98.6	95.6	97.7	98.0	98.5
	$G_{50\%}$	98.6	98.8	98.9	98.9	98.2	98.7	98.9	98.9	98.1	98.4	98.8	98.9
	$G_{90\%}$	97.5	98.5	98.7	98.8	96.0	98.3	98.7	98.7	94.5	97.7	97.9	98.6
	$G_{100\%}$	87.7	84.5	84.0	89.3	84.2	85.2	82.1	84.2	78.7	79.2	78.6	78.0

Taula 3.10: Percentatge de l'amplitud relativa de l'interval (3.7) sobre $[a_{-1}, a_0]$ (dades Weibull)

		$k=3$				$k=4$				$k=5$			
		n				n				n			
		200	400	700	1000	200	400	700	1000	200	400	700	1000
$\alpha=0.1$	$G_{(1)}$	82.5	80.3	82.9	83.1	75.8	80.3	80.1	81.0	79.4	78.4	78.1	79.2
	$G_{10\%}$	86.8	89.2	89.5	89.2	86.5	88.7	89.4	89.2	86.0	88.1	88.4	88.9
	$G_{50\%}$	89.3	89.7	89.6	89.8	88.7	89.3	89.7	89.8	88.6	88.9	89.5	89.7
	$G_{90\%}$	88.2	89.1	89.4	89.6	87.3	89.0	89.3	89.3	85.2	88.0	88.5	89.4
	$G_{100\%}$	80.3	77.4	79.4	81.9	78.4	79.8	75.6	78.4	73.7	74.4	76.1	73.6
$\alpha=0.05$	$G_{(1)}$	90.0	88.1	90.2	90.1	84.3	88.2	87.9	88.8	87.7	86.6	86.6	87.5
	$G_{10\%}$	93.0	94.6	94.7	94.7	92.9	94.3	94.7	94.6	92.6	93.9	94.0	94.4
	$G_{50\%}$	94.6	94.8	94.8	94.9	94.3	94.6	94.9	94.9	94.2	94.4	94.7	94.8
	$G_{90\%}$	94.0	94.5	94.7	94.8	93.4	94.4	94.6	94.6	91.8	93.8	94.1	94.6
	$G_{100\%}$	88.2	85.4	87.3	89.5	86.9	88.0	84.0	86.9	82.2	83.2	85.2	82.5
$\alpha=0.01$	$G_{(1)}$	97.5	96.6	97.6	97.1	94.4	96.7	96.5	97.1	96.7	95.7	96.3	96.5
	$G_{10\%}$	98.5	98.9	98.9	98.9	98.4	98.8	98.9	98.9	98.4	98.7	98.8	98.9
	$G_{50\%}$	98.9	99.0	99.0	99.0	98.8	98.9	99.0	99.0	98.8	98.9	98.9	99.0
	$G_{90\%}$	98.8	98.9	98.9	99.0	98.6	98.9	98.9	98.9	98.0	98.7	98.8	98.9
	$G_{100\%}$	96.7	94.7	96.0	97.3	96.4	96.9	94.1	96.4	92.6	93.9	95.7	93.9

Taula 3.11: Percentatge de l'amplitud relativa de l'interval (3.8) sobre $[a_0, a_1]$ (dades Weibull)

		$k=3$				$k=4$				$k=5$			
		n				n				n			
		200	400	700	1000	200	400	700	1000	200	400	700	1000
$\alpha=0.1$	$G_{(1)}$	82.9	80.5	83.3	82.5	75.8	80.3	80.1	81.0	79.6	78.1	78.9	79.8
	$G_{10\%}$	88.6	89.2	89.5	89.5	87.1	88.6	89.5	89.3	86.4	88.0	88.7	89.2
	$G_{50\%}$	89.1	89.3	89.7	89.9	88.3	89.5	89.8	89.8	88.3	88.9	89.6	89.7
	$G_{90\%}$	87.4	88.8	89.3	89.6	85.2	88.5	89.5	89.3	85.0	87.8	88.3	89.2
	$G_{100\%}$	75.5	72.2	72.9	78.0	78.4	79.8	75.6	78.5	69.4	70.2	70.3	68.8
$\alpha=0.05$	$G_{(1)}$	90.3	88.1	90.6	89.6	84.3	88.2	87.9	88.8	88.0	86.5	87.3	88.0
	$G_{10\%}$	94.2	94.6	94.7	94.7	93.2	94.2	94.7	94.6	92.8	93.9	94.2	94.6
	$G_{50\%}$	94.5	94.6	94.8	94.9	94.0	94.8	94.9	94.9	94.0	94.4	94.8	94.9
	$G_{90\%}$	93.4	94.3	94.6	94.8	91.9	94.1	94.7	94.6	91.8	93.7	94.0	94.6
	$G_{100\%}$	84.0	80.5	81.7	86.4	86.9	88.0	84.0	86.9	78.2	79.1	79.8	77.7
$\alpha=0.01$	$G_{(1)}$	97.7	96.4	97.7	96.6	94.4	96.7	96.5	97.1	96.9	95.7	96.5	96.9
	$G_{10\%}$	98.8	98.9	98.9	98.9	98.5	98.8	99.0	98.9	98.4	98.7	98.8	98.9
	$G_{50\%}$	98.9	98.9	99.0	99.0	98.8	98.9	99.0	99.0	98.8	98.9	99.0	99.0
	$G_{90\%}$	98.6	98.9	98.9	99.0	98.2	98.8	98.9	98.9	98.1	98.7	98.8	98.9
	$G_{100\%}$	94.2	91.2	92.6	95.7	96.4	96.9	94.1	96.4	89.8	91.0	92.7	90.0

Capítol 4

Mètodes DM i DMM per a microagregació multivariant

4.1 Microagregació Univariant i Multivariant

En **microagregació univariant**, el procediment a seguir per obtenir microagregats en un conjunt amb n microdades, és, primerament, ordenar les dades; a continuació, formar una partició de les n microdades en grups de *valors consecutius*, cadascun dels quals ha de tenir mida almenys k, puix que s'ha demostrat que, en microagregació univariant, la solució òptima que minimitza la pèrdua d'informació està formada per grups connexos d'elements consecutius. Finalment, es calcula la mitjana aritmètica sobre cada grup, la qual es fa servir per substituir cadascun dels valors que han intervingut en el seu càlcul. Un cop s'ha completat aquest procediment, les dades resultants (modificades) poden ser publicades.

En **microagregació multivariant**, el procediment a seguir serà una generalització del que hem explicat per al cas univariant. Això suposa, primerament, crear, per al conjunt dels n vectors de microdades, una partició amb grups de mida almenys k; després, es calcula el vector mitjana sobre cada grup, el qual es fa servir per substituir cadascun dels vectors que han intervingut en el seu càlcul; quedant, d'aquesta manera, modificades les dades vectorials originals per poder ser publicades.

Cal observar que, quan es treballa amb un conjunt de dades multivariant, els mètodes de microagregació mitjançant projecció de les dades sobre un eix, tenen l'inconvenient que la pèrdua d'informació deguda a la projecció s'afegeix a la pèrdua d'informació pròpia de la microagregació. D'altra banda, els mètodes de microagregació multivariant sense projecció de les dades, si bé disminueixen molt considerablement la pèrdua d'informació, tenen una gran dificultat a l'hora d'elegir el criteri per fer els agrupaments, ja que, fins ara, cap resultat ha estat provat que, de manera semblant a la propietat de connexió dels grups en el cas univariant, condueixi eficientment a la solució òptima que minimitza la pèrdua d'informació en el cas multivariant.

En aquest context, tot seguit descrivim dos mètodes heurístics: el mètode ja existent de la Distància Màxima (DM) (Domingo-Ferrer i Mateo-Sanz 2002) i el nou mètode de la Distància Màxima Modificat (DMM), com a criteris d'agrupament per a microagregació multivariant sense projecció de les dades.

4.2 Mètode DM de la Distància Màxima

Considerem un conjunt de microdades amb p variables mètriques i n vectors de dades (és a dir, el resultat d'observar p variables en n individus):

$$\mathbf{X_i} = (x_{i1}, x_{i2}, \ldots, x_{ip}) \qquad \text{per a} \qquad 1 \leq i \leq n$$

Sigui s un nombre natural fixat tal que $s \leq p$. Dividint p entre s, tindrem:

$$p = hs + r \tag{4.1}$$

essent $0 \leq r < s$.

Sigui $\mathbf{V} = \{V_1, V_2, \ldots, V_p\}$ el conjunt constituït per les p variables observades, i $\mathbf{G}$ una partició de $\mathbf{V}$ formada per $h-1$ conjunts amb s elements cadascun i un únic conjunt amb $s+r$ elements (en el cas que $r = 0$ la partició tindria h grups, tots de mida s).

Sigui $\mathbf{G} = \{G_1, G_2, G_3, \ldots, G_h\}$, on

$$G_l = \{V_{l_1}, V_{l_2}, \ldots, V_{l_s}\} \qquad \text{per a} \qquad 1 \leq l \leq h-1$$

$$G_h = \{V_{h_1}, V_{h_2}, \ldots, V_{h_{s+r}}\}$$

Per a cada l, $1 \leq l \leq h-1$, i per a h, siguin

$$\mathbf{Y_{il}} = (x_{il_1}, x_{il_2}, \ldots, x_{il_s}) \qquad \text{per a} \qquad 1 \leq i \leq n \qquad 1 \leq l \leq h-1$$

$$\mathbf{Y_{ih}} = (x_{ih_1}, x_{ih_2}, \ldots, x_{ih_{s+r}}) \qquad \text{per a} \qquad 1 \leq i \leq n$$

els vectors projecció de les microdades originals $\mathbf{X_i} = (x_{i1}, x_{i2}, \ldots, x_{ip})$, $1 \leq i \leq n$, sobre les coordenades corresponents als conjunts de variables

$$G_l = \{V_{l_1}, V_{l_2}, \ldots, V_{l_s}\} \qquad \text{per a} \qquad 1 \leq l \leq h-1$$

$$G_h = \{V_{h_1}, V_{h_2}, \ldots, V_{h_{s+r}}\}$$

respectivament.

En el mètode DM de microagregació multivariant, seguim els següents passos:

Començant per $l = 1$

Pas 1 Calculem les $\binom{n}{2}$ diferents distàncies euclidianes

$$\mathrm{d}(\mathbf{Y_{il}}, \mathbf{Y_{jl}}) = \sqrt{(x_{il_1} - x_{jl_1})^2 + (x_{il_2} - x_{jl_2})^2 + \cdots + (x_{il_s} - x_{jl_s})^2}$$

$$1 \leq i < j \leq n \tag{4.2}$$

entre cada dos dels punts $\mathbf{Y_{il}} = (x_{il_1}, x_{il_2}, \ldots, x_{il_s})$, $1 \leq i \leq n$.

Pas 2 Elegim els dos punts $\mathbf{Y_{al}}, \mathbf{Y_{bl}}$ més allunyats mútuament; és a dir, aquells dos punts la distància entre els quals és la més gran de totes les calculades a (4.2).

Pas 3 Formem dos microagregats projecció agrupant, primerament, $\mathbf{Y_{al}}$ amb els $k-1$ punts més propers a ell d'acord amb les distàncies calculades a (4.2); de manera semblant, a continuació, agrupem $\mathbf{Y_{bl}}$ amb els $k-1$ punts més propers a ell d'acord també amb les distàncies calculades a (4.2).

Pas 4 Si el número de vectors encara no agrupats és més gran o igual que $3k$, llavors tornem al pas **1**, agafant com a nou conjunt de dades el conjunt de punts projecció que queden sense agrupar.

Pas 5 Si el número de vectors encara no agrupats és més gran o igual que $2k$ i més petit que $3k$, llavors:

- calculem la distància euclidiana entre cada parella de punts encara no microagregats;
- elegim els dos punts $\mathbf{Y_{al}}, \mathbf{Y_{bl}}$ més allunyats mútuament;
- formem el grup que conté $\mathbf{Y_{al}}$ i els $k-1$ punts més propers a ell;
- formem un altre grup amb la resta de punts que queden per agrupar;
- anem al pas **7**.

Pas 6 Si el número de vectors encara no agrupats és més petit que $2k$, tots aquests vectors encara sense agrupar formaran un grup. Tot seguit anem al pas **7**.

Pas 7 Per a cada microagregat projecció M_{jl}, $1 \leq j \leq E[n/k]$, es calcula, per a cada variable de G_l, la seva mitjana aritmètica sobre els seus k valors del microagregat; la qual es fa servir per substituir, en les dades originals individuals $\mathbf{X_i}$ corresponents a les dades del microagregat projecció M_{jl}, cadascun dels valors individuals que ha intervingut en el seu càlcul.

Pas 8 Per a cadascun dels valors de l: $l = 2, l = 3, \ldots, l = h$, repetim tot l'anterior procés començant cada vegada pel pas **1**.

Un cop s'ha completat aquest procediment, els vectors de dades resultants (modificats) poden ser publicats.

4.3 Mètode DMM de la Distància Màxima Modificat

Considerem un conjunt de microdades amb p variables mètriques i n vectors de dades (és a dir, el resultat d'observar p variables en n individus):

$$\mathbf{X_i} = (x_{i1}, x_{i2}, \ldots, x_{ip}) \qquad \text{per a} \qquad 1 \leq i \leq n$$

Sigui s un nombre natural fixat tal que $s \leq p$. Dividint p entre s, tindrem:

$$p = hs + r \tag{4.3}$$

essent $0 \leq r < s$.

Sigui $\mathbf{V} = \{V_1, V_2, \ldots, V_p\}$ el conjunt constituït per les p variables observades, i $\mathbf{G}$ una partició de $\mathbf{V}$ formada per $h-1$ conjunts amb s elements cadascun i un únic conjunt amb $s+r$ elements (en el cas que $r = 0$ la partició tindria h grups, tots de mida s).

Sigui $\mathbf{G} = \{G_1, G_2, G_3, \ldots, G_h\}$, on

$$G_l = \{V_{l_1}, V_{l_2}, \ldots, V_{l_s}\} \qquad \text{per a} \qquad 1 \leq l \leq h-1$$

$$G_h = \{V_{h_1}, V_{h_2}, \ldots, V_{h_{s+r}}\}$$

Per a cada l, $1 \leq l \leq h-1$, i per a h, siguin

$$\mathbf{Y_{il}} = (x_{il_1}, x_{il_2}, \ldots, x_{il_s}) \qquad \text{per a} \qquad 1 \leq i \leq n \qquad 1 \leq l \leq h-1$$

$$\mathbf{Y_{ih}} = (x_{ih_1}, x_{ih_2}, \ldots, x_{ih_{s+r}}) \qquad \text{per a} \qquad 1 \leq i \leq n$$

els vectors projecció de les microdades originals $\mathbf{X_i} = (x_{i1}, x_{i2}, \ldots, x_{ip})$, $1 \leq i \leq n$, sobre les coordenades corresponents als conjunts de variables

$$G_l = \{V_{l_1}, V_{l_2}, \ldots, V_{l_s}\} \qquad \text{per a} \qquad 1 \leq l \leq h-1$$

$$G_h = \{V_{h_1}, V_{h_2}, \ldots, V_{h_{s+r}}\}$$

respectivament.

En el mètode DMM de microagregació multivariant, seguim els següents passos:

Començant per $l = 1$

Pas 1 Calculem el vector mitjana, $\overline{\mathbf{Y}} = (\overline{y}_1, \overline{y}_2, \ldots, \overline{y}_s)$, dels vectors projecció $\mathbf{Y_{il}}$, $1 \leq i \leq n$.

Tot seguit calculem les n diferents distàncies euclidianes

$$\mathrm{d}(\mathbf{Y_{il}}, \overline{\mathbf{Y}}_\mathbf{n}) = \sqrt{(x_{il_1} - \overline{y}_1)^2 + (x_{il_2} - \overline{y}_2)^2 + \cdots + (x_{il_s} - \overline{y}_s)^2}$$

$$1 \leq i \leq n \tag{4.4}$$

entre cada un dels punts $\mathbf{Y_{il}} = (x_{il_1}, x_{il_2}, \ldots, x_{il_s})$, $1 \leq i \leq n$, i el vector mitjana $\overline{\mathbf{Y}}$.

Pas 2 Elegim el punt $\mathbf{Y_{cl}}$ més allunyat del vector mitjana $\overline{\mathbf{Y}}$; és a dir, el punt que correspon a la distància més gran de totes les calculades a (4.4).

Tot seguit, calculem la distància euclidiana del punt $\mathbf{Y_{cl}}$ a cadascun dels punts projecció encara no microagregats en anteriors agrupaments. Mitjançant l'observació d'aquestes distàncies, agruparem $\mathbf{Y_{cl}}$ amb els $k-1$ punts més propers a ell.

Pas 3 Elegim el punt $\mathbf{Y_{dl}}$ més allunyat del punt $\mathbf{Y_{cl}}$ calculat al pas anterior.

De manera semblant al pas anterior, a continuació, calculem la distància euclidiana del punt $\mathbf{Y_{dl}}$ a cadascun dels punts projecció encara no microagregats en anteriors agrupaments. Mitjançant l'observació d'aquestes noves distàncies, agruparem $\mathbf{Y_{dl}}$ amb els $k-1$ punts més propers a ell.

Pas 4 Si el número de vectors encara no agrupats és més gran o igual que $3k$, llavors tornem al pas **1**, agafant com a nou conjunt de dades el conjunt de punts projecció que queden sense agrupar.

Pas 5 Si el número de vectors encara no agrupats és més gran o igual que $2k$ i més petit que $3k$, llavors:

- calculem el vector mitjana $\overline{\mathbf{Y}}$ de tots els vectors encara no microagregats;
- elegim el punt $\mathbf{Y_{cl}}$ més allunyat de $\overline{\mathbf{Y}}$;
- formem el grup que conté $\mathbf{Y_{cl}}$ i els $k-1$ punts més propers a ell;
- formem un altre grup amb la resta de punts que queden per agrupar;
- anem al pas **7**.

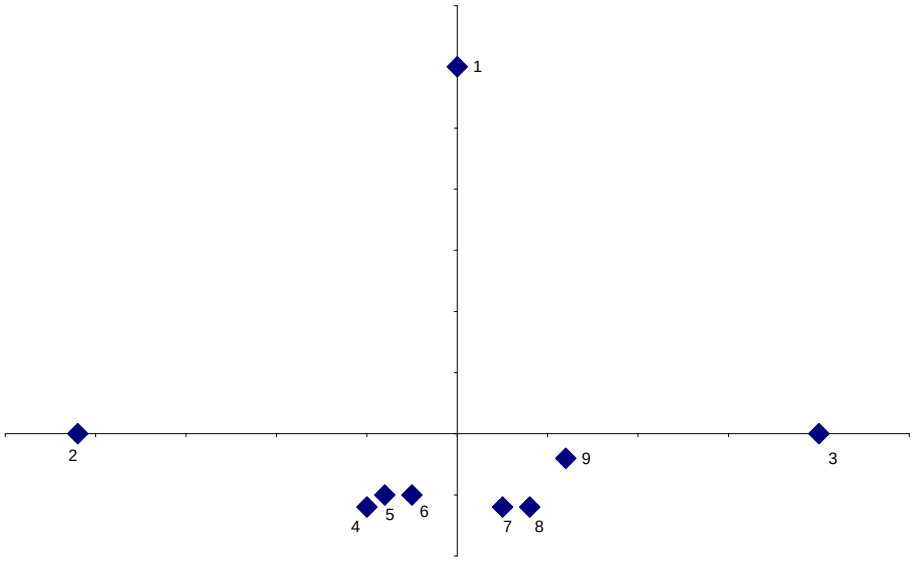


Figura 4.1: Representació dels 9 vectors de dades de l'exemple.

Pas 6 Si el número de vectors encara no agrupats és més petit que $2k$, tots aquests vectors encara sense agrupar formaran un grup. Tot seguit anem al pas **7**.

Pas 7 Per a cada microagregat projecció M_{jl}, $1 \leq j \leq E[n/k]$, es calcula, per a cada variable de G_l, la seva mitjana aritmètica sobre els seus k valors del microagregat; la qual es fa servir per substituir, en les dades originals individuals $\mathbf{X_i}$ corresponents a les dades del microagregat projecció M_{jl}, cadascun dels valors individuals que ha intervingut en el seu càlcul.

Pas 8 Per a cadascun dels valors de l: $l = 2, l = 3, \ldots, l = h$, repetim tot l'anterior procés començant cada vegada pel pas **1**.

Un cop s'ha completat aquest procediment, els vectors de dades resultants (modificats) poden ser publicats.

4.4 Exemple d'aplicació dels mètodes DM i DMM

Mitjançant l'aplicació dels dos mètodes de microagregació multivariant **DM** i **DMM** a un mateix conjunt de microdades amb $p = 2$ variables mètriques observades i $n = 9$ vectors de dades, comprovarem que els dos conjunts de microagregats formats per cadascun d'aquests dos mètodes, són diferents; per la qual cosa es tracta de dos mètodes de microagregació multivariant realment diferents.

Siguin els següents 9 vectors de dades amb 2 variables observades:

$$X_1 = (0, 3) \quad ; \quad X_2 = (-2.1, 0) \quad ; \quad X_3 = (2, 0)$$

$$X_4 = (-0.5, -0.6) \quad ; \quad X_5 = (-0.4, -0.5) \quad ; \quad X_6 = (-0.25, -0.5)$$

$$X_7 = (0.25, -0.6) \quad ; \quad X_8 = (0.4, -0.6) \quad ; \quad X_9 = (0.6, -0.2)$$

En l'aplicació dels dos mètodes considerarem els següents valors dels paràmetres: $s = p = 2$ i $k = 3$.

Aplicació del mètode DM

Després de calcular les $\begin{pmatrix} 9 \\ 2 \end{pmatrix} = 36$ diferents distàncies euclidianes entre els anteriors 9 vectors de dades agafats de dos en dos, es pot observar que els dos vectors mútuament més allunyats són: X_2 i X_3.

Considerant novament les 36 distàncies euclidianes esmentades, també es pot veure que els dos punts més propers al punt X_2 són: X_4 i X_5. De la mateixa manera podem comprovar que els dos punts més propers al punt X_3 són: X_8 i X_9.

Així doncs, mitjançant l'aplicació del mètode DM obtendríem finalment els tres següents microagregats:

$$\{X_2, X_4, X_5\} \quad ; \quad \{X_3, X_8, X_9\} \quad \text{i} \quad \{X_1, X_6, X_7\}$$

Aplicació del mètode DMM

Calculant primerament el vector mitjana, $\overline{\mathbf{Y}}$, dels anteriors 9 vectors de dades, tenim:

$$\overline{\mathbf{Y}} = \frac{1}{9} \sum_{i=1}^{9} X_i = (0, 0)$$

Calculant posteriorment les 9 distàncies euclidianes entre el vector mitjana i cadascun dels 9 vectors de dades, es pot observar que el vector més allunyat del centre $\overline{\mathbf{Y}} = (0, 0)$, és: X_1.

Tot seguit, calculant les 8 distàncies euclidianes entre X_1 i cadascun dels altres 8 vectors de dades, es pot comprovar que el punt més allunyat de X_1 és justament X_2.

Considerant novament les anteriors 8 distàncies, es pot observar que els dos punts més propers al punt X_1 són: X_6 i X_9.

Mitjançant càlcul de les distàncies entre el punt X_2 i cadascun dels restants 5 punts, es pot comprovar que els dos punts més propers al punt X_2 són: X_4 i X_5.

Així doncs, mitjançant l'aplicació del mètode DMM obtendríem finalment els tres següents microagregats:

$$\{X_1, X_6, X_9\} \quad ; \quad \{X_2, X_4, X_5\} \quad \text{i} \quad \{X_3, X_7, X_8\}$$

4.5 Implementació i complexitat computacional dels mètodes DM i DMM

4.5.1 Implementació dels mètodes DM i DMM

Existeixen dues maneres d'implementar els algorismes dels mètodes de microagregació de la Distància Màxima (DM) i de la Distància Màxima Modificat (DMM):

Emmagatzemant la matriu de distàncies

Donada una partició $\mathbf{G}$ del conjunt $\mathbf{V}$ de variables, es tracta de calcular, per a cada conjunt $G_l \in \mathbf{G}$, $1 \leq l \leq h$, la matriu que conté les distàncies euclidianes entre cada vector projecció

$\mathbf{Y_{il}} = (x_{il_1}, x_{il_2}, \ldots, x_{il_s})$, $1 \leq i \leq n$, i tots els altres. Aquesta matriu, per a cada G_l, és simètrica i té zeros a la diagonal principal; per la qual cosa, l'emmagatzematge que es necessita, quan treballem amb n vectors de dades, és una funció quadràtica d'n:

$$E_m = \left(\begin{array}{c} n \\ 2 \end{array} \right) = \frac{n\,(n-1)}{2}$$

L'emmagatzematge d'una matriu en la implementació d'un algorisme és una opció pràctica només quan es treballa amb conjunts que tenen un número de dades no excessivament gran.

Sense emmagatzemar la matriu de distàncies

Quan es treballa amb conjunts que tenen un número n de dades inicials molt gran, la millor alternativa, per evitar problemes de memòria del computador, és emmagatzemar només les n dades inicials; la qual cosa requereix senzillament un emmagatzematge $E_m = n$.

Tot i això, en el cas dels algorismes per als mètodes de microagregació DM i DMM, aquest estalvi de memòria, obviament incrementa la complexitat computacional, ja que totes les distàncies implicades en el desenvolupament dels algorismes es calculen únicament quan es necessiten, en lloc de ser guardades a la memòria del computador. A les següents subseccions estudiem les complexitats computacionals dels algorismes DM i DMM quan s'utilitzen sense emmagatzemar la matriu de distàncies.

4.5.2 Càlcul del número de distàncies i complexitat computacional de l'algorisme DM sense emmagatzemar la matriu de distàncies

Primerament, calculem el número total, D_p, de distàncies computades durant l'execució de l'algorisme del mètode DM:

Desenvoluparem el càlcul considerant el següent número n de registres:

$$n = 2f \cdot k + t \qquad 0 \leq t < k \tag{4.5}$$

essent f un nombre natural,

puix que els resultats obtinguts per a $n = (2f+1) \cdot k + t\ ,\ 0 \leq t < k$, són molt similars.

Distingim dos tipus de distàncies calculades durant l'execució de l'algorisme DM:

1. Distàncies calculades amb l'objectiu de seleccionar els punts $\mathbf{Y_{al}}$ i $\mathbf{Y_{bl}}$ mútuament més allunyats.
2. Distàncies calculades per agrupar $\mathbf{Y_{al}}$ i $\mathbf{Y_{bl}}$ amb els respectius $k-1$ punts més propers a cadascun d'ells.

Distàncies calculades per seleccionar els dos punts més allunyats

Calculem de la següent manera el número de distàncies, D_{p1}, que l'algorisme DM ha de computar per seleccionar els dos punts més allunyats quan no està emmagatzemada la matriu de distàncies:

$$D_{p1} = \sum_{i=0}^{\frac{n-2k-t}{2k}} \left(\begin{array}{c} n-2ik \\ 2 \end{array} \right) =$$

$$= \left(\begin{array}{c} n \\ 2 \end{array} \right) + \left(\begin{array}{c} n-2k \\ 2 \end{array} \right) + \left(\begin{array}{c} n-4k \\ 2 \end{array} \right) + \cdots + \left(\begin{array}{c} 2k+t \\ 2 \end{array} \right) =$$

$$= \frac{n\,(n-1)}{2} + \frac{(n-2k)\,(n-2k-1)}{2} + \frac{(n-4k)\,(n-4k-1)}{2} + \cdots + \frac{(2k+t)\,(2k+t-1)}{2} =$$

$$= \frac{1}{2} \cdot \left[\frac{n-t}{2k}\, n^2 - \left(4k \sum_{i=1}^{\frac{n-2k-t}{2k}} i + \frac{n-t}{2k} \right) n + 2k \sum_{i=1}^{\frac{n-2k-t}{2k}} i + 4k^2 \sum_{i=1}^{\frac{n-2k-t}{2k}} i^2 \right] =$$

$$= \frac{1}{2} \cdot \left[\frac{n-t}{2k}\, n^2 - \left(4k\, \frac{1+\frac{n-2k-t}{2k}}{2}\, \frac{n-2k-t}{2k} + \frac{n-t}{2k} \right) n + 2k\, \frac{1+\frac{n-2k-t}{2k}}{2}\, \frac{n-2k-t}{2k} + \right.$$

$$\left. + 4k^2\, \frac{\frac{n-2k-t}{2k} \left(\frac{n-2k-t}{2k} + 1 \right) \left(\frac{n-2k-t}{k} + 1 \right)}{6} \right] =$$

$$= \frac{1}{12k}\, n^3 + \frac{2k-1}{8k}\, n^2 + \frac{2k^2 - 2kt - 3k + 2t}{12k}\, n - \frac{2t^3 - 3t^2 + 6kt^2 + 4k^2 t - 6tk}{24k}$$

Distàncies calculades per microagregar cadascun dels dos punts mútuament més allunyats

Calculem de la següent manera el número de distàncies, D_{p2}, que l'algorisme DM ha de computar per microagregar els punts més allunyats quan no està emmagatzemada la matriu de distàncies:

$$D_{p2} = (2k+t-2) + \sum_{i=0}^{\frac{n-4k-t}{2k}} \{(n-2ik-2) + [n-(2i+1)k-1]\} =$$

$$= [(n-2)+(n-k-1)] + [(n-2k-2)+(n-3k-1)] + [(n-4k-2)+(n-5k-1)] + \cdots + [(2k+t-2)+0] =$$

$$= [(n-2)+(n-2k-2)+(n-4k-2)+\cdots+(2k+t-2)] + [(n-k-1)+(n-3k-1)+(n-5k-1)+\cdots+0] =$$

$$= \left(\frac{n-t}{2k}\, n - k \sum_{i=1}^{\frac{n-2k-t}{2k}} 2i - \frac{n-t}{2k}\, 2 \right) + \left[\left(\frac{n-t}{2k} - 1 \right) n - k \sum_{i=1}^{\frac{n-2k-t}{2k}} (2i-1) - \left(\frac{n-t}{2k} - 1 \right) \right] =$$

$$= \left(\frac{n-t}{2k}\, n - k\, \frac{2+\frac{n-2k-t}{k}}{2}\, \frac{n-2k-t}{2k} - \frac{n-t}{k} \right) +$$

$$+ \left[\left(\frac{n-t}{2k} - 1 \right) n - k\, \frac{1+\frac{n-3k-t}{k}}{2}\, \frac{n-2k-t}{2k} - \left(\frac{n-t}{2k} - 1 \right) \right] =$$

$$= \left(\frac{1}{4k}\, n^2 + \frac{k-2}{2k}\, n - \frac{t^2 + 2kt - 4t}{4k} \right) + \left(\frac{1}{4k}\, n^2 - \frac{1}{2k}\, n - \frac{t^2 + 4k^2 + 4kt - 2t - 4k}{4k} \right) =$$

$$= \frac{1}{2k}\, n^2 + \frac{k-3}{2k}\, n - \frac{2t^2 + 4k^2 + 6kt - 6t - 4k}{4k}$$

Complexitat computacional de l'algorisme DM

D'aquesta manera, el número total de distàncies, D_p, computades durant l'execució de l'algorisme DM és:

$$D_p = D_{p1} + D_{p2} = \left(\frac{1}{12k}\, n^3 + \frac{2k-1}{8k}\, n^2 + \frac{2k^2 - 2kt - 3k + 2t}{12k}\, n - \frac{2t^3 - 3t^2 + 6kt^2 + 4k^2 t - 6tk}{24k} \right) +$$

$$+ \left(\frac{1}{2k}\, n^2 + \frac{k-3}{2k}\, n - \frac{2t^2 + 4k^2 + 6kt - 6t - 4k}{4k} \right) =$$

$$= \frac{1}{12k}\, n^3 + \frac{2k+3}{8k}\, n^2 + \frac{2k^2 - 2kt + 3k + 2t - 18}{12k}\, n -$$

$$- \frac{2t^3 + 9t^2 + 24k^2 + 6kt^2 + 4k^2t + 30tk - 36t - 24k}{24k}$$

Desglossarem la complexitat computacional del mètode DM en tres apartats, en els que el computador realitza diferents operacions, que suposen també diferents economies de temps computacional:

Número total de sumes i restes implicades en l'algorisme DM

Com que totes les distàncies de l'algorisme DM estan computades entre vectors projecció dels vectors de dades inicials sobre les variables de la partició:

$$G_l = \{V_{l_1}, V_{l_2}, \ldots, V_{l_s}\} \qquad \text{per a} \qquad 1 \leq l \leq h-1$$

$$G_h = \{V_{h_1}, V_{h_2}, \ldots, V_{h_{s+r}}\}$$

i cada distància entre dos punts projecció corresponents als conjunts $G_l = \{V_{l_1}, V_{l_2}, \ldots, V_{l_s}\}$, $1 \leq l \leq h-1$, requereix efectuar s diferències i $s-1$ sumes; a més a més de les $s+r$ diferències i $s+r-1$ sumes de la distància entre cada dos punts projecció corresponents al conjunt $G_h = \{V_{h_1}, V_{h_2}, \ldots, V_{h_{s+r}}\}$, llavors, el número total de sumes i restes S_p de l'algorisme, és:

$$S_p = [(2s-1)\cdot(h-1) + (2s+2r-1)]\cdot D_p =$$

$$= [(2s-1)\,h + 2r]\cdot\left(\frac{1}{12k}\, n^3 + \frac{2k+3}{8k}\, n^2 + \frac{2k^2 - 2kt + 3k + 2t - 18}{12k}\, n -\right.$$

$$\left. - \frac{2t^3 + 9t^2 + 24k^2 + 6kt^2 + 4k^2t + 30tk - 36t - 24k}{24k}\right)$$

Número de multiplicacions implicades en l'algorisme DM

Com que cada distància entre dos punts projecció corresponents als conjunts $G_l = \{V_{l_1}, V_{l_2}, \ldots, V_{l_s}\}$, $1 \leq l \leq h-1$, requereix efectuar s multiplicacions; a més a més de les $s+r$ multiplicacions de la distància entre cada dos punts projecció corresponents al conjunt $G_h = \{V_{h_1}, V_{h_2}, \ldots, V_{h_{s+r}}\}$, llavors, el número de multiplicacions M_p de l'algorisme, és:

$$M_p = [s\,(h-1) + (s+r)]\cdot D_p =$$

$$= (sh+r)\cdot\left(\frac{1}{12k}\, n^3 + \frac{2k+3}{8k}\, n^2 + \frac{2k^2 - 2kt + 3k + 2t - 18}{12k}\, n -\right.$$

$$\left. - \frac{2t^3 + 9t^2 + 24k^2 + 6kt^2 + 4k^2t + 30tk - 36t - 24k}{24k}\right)$$

Número de comparacions per ordenar distàncies a l'algorisme DM

Donada l'estreta relació entre el número de comparacions C_p i el número de distàncies calculades, el número de comparacions que ha de comprovar l'algorisme DM, per ordenar distàncies, és molt similar al número total de distàncies calculades.

4.5.3 Càlcul del número de distàncies i complexitat computacional de l'algorisme DMM sense emmagatzemar la matriu de distàncies

Primerament, calculem el número total, D_m, de distàncies computades durant l'execució de l'algorisme del mètode DMM:

Desenvoluparem el càlcul considerant el següent número n de registres:

$$n = 2f \cdot k + t \qquad 0 \leq t < k \tag{4.6}$$

essent f un nombre natural,

puix que els resultats obtinguts per a $n = (2f+1) \cdot k + t\ ,\ 0 \leq t < k$, són molt similars.

Perquè aquest desenvolupament resulti més entenedor, diferenciarem dos tipus de distàncies calculades durant l'execució de l'algorisme DMM:

1. Distàncies calculades amb l'objectiu de seleccionar el punt $\mathbf{Y_{cl}}$ més allunyat del punt mitjana $\overline{\mathbf{Y}}$.
2. Distàncies calculades per trobar el segon punt $\mathbf{Y_{dl}}$ (el més allunyat de l'anterior punt $\mathbf{Y_{cl}}$), i agrupar $\mathbf{Y_{cl}}$ i $\mathbf{Y_{dl}}$ amb els respectius $k-1$ punts més propers a cadascun d'ells.

Distàncies calculades per seleccionar el punt més allunyat del punt mitjana

Calculem de la següent manera el número de distàncies, D_{m1}, que l'algorisme DMM ha de computar per seleccionar el punt $\mathbf{Y_{cl}}$ més allunyat del punt mitjana $\overline{\mathbf{Y}}$:

$$D_{m1} = \sum_{i=0}^{\frac{n-2k-t}{2k}} (n-2ik) =$$

$$= n + (n-2k) + (n-4k) + \cdots + (2k+t) =$$

$$= \frac{n-t}{2k}\, n - k \sum_{i=1}^{\frac{n-2k-t}{2k}} 2i =$$

$$= \frac{n-t}{2k}\, n - k\, \frac{2 + \frac{n-2k-t}{k}}{2}\, \frac{n-2k-t}{2k} =$$

$$= \frac{1}{4k}\, n^2 + \frac{1}{2}\, n - \frac{t^2+2kt}{4k}$$

Distàncies calculades per trobar el segon punt $\mathbf{Y_{dl}}$ i formar dos microagregats

Calculem de la següent manera el número de distàncies, D_{m2}, que l'algorisme DMM ha de computar per trobar el segon punt $\mathbf{Y_{dl}}$ (el més allunyat de l'anterior punt $\mathbf{Y_{cl}}$), i agrupar $\mathbf{Y_{cl}}$ i $\mathbf{Y_{dl}}$ amb els respectius $k-1$ punts més propers a cadascun d'ells, quan no està emmagatzemada la matriu de distàncies:

$$D_{m2} = (2k+t-1) + \sum_{i=0}^{\frac{n-4k-t}{2k}} \{(n-2ik-1) + [n-(2i+1)k-1]\} =$$

$$= [(n-1)+(n-k-1)] + [(n-2k-1)+(n-3k-1)] + [(n-4k-1)+(n-5k-1)] + \cdots + [(2k+t-1)+0] =$$

$$= [(n-1)+(n-2k-1)+(n-4k-1)+\cdots+(2k+t-1)]+[(n-k-1)+(n-3k-1)+(n-5k-1)+\cdots+0] =$$

$$= \left(\frac{n-t}{2k}\,n - k\sum_{i=1}^{\frac{n-2k-t}{2k}} 2i - \frac{n-t}{2k}\right) + \left[\left(\frac{n-t}{2k}-1\right)n - k\sum_{i=1}^{\frac{n-2k-t}{2k}}(2i-1) - \left(\frac{n-t}{2k}-1\right)\right] =$$

$$= \left(\frac{n-t}{2k}\,n - k\,\frac{2+\frac{n-2k-t}{k}}{2}\,\frac{n-2k-t}{2k} - \frac{n-t}{2k}\right) +$$

$$+ \left[\left(\frac{n-t}{2k}-1\right)n - k\,\frac{1+\frac{n-3k-t}{k}}{2}\,\frac{n-2k-t}{2k} - \left(\frac{n-t}{2k}-1\right)\right] =$$

$$= \left(\frac{1}{4k}\,n^2 + \frac{k-1}{2k}\,n - \frac{t^2+2kt-2t}{4k}\right) + \left(\frac{1}{4k}\,n^2 - \frac{1}{2k}\,n - \frac{t^2+4k^2+4kt-2t-4k}{4k}\right) =$$

$$= \frac{2}{4k}\,n^2 + \frac{k-2}{2k}\,n - \frac{2t^2+4k^2+6kt-4t-4k}{4k}$$

Complexitat computacional de l'algorisme DMM

D'aquesta manera, el número total de distàncies, D_m, computades durant l'execució de l'algorisme DMM és:

$$D_m = D_{m1} + D_{m2} = \left(\frac{1}{4k}\,n^2 + \frac{1}{2}\,n - \frac{t^2+2kt}{4k}\right) +$$

$$+ \left(\frac{2}{4k}\,n^2 + \frac{k-2}{2k}\,n - \frac{2t^2+4k^2+6kt-4t-4k}{4k}\right) =$$

$$= \frac{3}{4k}\,n^2 + \frac{k-1}{k}\,n - \frac{3t^2+4k^2+8kt-4t-4k}{4k}$$

Desglossarem igualment la complexitat computacional del mètode DMM en tres apartats, en els que el computador realitza diferents operacions, que suposen també diferents economies de temps computacional:

Número total de sumes i restes implicades en l'algorisme DMM

El número de sumes necessàries per calcular cada una de les s components dels vectors mitjana corresponents als conjunts $G_l\,,\ 1 \le l \le h-1$ de la partició; i cada una de les $s+r$ components dels vectors mitjana corresponents al conjunt G_h de la partició, és:

$$(n-1) + (n-2k-1) + (n-4k-1) + \cdots + (2k+t-1) = \frac{1}{4k}\,n^2 + \frac{k-1}{2k}\,n - \frac{t^2+2kt-2t}{4k}$$

tal com ha estat provat a l'anterior subsecció 4.5.3.

Si afegim també el número, $[(2s-1)\cdot(h-1)+(2s+2r-1)]\cdot D_m$, de sumes i restes implicades en totes les distàncies calculades (de forma similar a la subsecció 4.5.2), obtenim que el número total de sumes i restes computades per l'algorisme DMM, és:

$$S_m = [(h-1)s+(s+r)]\cdot\left(\frac{1}{4k}\,n^2 + \frac{k-1}{2k}\,n - \frac{t^2+2kt-2t}{4k}\right)+[(2s-1)\cdot(h-1)+(2s+2r-1)]\cdot D_m =$$

$$= (hs+r)\cdot\left(\frac{1}{4k}\,n^2 + \frac{k-1}{2k}\,n - \frac{t^2+2kt-2t}{4k}\right) + [(2s-1)\,h+2r]\cdot\left(\frac{3}{4k}\,n^2 + \frac{k-1}{k}\,n -\right.$$

$$- \frac{3t^2 + 4k^2 + 8kt - 4t - 4k}{4k}\Bigg) =$$

$$= \frac{7hs - 3h + 7r}{4k}\, n^2 + \frac{(k-1)\,(5hs - 2h + 5r)}{2k}\, n -$$

$$- \frac{(3t^2 + 4k^2 + 8kt - 4t - 4k)\,(2hs - h + 2r) + (t^2 + 2kt - 2t)\,(hs + r)}{4k}$$

Número de multiplicacions i divisions implicades en l'algorisme DMM

L'algorisme ha d'efectuar s divisions, corresponents a les s components dels vectors mitjana dintre els conjunts $G_l\,,\ 1 \leq l \leq h-1$ de la partició; i $s+r$ divisions, corresponents a les $s+r$ components dels vectors mitjana dintre el conjunt $G_h\,$. A més a més, el número de vectors mitjana que s'han de calcular, per a cada conjunt $G_l\,,\ 1 \leq l \leq h$ de la partició, és $\frac{n-t}{2k}$.

Si afegim també el número, $[s \cdot (h-1) + (s+r)] \cdot D_m$, de multiplicacions implicades en totes les distàncies calculades (de forma similar a la subsecció 4.5.2), obtenim que el número total de multiplicacions i divisions computades per l'algorisme DMM, és:

$$M_m = \frac{n-t}{2k} \cdot [s \cdot (h-1) + (s+r)] + D_m \cdot [s \cdot (h-1) + (s+r)] =$$

$$= \frac{n-t}{2k}\,(sh + r) + \left(\frac{3}{4k}\, n^2 + \frac{k-1}{k}\, n - \frac{3t^2 + 4k^2 + 8kt - 4t - 4k}{4k}\right)(sh + r) =$$

$$= \frac{3(hs + r)}{4k}\, n^2 + \frac{(hs + r)\,(2k-1)}{2k}\, n - \frac{(hs + r)\,(3t^2 + 4k^2 + 8kt - 2t - 4k)}{4k}$$

Número de comparacions per ordenar distàncies a l'algorisme DMM

Donada l'estreta relació entre el número de comparacions C_m i el número de distàncies calculades, el número de comparacions que ha de comprovar l'algorisme DMM, per ordenar distàncies, és molt similar al número total de distàncies calculades.

4.5.4 Conclusió

Així doncs, hem provat que el número de distàncies calculades i la complexitat computacional de l'algorisme DM (Mètode de microagregació de la "Distància Màxima"), sense emmagatzemar la matriu de distàncies, són funcions polinòmiques de tercer grau respecte del número de registres n que es microagreguen. Tanmateix, hem vist que el número de distàncies calculades i la complexitat computacional de l'algorisme DMM (Mètode de microagregació de la "Distància Màxima Modificat"), sense emmagatzemar la matriu de distàncies, són funcions polinòmiques quadràtiques respecte del número de registres n que es microagreguen. Per tant, en el nou mètode de microagregació DMM hi ha una reducció de la complexitat computacional molt considerable respecte del mètode de microagregació DM existent.

Capítol 5

Mesures de qualitat per comparar mètodes pertorbatius

5.1 Qualitat d'un mètode de control de la revelació pertorbatiu

Donada la gran diversitat de mètodes de control de la revelació estadística pertorbatius existents, se'ns presenta el repte de mesurar la seva qualitat. Els dos aspectes més importants a tenir en compte per caracteritzar la qualitat d'un mètode de control de la revelació pertorbatiu, són:

- La pèrdua d'informació produïda per la no publicació exacta de les dades originals.
- La pèrdua de confidencialitat que també es produeix en la publicació de dades que, tot i no ser idèntiques a les dades originals, sí que hauran de ser prou properes a la realitat.

Direm que la pèrdua d'informació és petita si l'estructura de les dades emmascarades publicades és molt similar a l'estructura del conjunt de dades original. Per poder assegurar que el conjunt de dades publicades és analíticament vàlid i analíticament interessant, serà molt important conservar l'estructura de les dades originals. Segons Winkler (1998) tenim que:

- Un conjunt de microdades és analíticament vàlid si aproximadament es conserva tot el que segueix (algunes de les següents condicions només seran aplicables a variables contínues):
  1. Mitjanes i covariàncies de petits conjunts de dades.
  2. Valors marginals d'algunes taules de dades.
  3. Almenys una característica distribucional.
- Un conjunt de microdades és analíticament interessant si conté un nombre suficient de variables per proveir mínimament les necessitats d'una determinada recerca.

Certament, la pèrdua d'informació dependrà principalment de l'ús i la utilitat que se'n faci de les dades publicades. Tanmateix, les possibles utilitats de les dades publicades són tan diverses, que serà molt difícil identificar-les totes al moment de publicar les dades. Per la qual cosa, apareix la imperant necessitat de disposar de quantificadors que mesuren d'una manera genèrica la pèrdua d'informacó,

reflectint, de la forma més exhaustiva possible, tot el dany produït en les dades originals pel mètode d'emmascarament; i, a la vegada, donada la seva generalitat, puguin ser utilitzats per comparar les diferents pèrdues d'informació provocades pels seus corresponents mètodes de control de la revelació estadística.

Tot i això, una completa valoració de la qualitat d'un mètode de control de la revelació estadística no pot limitar-se a mesurar solament la pèrdua d'informació, puix que també ha de ser mesurada una segona característica, tan important com aquella, de les dades publicades: la pèrdua de confidencialitat. Pèrdua de confidencialitat produïda per la publicació d'unes dades, que, si bé no són exactament iguals a les originals, sí que són suficientment properes a aquestes.

La millor opció serà el mètode de control de la revelació estadística que optimitzi l'equilibri entre la pèrdua d'informació i la pèrdua de confidencialitat, aspectes, per altra banda, ben contraposats entre ells.

Les mesures de pèrdua d'informació i de risc de revelació exposades en aquest capítol es basen en el treball previ Oganian (2003). Com a innovacions, introduïm:

- Un nou interval de confidencialitat basat en la desviació típica (en el treball previ l'únic interval que s'utilitzava era basat en rangs). El nou interval permet mesurar el risc de revelació de manera més acurada en dades asimètriques, com solen ser-ho les dades econòmiques.
- Una nova ponderació de les mesures de pèrdua d'informació i de pèrdua de confidencialitat a l'hora de calcular la mesura global de qualitat d'un mètode. Aquesta nova ponderació té en compte les diverses naturaleses de les mesures emprades.

Les innovacions anteriors han estat publicades a Domingo-Ferrer, Mateo-Sanz i Torres (2003).

5.2 Caracterització d'un conjunt de microdades contínues

En aquest capítol ens centrarem en l'anàlisi de microdades, on les variables publicades tenen un caràcter eminentment continu. Per al cas de variables categòriques, vegeu el treball Domingo-Ferrer i Torra (2001b).

De la definició de Winkler sobre la validesa analítica d'un conjunt de dades, se'n dedueixen les següents formes de comparar quantitativament l'estructura dels dos conjunts de dades (dades originals versus dades pertorbades):

- Comparació directa dels dos conjunts de dades.
- Comparació de les covariàncies dels dos conjunts de dades. Una petita pèrdua d'informació hauria de traduir-se amb petites diferències entre les matrius de covariàncies dels dos conjunts de dades.
- Comparació entre les matrius de correlació dels dos conjunts de dades. La interpretació seria la mateixa que per a les matrius de covariància.
- Comparar les correlacions entre variables i components principals en els dos conjunts de dades. Si hi ha poca pèrdua d'informació, llavors, el model de correlació ha de ser similar en els dos conjunts de dades.
- Considerar el percentatge de variabilitat de cada variable explicada per la primera component principal. Aquest percentatge s'anomena comunalitat; i no tindria que ser molt diferent en els dos conjunts de dades, cas que suposéssim poca pèrdua d'informació.

- Comparar les coordenades de les dades individuals en termes de les components principals. Si la pèrdua d'informació és petita, aquestes coordenades haurien de ser semblants per als dos conjunts de dades.

Com que totes les anteriors caracteritzacions són condicions importants per assegurar una petita pèrdua d'informació, llavors, cap d'elles, considerada aïlladament, pot considerar-se suficient; sinó que, únicament, quan totes aquestes caracteritzacions mostrin petites diferències entre les dades originals i les dades pertorbades, podrem concluir realment que la pèrdua d'informació és petita.

5.2.1 Mesures per a la caracterització

Suposem un conjunt de microdades amb n individus (registres) $I_1, I_2, \ldots, I_n$ i p variables contínues $Z_1, Z_2, \ldots, Z_p$.

Sigui $\mathbf{X}$ la matriu de n files i p columnes que representa el conjunt de microdades original (les files són els registres, que es corresponen amb els individus, i les columnes són les variables).

Sigui $\mathbf{X'}$ la matriu de n files i p columnes que representa el conjunt publicat de microdades modificades.

Les magnituds i les corresponents matrius que utilitzarem per caracteritzar la informació continguda en els conjunts de dades, seran les següents:

- Els vectors mitjana $\overline{\mathbf{X}}$ i $\overline{\mathbf{X}}'$ de cada conjunt de dades (originals i modificades), amb p coordenades cadascun, que es corresponen amb les mitjanes de les p variables estudiades.
- Les matrius de covariància $\mathbf{V}$ i $\mathbf{V'}$, calculades a partir de $\mathbf{X}$ i de $\mathbf{X'}$ respectivament.
- Les matrius de correlació $\mathbf{R}$ i $\mathbf{R'}$, amb p files i p columnes cadascuna, calculades a partir de $\mathbf{X}$ i de $\mathbf{X'}$ respectivament.
- Les matrius de correlació $\mathbf{RF}$ i $\mathbf{RF'}$ entre les p variables i els p factors principals $\mathbf{PC_1}, \ldots, \mathbf{PC_p}$ obtinguts a través de l'anàlisi en components principals.
- La comunalitat entre cada una de les p variables i la primera component principal $\mathbf{PC_1}$ (o altres $\mathbf{PC_i}$'s). La comunalitat és el percentatge de variabilitat de cada variable explicada per $\mathbf{PC_1}$ (o $\mathbf{PC_i}$). Sigui $\mathbf{C}$ el vector de comunalitats per a $\mathbf{X}$ i $\mathbf{C'}$ el corresponent vector per a $\mathbf{X'}$.
- Les matrius dels factors principals $\mathbf{F}$ i $\mathbf{F'}$. Les columnes de la matriu $\mathbf{F}$ són els factors principals, pels quals han de multiplicar-se el valors de les variables de la matriu $\mathbf{X}$ per obtenir la projecció dels registres sobre els eixos de les components principals. $\mathbf{F'}$ és la matriu corresponent a $\mathbf{X'}$.

5.2.2 Format de les matrius

La matriu $\mathbf{X}$ té el següent format:

	Z_1	$\cdots$	Z_j	$\cdots$	Z_p
I_1					
$\vdots$					
I_i			x_{ij}		
$\vdots$					
I_n					

La matriu **V** té el següent format:

	Z_1	$\cdots$	Z_j	$\cdots$	Z_p
Z_1					
$\vdots$					
Z_i			v_{ij}		
$\vdots$					
Z_p					

La matriu **R** té el següent format:

	Z_1	$\cdots$	Z_j	$\cdots$	Z_p
Z_1					
$\vdots$					
Z_i			r_{ij}		
$\vdots$					
Z_p					

La matriu **RF** té el següent format:

	PC_1	$\cdots$	PC_j	$\cdots$	PC_p
Z_1					
$\vdots$					
Z_i			rf_{ij}		
$\vdots$					
Z_p					

El vector de comunalitats **C** té el següent format:

	Comunalitat
Z_1	c_1
$\vdots$	
Z_i	c_i
$\vdots$	
Z_p	c_p

La matriu dels factors principals **F** té el següent format:

	PC_1	$\cdots$	PC_j	$\cdots$	PC_p
Z_1					
$\vdots$					
Z_i			f_{ij}		
$\vdots$					
Z_p					

5.3 Mesures per a la pèrdua d'informació

Com ha estat explicat anteriorment, podem mesurar la pèrdua d'informació en funció de les diferències estructurals entre el conjunt de dades original i el conjunt de dades pertorbades.

Puix que no sembla existir una única mesura quantitativa que reflecteixi completament totes les diferències estructurals, proposem mesurar la pèrdua d'informació a través de les discrepàncies entre les matrius $\mathbf{X}$, $\overline{\mathbf{X}}$, $\mathbf{V}$, $\mathbf{R}$, $\mathbf{RF}$, $\mathbf{C}$ i $\mathbf{F}$, calculades a partir de les dades originals, i les corresponents matrius $\mathbf{X}'$, $\overline{\mathbf{X}}'$, $\mathbf{V}'$, $\mathbf{R}'$, $\mathbf{RF}'$, $\mathbf{C}'$ i $\mathbf{F}'$, calculades a partir de les dades pertorbades.

Aquestes discrepàncies, que reflecteixen les diferències estructurals entre els dos conjunts de dades, poden ser mesurades, almenys, de tres maneres diferents (Domingo-Ferrer i Mateo-Sanz 1999):

Mitjana de l'error quadràtic. És la suma dels quadrats de les diferències entre les cel·les corresponents de les dues matrius, dividida pel número de cel·les que han intervingut en la suma.

Mitjana de l'error absolut. És la suma dels valors absoluts de les diferències entre les cel·les corresponents de les dues matrius, dividida pel número de cel·les que han intervingut en la suma.

Variació mitjana. És la suma del percentatge de variació, en valor absolut, de les cel·les de la matriu calculada a partir de les dades pertorbades, respecte les cel·les de la matriu calculada a partir de les dades originals, dividida pel número de cel·les que han intervingut en la suma. Aquesta tercera mesura té l'avantatge de no estar afectada per canvis d'escala de les variables.

5.3.1 Mesures per a les discrepàncies $\mathbf{X} - \mathbf{X}'$

De les tres maneres diferents de mesurar detallades anteriorment, se'n dedueixen les tres següents fórmules per a les discrepàncies entre les dades originals i les dades pertorbades:

Mitjana de l'error quadràtic

$$\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}(x_{ij}-x'_{ij})^2}{np} \tag{5.1}$$

Mitjana de l'error absolut

$$\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}|x_{ij}-x'_{ij}|}{np} \tag{5.2}$$

Variació mitjana

$$\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}\frac{|x_{ij}-x'_{ij}|}{|x_{ij}|}}{np} \tag{5.3}$$

Nota 1 *A l'expressió (5.3), si $x_{ij} = 0$ i $x'_{ij} \neq 0$, llavors dividiríem per $|x'_{ij}|$ en lloc de dividir per $|x_{ij}|$; cas que els dos siguin zero, $x_{ij} = x'_{ij} = 0$, llavors, el corresponent terme de la suma s'igualaria a zero. Per evitar la divisió per zero, es pot seguir aquesta mateixa estratègia a totes les següents expressions on aparegui la variació mitjana.*

5.3.2 Mesures per a les discrepàncies $\overline{\mathbf{X}} - \overline{\mathbf{X}}'$

Les discrepàncies entre els vectors mitjana dels dos conjunts de dades poden ser mesurades per:

Mitjana de l'error quadràtic

$$\frac{\sum_{j=1}^{p}(\bar{x}_j - \bar{x}'_j)^2}{p} \tag{5.4}$$

Mitjana de l'error absolut

$$\frac{\sum_{j=1}^{p}|\bar{x}_j - \bar{x}'_j|}{p} \tag{5.5}$$

Variació mitjana

$$\frac{\sum_{j=1}^{p}\frac{|\bar{x}_j - \bar{x}'_j|}{|\bar{x}_j|}}{p} \tag{5.6}$$

5.3.3 Mesures per a les discrepàncies $\mathbf{V} - \mathbf{V}'$

Podem mesurar les discrepàncies entre les matrius de covariància dels dos conjunts de dades, a través de:

Mitjana de l'error quadràtic

$$\frac{\sum_{j=1}^{p}\sum_{1\leq i\leq j}(v_{ij} - v'_{ij})^2}{\frac{p(p+1)}{2}} \tag{5.7}$$

Mitjana de l'error absolut

$$\frac{\sum_{j=1}^{p}\sum_{1\leq i\leq j}|v_{ij} - v'_{ij}|}{\frac{p(p+1)}{2}} \tag{5.8}$$

Variació mitjana

$$\frac{\sum_{j=1}^{p}\sum_{1\leq i\leq j}\frac{|v_{ij} - v'_{ij}|}{|v_{ij}|}}{\frac{p(p+1)}{2}} \tag{5.9}$$

Cal observar que només hem comparat els elements que ocupen el triangle superior de les matrius $\mathbf{V}$ i $\mathbf{V}'$, perquè aquestes dues matrius són simètriques.

Cas que vulguéssim concentrar el nostre estudi únicament sobre les discrepàncies entre les variàncies de les variables; és a dir, sobre els elements diagonals de les matrius de covariància, tindríem:

Mitjana de l'error quadràtic

$$\frac{\sum_{j=1}^{p}(v_{jj} - v'_{jj})^2}{p} \tag{5.10}$$

Mitjana de l'error absolut

$$\frac{\sum_{j=1}^{p}|v_{jj} - v'_{jj}|}{p} \tag{5.11}$$

Variació mitjana

$$\frac{\sum_{j=1}^{p}\frac{|v_{jj} - v'_{jj}|}{v_{jj}}}{p} \tag{5.12}$$

5.3.4 Mesures per a les discrepàncies $\mathbf{R} - \mathbf{R}'$

Les discrepàncies entre les matrius de correlació poden ser mesurades a través de:

Mitjana de l'error quadràtic

$$\frac{\sum_{j=1}^{p}\sum_{1\leq i<j}(r_{ij}-r'_{ij})^2}{\frac{p(p-1)}{2}} \tag{5.13}$$

Mitjana de l'error absolut

$$\frac{\sum_{j=1}^{p}\sum_{1\leq i<j}|r_{ij}-r'_{ij}|}{\frac{p(p-1)}{2}} \tag{5.14}$$

Variació mitjana

$$\frac{\sum_{j=1}^{p}\sum_{1\leq i<j}\frac{|r_{ij}-r'_{ij}|}{|r_{ij}|}}{\frac{p(p-1)}{2}} \tag{5.15}$$

Només han estat comparats els elements que ocupen el triangle superior de les matrius $\mathbf{R}$ i $\mathbf{R}'$, exceptuant els elements diagonals, ja que es tracta de matrius simètriques amb tots els elements de la diagonal principal iguals a la unitat ($r_{ii} = r'_{ii} = 1$ per a $1 \leq i \leq p$).

5.3.5 Mesures per a les discrepàncies $\mathbf{RF} - \mathbf{RF}'$

Sigui $(w_1, \ldots, w_p)$ el vector de pesos, on w_j és el percentatge de la inèrcia total explicada per la j-èsima component principal en les dades originals.

Podem utilitzar aquests pesos per construir les següents mesures que quantifiquen les discrepàncies entre les matrius $\mathbf{RF}$ i $\mathbf{RF}'$:

Mitjana de l'error quadràtic

$$\frac{\sum_{j=1}^{p} w_j \sum_{i=1}^{p}(rf_{ij}-rf'_{ij})^2}{p^2}$$

Mitjana de l'error absolut

$$\frac{\sum_{j=1}^{p} w_j \sum_{i=1}^{p}|rf_{ij}-rf'_{ij}|}{p^2}$$

Variació mitjana

$$\frac{\sum_{j=1}^{p} w_j \sum_{i=1}^{p}\frac{|rf_{ij}-rf'_{ij}|}{|rf_{ij}|}}{p^2}$$

5.3.6 Mesures per a les discrepàncies $\mathbf{C} - \mathbf{C}'$

Podem mesurar la discrepància entre les comunalitats de la següent manera:

Mitjana de l'error quadràtic

$$\frac{\sum_{i=1}^{p}(c_i-c'_i)^2}{p}$$

Mitjana de l'error absolut

$$\frac{\sum_{i=1}^{p} |c_i - c_i'|}{p}$$

Variació mitjana

$$\frac{\sum_{i=1}^{p} \frac{|c_i - c_i'|}{|c_i|}}{p}$$

5.3.7 Mesures per a les discrepàncies $\mathbf{F} - \mathbf{F}'$

Sigui $(w_1, \ldots, w_p)$ el vector de pesos de la Subsecció 5.3.5. Llavors obtenim les següents mesures per a les discrepàncies $\mathbf{F} - \mathbf{F}'$:

Mitjana de l'error quadràtic

$$\frac{\sum_{j=1}^{p} w_j \sum_{i=1}^{p} (f_{ij} - f_{ij}')^2}{p^2}$$

Mitjana de l'error absolut

$$\frac{\sum_{j=1}^{p} w_j \sum_{i=1}^{p} |f_{ij} - f_{ij}'|}{p^2}$$

Variació mitjana

$$\frac{\sum_{j=1}^{p} w_j \sum_{i=1}^{p} \frac{|f_{ij} - f_{ij}'|}{|f_{ij}|}}{p^2}$$

5.4 Mesures per a la pèrdua de confidencialitat

Una valoració de la qualitat d'un mètode de control de la revelació estadística, no pot limitar-se a mesurar únicament la pèrdua d'informació, ja que també haurà d'avaluar una segona característica de les dades publicades, tan important com la pèrdua d'informació: la pèrdua de confidencialitat produïda pel lliurament d'unes dades, si bé no exactament iguals a les dades originals, sí, certament, prou properes i semblants a elles. La millor opció serà el mètode de control de la revelació que optimitzi l'equilibri entre la pèrdua d'informació i la pèrdua de confidencialitat (Sebé, Domingo-Ferrer, Mateo-Sanz i Torra 2002)(Dandekar, Domingo-Ferrer i Sebé 2002).

Per comprendre millor l'equilibri entre aquestes dues magnituds, considerem els dos casos extrems dintre el marc de la publicació de dades:

- Encriptar les dades estadístiques, de manera que sigui pràcticament impossible la seva revelació; però, com a conseqüència, no donar cap tipus d'informació (màxima pèrdua d'informació).
- L'altre cas extrem seria no fer cap emmascarament de les dades, perquè els usuaris puguessin deduir qualssevol tipus de resultats completament exactes; però, en aquest cas, facilitaríem excessivament la revelació de dades confidencials, especialment en el cas de microdades.

Tots els estudis sobre el risc de revelació estan centrats bàsicament en mètodes de mostreig, on es publica una mostra de les dades originals. En aquest cas, el risc de revelació es mesura com la probabilitat que un individu amb unes característiques úniques a la mostra, també sigui únic a la població (Elliot, Skinner i Dale 1999, Skinner, Marsh, Openshaw i Wymer 1994). Quan la mida de la mostra és semblant a la mida de la població, aquesta probabilitat pot ser perillosament alta, de manera que un intrús que localitzés un individu amb característiques úniques a la mostra publicada, podria estar quasibé segur que el mateix individu també seria únic a la població; la qual cosa conduiria a la identificació de l'individu.

La propietat d'unicitat no és tan rellevant quan s'aplica a un mètode pertorbatiu, ja que, en aquest cas, es publiquen totes les microdades distorsionades. Per altra banda, no hi ha molta literatura sobre el risc de revelació que pugui ser aplicada a un ventall gran de mètodes pertorbatius, sinó que les mesures sobre el risc de revelació tendeixen a ser més bé específiques per a cada mètode particular.

Tot i això, existeixen aproximacions empíriques, com les tècniques d'enllaç de registres, que, certament, aporten una visió generalitzadora per avaluar la pèrdua de confidencialitat en el marc dels mètodes pertorbatius.

Tot seguit descrivim dues aproximacions a l'enllaç de registres, que produeixen mesures de pèrdua de confidencialitat, i dues mesures més basades en intervals de revelació:

Enllaç de registres basat en distàncies (**ERD**)

Aquesta aproximació, descrita a Pagliuca i Seri (1999) per al cas de microagregació i utilitzant la distància euclidiana, es pot generalitzar per a qualsevol mètode pertorbatiu en el que es pugui definir una distància entre els valors de les dades originals i els valors de les dades pertorbades.

Dintre d'un context d'enllaç de registres, aquest mètode suposa que un intrús ha aconseguit obtenir un conjunt extern de dades que conté un subconjunt de variables clau de les mateixes variables presents en el conjunt de dades modificades.

També suposa que l'intrús intenta enllaçar el conjunt de dades modificades amb el conjunt extern de dades que posseeix, utilitzant el subconjunt de variables comunes, per descobrir dades individuals originals.

L'enllaç s'obté buscant els registres més propers, mitjançant distàncies, entre els registres del conjunt extern de dades i els registres del conjunt de dades modificades.

Per mesurar el risc de revelació en aquest context, se segueix el següent procés:

1. Per a cada registre del conjunt original de dades extern, es calcula la distància a cada registre del conjunt modificat de dades, utilitzant únicament el subconjunt de variables comunes als dos conjunts de dades (les distàncies s'estandarditzen per evitar problemes d'escala).
2. Se seleccionen els registres *més proper* i *segon més proper* del conjunt modificat de dades.

Un registre del conjunt original s'etiqueta com a "enllaçat" quan el registre més proper del conjunt modificat de dades coincideix amb el corresponent registre original. Un registre del conjunt original s'etiqueta com a "enllaçat de segon nivell" quan el segon registre més proper del conjunt modificat de dades coincideix amb el corresponent registre original (es podrien considerar successivament els tercer, quart, ... nivells).

La mesura de pèrdua de confidencialitat (ERD) s'avalua, doncs, mitjançant el percentatge de registres "enllaçats" i "enllaçats de segon nivell".

Nota1. Les distàncies es calculen utilitzant únicament el subconjunt de variables comunes al conjunt extern de dades i al conjunt modificat.

Nota2. Recordem que aquest mètode, per calcular les distàncies, requereix reescalar les variables i, de vegades, definir diferents pesos per a les variables: per exemple, a l'aplicació proposada a Pagliuca i Seri (1999), totes les variables tenen el mateix pes.

Enllaç de registres probabilístic (**ERP**)

Aquesta aproximació probabilística a l'enllaç de registres, que va ser aplicada al *Census of Tampa, Florida (1985)*, es descriu a Jaro (1989).

Aquest algorisme d'aparellament utilitza un model d'assignació de suma lineal per enllaçar registres dels dos conjunts de dades (el conjunt extern i el conjunt modificat). Llavors, una mesura de pèrdua de confidencialitat és el percentatge de registres correctament aparellats. Només s'utilitzen les variables comunes al conjunt extern i al conjunt modificat de dades.

Tot i que aquest mètode no és tan senzill com el basat en distàncies, és prou atractiu perquè només requereix que l'usuari introdueixi dues probabilitats com a paràmetres d'entrada: una, és el límit superior de la probabilitat d'un enllaç fals, i l'altra, és el límit superior de la probabilitat d'un no-enllaç fals.

Interval de confidencialitat sobre el número de registres (**ICN**)

Aquesta mesura del risc de revelació no està basada en enllaç de registres, sinó que la idea és valorar el nivell de revelació inherent a intervals construïts sobre el conjunt modificat de dades.

Per a cada registre del conjunt de dades modificades es calculen intervals per rangs de la següent manera: cada variable s'ordena independentment de les altres i es defineixen intervals per rangs centrats en cadascun dels valors que la variable té a cada registre, de manera que l'amplitud de cada interval sigui igual o més petita que un $p\%$ (p constant fixada) del número total de registres.

Així, doncs, una mesura de la pèrdua de confidencialitat (ICN) seria la proporció de registres originals que es troben totalment dins de l'interval per rangs centrat en el seu corresponent registre modificat. Una proporció del 100% significaria que un intrús està completament segur que el valor original de cada variable d'un determinat registre es troba dintre d'un interval ben concret centrat en el valor modificat de la variable.

Nota. Aquest procediment es desenvolupa per a totes les variables del conjunt modificat de dades, de manera que considerem totes les variables comunes als dos conjunts de dades.

Interval de confidencialitat sobre la desviació típica (**ICD**)

Aquesta altra aproximació a la mesura del risc de revelació a través de la valoració del nivell de revelació inherent a intervals, és prou semblant a l'anterior, ja que també suposa ordenar cada variable independentment de les altres en el conjunt modificat de dades, i definir intervals centrats en cadascun dels valors que la variable té a cada registre; però, a diferència de l'anterior aproximació, ara, l'amplitud de cada interval serà igual o més petita que un $p\%$ (p constant també fixada) del valor de la desviació típica de cada variable.

Nota. Aquest procediment es desenvolupa també per a totes les variables del conjunt modificat de dades, de manera que considerem totes les variables comunes als dos conjunts de dades.

5.4.1 Escenaris de revelació

Suposem un conjunt de microdades amb n individus (registres) $I_1, I_2, \ldots, I_n$ i p variables contínues $Z_1, Z_2, \ldots, Z_p$.

Anomenarem escenari de revelació, cadascun dels possibles camps de coneixement que pot tenir un intrús respecte dades confidencials. Així doncs, els escenaris de revelació es correspondran amb els diferents graus de coneixement que l'intrús té d'aquestes dades confidencials.

Podem diferenciar dos aspectes d'aquest grau de coneixement:

1. Número de variables conegudes.
2. Per a cada número de variables conegudes, combinacions possibles de variables.

Respecte la primera diferenciació: número k de variables conegudes, totes les possibilitats són:

$$k = 1 \ ; \ k = 2 \ ; \ k = 3 \ ; \ \ldots \ ; \ k = p$$

Respecte la segona diferenciació: suposant k variables conegudes; llavors, el número de combinacions de k variables cadascuna, respecte un total de p variables, és:

$$\left(\begin{array}{c} p \\ k \end{array} \right)$$

D'aquesta manera, resulta que el número total, E, d'escenaris de revelació, serà:

$$E = \left(\begin{array}{c} p \\ 1 \end{array} \right) + \left(\begin{array}{c} p \\ 2 \end{array} \right) + \left(\begin{array}{c} p \\ 3 \end{array} \right) + \cdots + \left(\begin{array}{c} p \\ p \end{array} \right) = \sum_{k=1}^{p} \left(\begin{array}{c} p \\ k \end{array} \right) =$$

$$= (1+1)^p - 1 = 2^p - 1$$

Per a cada un d'aquests $2^p - 1$ escenaris de revelació, es pot calcular el seu corresponent risc de revelació mitjançant les quatre mesures descrites en aquesta secció.

5.5 Mesura global per comparar mètodes pertorbatius

Pensem que una mesura global, **MG**, per valorar la qualitat d'un mètode de control de la revelació pertorbatiu ha de donar la mateixa importància a la pèrdua d'informació, **PI**, que a la pèrdua de confidencialitat, **PC**. Per això ponderarem al 50% cada una d'aquestes pèrdues:

$$\mathbf{MG} = 0.5 \cdot \mathbf{PI} + 0.5 \cdot \mathbf{PC}$$

Respecte la pèrdua d'informació, considerarem les següents mesures no basades en components principals:

PI1 : Variació mitjana de les discrepàncies entre les dades $\mathbf{X} - \mathbf{X}'$

PI2 : Variació mitjana de les discrepàncies entre mitjanes $\overline{\mathbf{X}} - \overline{\mathbf{X}}'$

PI3 : Variació mitjana de les discrepàncies entre covariàncies $\mathbf{V} - \mathbf{V}'$

PI4 : Variació mitjana de les discrepàncies entre variàncies $\mathbf{S} - \mathbf{S}'$

PI5 : Mitjana de l'error absolut de les discrepàncies entre correlacions $\mathbf{R} - \mathbf{R}'$

Hem preferit la variació mitjana en quatre de les anteriors mesures perquè no està afectada pels canvis d'escala de les variables. Tot i això, per mesurar les discrepàncies entre els coeficients de correlació hem triat la mitjana de l'error absolut per dos raons: les correlacions no estan afectades pels canvis d'escala de les variables; i, d'altra banda, la variació mitjana podria ser enormement gran quan els coeficients de correlació originals tinguessin valors molt propers a 0.

Per trobar una mesura global que ponderi adequadament cada una de les diferents discrepàncies reflectides per les anteriors cinc mesures, definirem les tres següents naturaleses, que ens serviran per classificar aquestes mesures:

Discrepància directa entre dades

En aquesta naturalesa inclourem:

- PI1 : Variació mitjana de les discrepàncies entre les dades $\mathbf{X}-\mathbf{X}'$, que reflecteix de manera immediata la distorsió de les dades publicades respecte les dades originals.

Discrepància afectada per la centralització i la dispersió de cada variable

En aquesta naturalesa inclourem:

- PI2 : Variació mitjana de les discrepàncies entre mitjanes $\overline{\mathbf{X}}-\overline{\mathbf{X}}'$.
- PI4 : Variació mitjana de les discrepàncies entre variàncies $\mathbf{S}-\mathbf{S}'$, les quals suposen considerar cada variable independentment de les altres.

Discrepància afectada per la relació entre cada dues variables

En aquesta naturalesa inclourem:

- PI3 : Variació mitjana de les discrepàncies entre covariàncies $\mathbf{V}-\mathbf{V}'$.
- PI5 : Mitjana de l'error absolut de les discrepàncies entre correlacions $\mathbf{R}-\mathbf{R}'$, les quals suposen considerar relacions entre parelles de variables.

Per equilibrar la influència de cadascuna d'aquestes tres naturaleses en la mesura global sobre la pèrdua d'informació, cadascuna tindrà un pes igual a $\frac{1}{3}$; la qual cosa suposa assignar un pes de $\frac{1}{6}$ a PI2, PI4, PI3 i PI5, per valorar equitativament cada mesura.

Així doncs, resulta la següent mesura que engloba tota la pèrdua d'informació:

$$\mathbf{PI} = \frac{1}{3}\cdot PI1 + \frac{1}{6}\cdot PI2 + \frac{1}{6}\cdot PI4 + \frac{1}{6}\cdot PI3 + \frac{1}{6}\cdot PI5$$

Pel que fa a la mesura sobre la pèrdua de confidencialitat, distingirem també dues naturaleses:

Pèrdua de confidencialitat basada en enllaç de registres

Aquesta naturalesa inclourà l'enllaç de registres basat en distàncies ERD, que reflecteix el percentatge de registres "enllaçats" mitjançant el càlcul de distàncies entre les variables del conjunt extern de dades conegudes i el conjunt modificat de dades.

Pèrdua de confidencialitat basada en intervals d'estimació

Aquesta naturalesa inclourà les dues aproximacions referides a intervals de confidencialitat: ICN i ICD, les quals valoren el nivell de revelació inherent a intervals construïts sobre el conjunt modificat de dades.

Com que donarem la mateixa importància a cadascuna d'aquestes dues naturaleses, tindran totes dues el mateix pes $\frac{1}{2}$; la qual cosa suposa donar un pes d'$\frac{1}{4}$ a ICN i a ICD perquè siguin equitativament valorades. Per tot això, la mesura global de pèrdua de confidencialitat que considerarem en el nostre estudi, és:

$$\mathbf{PC} = 0.50\cdot ERD + 0.25\cdot ICN + 0.25\cdot ICD$$

Capítol 6

Comparació de mètodes pertorbatius

Aquest capítol presenta un estudi comparatiu sobre la qualitat dels mètodes de control de la revelació pertorbatius més rellevants, actualment existents, per a la protecció de microdades, on les variables publicades tenen caràcter continu. Una valoració sobre la qualitat d'aquests mètodes suposa, com ha estat ja comentat anteriorment, avaluar la pèrdua d'informació i el risc de revelació inherents a l'aplicació de cada un d'aquests mètodes.

El risc de revelació ha estat mesurat empíricament a través de dues aproximacions: enllaç de registres basat en distàncies i construcció d'intervals de confidencialitat. Per altra banda, la pèrdua d'informació ha estat valorada a partir d'un conjunt de mesures desenvolupades al capítol anterior.

Les dades individuals que han servit per realitzar l'estudi han estat tretes mitjançant el *Census's Data Extraction System*.

6.1 Mètodes i paràmetres usats

Quan treballem amb variables contínues, no és molt adient l'aplicació de mètodes de mostreig per emmascarar dades estadístiques confidencials, puix que aquests mètodes deixen sense modificar les variables de tots els informants seleccionats a la mostra; de manera que si un determinat valor v_i d'una variable apareix en un fitxer publicat, llavors, com que és molt improbable que una variable contínua tingui valors iguals en individus diferents, la identificació de la persona informant és prou fàcil, donada la unicitat del valor de la variable.

Per això, quan treballem amb variables contínues que contenen dades confidencials, creiem que els mètodes de control de la revelació pertorbatius, mitjançant els quals podem publicar totes les microdades modificades, són molt més adients i segurs.

Els mètodes pertorbatius considerats en aquest estudi han estat els següents:

Pertorbació additiva aleatòria (`Addt`p abreviat)

Les dades originals es modifiquen afegint una pertorbació aleatòria que segueix una distribució normal (Kim 1986). Si la desviació típica de la variable original és igual a s, llavors, la distribució normal que s'afegeix té mitjana 0, i desviació típica ps.

Els valors de p considerats han estat 0.01, 0.02, 0.04, 0.06, 0.08, fins a 0.2, amb increments de 0.02.

Pertorbació de dades segons una distribució de probabilitat (`Distr` abreviat)

Per a cada variable del fitxer original, es busca la distribució de probabilitat que millor s'ajusta a les seves dades. Aquesta distribució s'utilitza, després, per generar el conjunt modificat de dades (K. Liew, Choi i J. Liew 1985).

Aquest mètode no depèn de cap paràmetre. En la nostra experimentació hem usat el software Crystal Ball (`http://www.cbpro.com`) per trobar la distribució que millor ajusta unes dades. Les distribucions que compara aquest software són: lognormal, Weibull, uniforme, beta, gamma, logística, Pareto i valor extrem.

El criteri emprat per trobar la millor distribució ha estat la minimització de l'estadístic de Kolmogorov-Smirnov.

Remostreig

Prenem t mostres aleatòries independents $X_1, \ldots, X_t$ de valors d'una variable original V_i. Ordenem totes les mostres obtingudes, i construïm la variable modificada V_i', agafant, com a primer valor, la mitjana de tots els primers valors de les t mostres; com a segon valor, la mitjana de tots els segons valors de les t mostres; i així successivament ....

Aquesta tècnica ha estat estudiada per a $t = 1$ (`Remost1`), i per a $t = 3$ (`Remost3`).

Microagregació

Els registres individuals s'ajunten en petits grups de mida almenys k (Defays i Nanopoulos 1993, Domingo-Ferrer i Mateo-Sanz 2002). Les dades modificades que es publiquen són les mitjanes dels valors de cada variable corresponents als individus que formen cada grup.

S'han considerat diferents variants de la microagregació: ordenació individual (`MicOI`k); projecció de les dades mitjançant les z-puntuacions (`MicZ`k), o bé mitjançant la primera component principal (`MicPCP`k); microagregació sense projectar les dades, utilitzant 2 variables simultàniament (`Mic2mul`k), o bé 3 variables simultàniament (`Mic3mul`k), o bé 4 variables simultàniament (`Mic4mul`k), o bé 5 variables simultàniament (`Mic5mul`k), o bé 6 variables simultàniament (`Mic6mul`k), o finalment, totes les variables simultàniament (`Micmul`k).

S'han provat valors de k entre 3 i 20, per a cada variant.

Hem utilitzat el nou algorisme de microagregació multivariant de la Distància Màxima Modificat (DMM) per microagregar sense projectar les dades a les següents variants: `Mic2mul`k, `Mic3mul`k, `Mic4mul`k, `Mic5mul`k, `Mic6mul`k i `Micmul`k. A més a més, en totes aquestes variants, l'algorisme de la Distància Màxima Modificat (DMM) ha estat aplicat solament a la partició natural del conjunt de variables. Anomenem partició natural del conjunt de variables, la partició $\mathbf{G}$ formada de la següent manera:

Sigui $\mathbf{V} = \{V_1, V_2, V_3, \ldots, V_p\}$ el conjunt format per les p variables observades; i suposem que estem desenvolupant l'algorisme DMM per fer microagregació multivariant amb grups de s variables.

Si $p = hs + r$, essent $0 \leq r < s$; llavors la partició natural és $\mathbf{G} = \{G_1, G_2, G_3, \ldots, G_h\}$, on

$$G_1 = \{V_1, V_2, V_3, \ldots, V_s\}$$

$$G_2 = \{V_{s+1}, V_{s+2}, V_{s+3}, \ldots, V_{2s}\}$$

$$\vdots$$

$$G_i = \{V_{(i-1)s+1}, V_{(i-1)s+2}, V_{(i-1)s+3}, \ldots, V_{is}\}$$

$$\vdots$$

$$G_h = \{V_{(h-1)s+1}, V_{(h-1)s+2}, V_{(h-1)s+3}, \ldots, V_p\}$$

Pèrdua per compressió (`JPEG`q abreviat)

Aquest mètode és prou recent i ha estat proposat pels seus mateixos autors per al tractament de dades contínues. La idea del mètode és interpretar la matriu de microdades inicial, on les files corresponen als registres individuals i les columnes corresponen a les variables, com una imatge. La pèrdua per compressió, i més específicament, l'algorisme JPEG (Joint Photographic Experts Group `http://www.jpeg.org`), s'aplica sobre aquesta imatge; i la imatge comprimida resultant serà el conjunt modificat de dades.

Caldrà especificar el nivell de compressió adequat, d'acord amb el nivell de qualitat q que desitgem. Els valors provats del paràmetre de qualitat q han estat des del 10% fins al 95%, amb increments del 5%.

Intercanvi de dades (*Rank swapping*) (`Rank`p abreviat)

Encara que aquest mètode fou inicialment pensat per aplicar-se únicament a variables ordinals, també es pot aplicar a qualsevol variable numèrica (Moore 1996).

Primerament, els valors d'una variable V_i s'ordenen de forma ascendent; i, posteriorment, cada valor ordenat de V_i s'intercanvia amb un altre valor ordenat de la mateixa variable, escollit aleatòriament entre un rang de valors restringit; la qual cosa significa que la diferència entre els dos valors intercanviats no pot ser més gran que un p% del número total de registres.

Els valors de p provats en aquest estudi varien entre 1 i 20, amb increments d'una unitat.

6.2 Descripció del conjunt de dades

El conjunt de dades que ha servit per fer aquest estudi comparatiu es va aconseguir a través del *Data Extraction System of the U. S. Bureau of the Census* (`http://www.census.gov/DES/www/welcome.html`).

6.2.1 Procediment d'extracció de dades

L'extracció de les dades a través del *Data Extraction System* (*DES*) es va desenvolupar de la següent manera:

Nivell 1 De totes les dades disponibles, es va triar "the Current Population Survey".

Nivell 2 Concretament es van escollir les dades corresponents a l'any 1995.

Nivell 3 Aquestes dades es van treure del grup de fitxers "March Questionnaire Supplement - Person Data Files".

Nivell 4 Per trobar les variables numèriques es va usar l'opció "File content documentation". I es van seleccionar els registres MAX, amb les 54 variables numèriques (alfabèticament ordenades, la primera i l'última variables seleccionades van ser AERNLWT i WSWAL, respectivament).

El resultat final de tot aquest procediment va ser un fitxer amb 149642 registres ASCII.

6.2.2 Selecció de variables

De les 54 variables obtingudes, 38 variables tenien valor 0 per a més de 120000 registres escollits; per la qual cosa, van ser eliminades.

De les 16 variables que van quedar, van ser excloses tres variables més perquè els seus valors estaven dintre d'un conjunt molt restringit, de manera que quasibé podien ser considerades com a categòriques.

Finalment, les 13 variables que van quedar, són: AFNLWGT, AGI, EMCONTRB, ERNVAL, FEDTAX, FICA, INTVAL, PEARNVAL, POTHVAL, PTOTVAL, STATETAX, TAXINC, WSALVAL.

6.2.3 Selecció de registres

Dels 149642 registres obtinguts, n'hi havia que tenien, dintre el rang de les 13 variables, un considerable número de valors 0. En una variable contínua, generalment no s'espera que el valor 0 aparegui amb molta freqüència. També hi havia valors desapareguts de les variables entre els 149642 registres. Per tot això, els registres amb un considerable número de zeros o bé els registres per als que mancava valor en alguna de les 13 variables, van ser eliminats.

A més a més, es va decidir que el número de registres del conjunt de dades d'aquest estudi no superés els 1200 registres perquè, cas que s'utilitzés l'experimentació amb el software de l'enllaç de registres probabilístic, pugués ser repetida amb un temps computacional raonable.

Així doncs, per acabar de reduir el conjunt de registres, es va fer el que tot seguit expliquem:

1. Per cada valor diferent de la variable FEDTAX, només es va conservar un únic registre, de manera que el nou conjunt reduït de registres ja no tenia valors repetits de la variable FEDTAX. El registre conservat, per cada valor diferent de la variable FEDTAX, va ser el que tenia número de seqüència més baix.

2. Es va repetir el pas anterior per a les variables AFNLWGT, AGI, EMCONTRB, PTOTVAL, TAXINC, STATETAX. Després d'això, cap d'aquestes variables tenia valors repetits.

3. Per rebaixar el número de registres a menys de 1100, calia eliminar encara algunes repeticions més d'alguna altra variable. Es van ordenar els valors de la variable POTHVAL, i es van anar eliminant els seus valors repetits fins que van quedar 1080 registres.

Així, finalment, vam obtenir un conjunt de microdades amb les següents propietats:

- El número de registres era inferior a 1200.

- Hi havia set variables sense valors repetits: FEDTAX, AFNLWGT, AGI, EMCONTRB, PTOTVAL, TAXINC, STATETAX. La qual cosa és una característica general de les variables contínues.

- El número definitiu de registres, 1080, és l'enter més gran, inferior a 1200, que a la vegada és múltiple de 5, 8 i 9. Així, el conjunt de dades pot ser microagregat per a $k = 3, 4, 5, 6, 8, 9, 10, 12, 15, 18$ i 20, de manera que tots els grups tinguin la mateixa mida k.

6.3 Pèrdua d'informació: Mesures utilitzades

Seguint la nomenclatura de la secció 5.5, les mesures sobre la pèrdua d'informació que hem utilitzat en aquest estudi comparatiu han estat les següents:

PI1 : Variació mitjana de les discrepàncies entre les dades $\mathbf{X} - \mathbf{X}'$ multiplicada per 100.

PI2 : Variació mitjana de les discrepàncies entre mitjanes $\overline{\mathbf{X}} - \overline{\mathbf{X}}'$ multiplicada per 100.

PI3 : Variació mitjana de les discrepàncies entre covariàncies $\mathbf{V} - \mathbf{V}'$ multiplicada per 100.

PI4 : Variació mitjana de les discrepàncies entre variàncies $\mathbf{S} - \mathbf{S}'$ multiplicada per 100.

PI5 : Mitjana de l'error absolut de les discrepàncies entre correlacions $\mathbf{R} - \mathbf{R}'$ multiplicada per 100.

Hem preferit la variació mitjana en quatre de les anteriors mesures perquè no està afectada pels canvis d'escala de les variables. Tot i això, per mesurar les discrepàncies entre els coeficients de correlació hem triat la mitjana de l'error absolut per dues raons: les correlacions no estan afectades pels canvis d'escala de les variables; i, d'altra banda, la variació mitjana podria ser enormement gran quan els coeficients de correlació originals tinguessin valors molt propers a 0.

Finalment, per trobar una mesura que engloba tota la pèrdua d'informació **PI**, hem ponderat $PI1$, $PI2$, $PI3$, $PI4$ i $PI5$ d'acord amb les tres naturaleses descrites al capítol anterior:

1. Discrepància directa entre dades: $PI1$.
2. Discrepància entre estadístics univariants: $PI2$ i $PI4$.
3. Discrepància afectada per la relació entre cada dues variables: $PI3$ i $PI5$.

donant el mateix pes (1/3) a cada una d'elles.

Així doncs, resulta la següent mesura que engloba tota la pèrdua d'informació:

$$\mathbf{PI} = \frac{1}{3}\cdot PI1 + \frac{1}{6}\cdot PI2 + \frac{1}{6}\cdot PI4 + \frac{1}{6}\cdot PI3 + \frac{1}{6}\cdot PI5 =$$

$$= \frac{100}{3np}\sum_{i=1}^{n}\sum_{j=1}^{p}\frac{|x_{ij}-x'_{ij}|}{|x_{ij}|} + \frac{100}{6p}\sum_{j=1}^{p}\frac{|\overline{x}_j-\overline{x}'_j|}{|\overline{x}_j|} + \frac{2\cdot 100}{6p(p-1)}\sum_{i=1}^{p}\sum_{j=i+1}^{p}\frac{|v_{ij}-v'_{ij}|}{|v_{ij}|} +$$
$$+ \frac{100}{6p}\sum_{j=1}^{p}\frac{|v_j-v'_j|}{|v_j|} + \frac{2\cdot 100}{6p(p-1)}\sum_{i=1}^{p}\sum_{j=i+1}^{p}|r_{ij}-r'_{ij}|$$

6.4 Risc de revelació: Mesures utilitzades

Un mesurament empíric del risc de revelació s'ha fet mitjançant una aproximació a l'enllaç de registres descrita al capítol anterior: Enllaç de registres basat en distàncies (ERD).

Un segon mesurament empíric del risc de revelació s'ha fet mitjançant càlcul d'intervals de confidencialitat sobre el número de registres (ICN) i sobre la desviació típica (ICD), tal com ha estat descrit també al capítol anterior.

De les tretze variables utilitzades en aquest estudi, combinant les set variables sense valors repetits: FEDTAX, AFNLWGT, AGI, EMCONTRB, PTOTVAL, TAXINC i STATETAX, s'han considerat set escenaris de revelació per a l'aproximació a l'enllaç de registres, d'acord amb el número de variables comunes al conjunt extern i al conjunt modificat de dades:

1. *Una variable comuna.* Únicament la variable FEDTAX és comuna al conjunt extern i al conjunt modificat de dades.
2. *Dues variables comunes.* Les variables FEDTAX, AFNLWGT són comunes als dos conjunts de dades.
3. *Tres variables comunes.* Les variables FEDTAX, AFNLWGT, AGI són comunes als dos conjunts de dades.
4. *Quatre variables comunes.* Les variables FEDTAX, AFNLWGT, AGI, EMCONTRB són comunes als dos conjunts de dades.
5. *Cinc variables comunes.* Les variables FEDTAX, AFNLWGT, AGI, EMCONTRB, PTOTVAL són comunes als dos conjunts de dades.
6. *Sis variables comunes.* Les variables FEDTAX, AFNLWGT, AGI, EMCONTRB, PTOTVAL, TAXINC són comunes als dos conjunts de dades.
7. *Set variables comunes.* Les variables FEDTAX, AFNLWGT, AGI, EMCONTRB, PTOTVAL, TAXINC, STATETAX són comunes als dos conjunts de dades.

Aquests set escenaris de revelació es corresponen amb diferents graus de coneixement que un intrús pot tenir respecte les dades originals.

En el nostre estudi, les tres mesures sobre el risc de revelació, anteriorment referides, han estat calculades de la següent manera:

- La mesura sobre el risc de revelació ERD es correspon amb el percentatge de registres correctament enllaçats mitjançant l'enllaç de registres basat en distàncies. Aquesta mesura ERD, tal com apareix als resultats d'aquesta anàlisi comparativa, és la mitjana dels ERD- i's obtinguts per a cada un dels set anteriors escenaris de revelació (anomenem ERD-i, el ERD corresponent a l'i-èsim escenari de revelació).
- La mesura sobre el risc de revelació ICN, basada en intervals de confidencialitat sobre el número total de registres, s'ha calculat com el percentatge mitjà dels valors originals que es troben dintre dels intervals centrats en els seus corresponents valors modificats. Aquesta mitjana s'ha calculat sobre tot el rang de percentatges considerats (des de $p = 1\%$ fins a $p = 10\%$, amb increments d'1%).
- La mesura sobre el risc de revelació ICD, basada en intervals de confidencialitat sobre la desviació típica de cada variable, s'ha calculat com el percentatge mitjà dels valors originals que es troben dintre dels intervals centrats en els seus corresponents valors modificats. Aquesta mitjana s'ha calculat sobre tot el rang de percentatges considerats (Des de $p = 1\%$ fins a $p = 10\%$, amb increments d'1%).

Finalment, la mesura, **PC**, que engloba tota la pèrdua de confidencialitat, ha estat calculada considerant les dues naturaleses sobre el risc de revelació descrites al capítol anterior:

1. Pèrdua de confidencialitat basada en enllaç de registres: ERD.

2. Pèrdua de confidencialitat basada en intervals d'estimació: ICN i ICD.

i donant la mateixa ponderació a cada una d'elles:

$$\mathbf{PC} = 0.50 \cdot ERD + 0.25 \cdot ICN + 0.25 \cdot ICD$$

6.5 Resultats de l'estudi comparatiu i conclusions

6.5.1 Mesura global sobre la qualitat d'un mètode pertorbatiu

Creiem que una mesura global, representativa de la qualitat d'un mètode de control de la revelació estadística, ha de donar la mateixa importància a la pèrdua d'informació que al risc de revelació. Per això, la mesura global, **MG**, sobre la qualitat dels mètodes pertorbatius comparats en aquest estudi ha estat calculada de la següent manera:

$$\begin{aligned} \mathbf{MG} &= 0.50 \cdot \mathrm{PI} + 0.50 \cdot \mathrm{PC} = \\ &= 0.5 \cdot \mathrm{PI} + 0.25 \cdot \mathrm{ERD} + 0.125 \cdot \mathrm{ICN} + 0.125 \cdot \mathrm{ICD} \end{aligned} \tag{6.1}$$

Com podem observar a l'expressió (6.1), hem donat la mateixa ponderació a la pèrdua d'informació (0.5) que al risc de revelació (0.5). El pes 0.5 assignat al risc de revelació ha estat igualment repartit entre l'aproximació empírica per enllaç de registres (0.25), i les aproximacions empíriques per intervals de confidencialitat (0.25) ; tot i que aquest últim pes s'ha repartit també igualment entre els intervals de confidencialitat basats en el número de registres ICN (0.125) i els intervals de confidencialitat basats en la desviació típica ICD (0.125).

Els valors de tots els sumants que formen la mesura global **MG**, excepte la pèrdua d'informació (PI), pertanyen a l'interval $[0, 100]$. La pèrdua d'informació és un percentatge de variació que podria ser superior a 100. Tot i això, la següent regla pot ser útil per interpretar el resultat de la mesura global d'un determinat mètode pertorbatiu:

> No utilitzar cap mètode de control de la revelació, és a dir, publicar totes les dades originals sense cap tipus de modificació, donaria una mesura global igual a 50, perquè no hi hauria pèrdua d'informació, però el risc de revelació seria el màxim 100. Per la qual cosa, *un mètode de control de la revelació amb una puntuació de la mesura global superior a 50, no serà útil.*

6.5.2 Taules incloses a l'apèndix

Taula A.1

La Taula A.1 de l'apèndix és una taula resum que conté una ordenació dels mètodes comparats en aquest estudi. Per a cada mètode apareixen les següents mesures: PI, ERD, ICN, ICD i **MG**. Els mètodes estan ordenats d'acord amb el valor de la corresponent mesura global **MG**, en ordre ascendent (com més petit és el valor de la mesura global, el mètode té més qualitat).

A més a més, les columnes PI Rank, ERD Rank, ICN Rank i ICD Rank contenen la diferent posició que ocupa cada mètode segons les ordenacions respecte PI, ERD, ICN i ICD (com més petit és el número que representa la posició, més qualitat té el mètode per a la corresponent mesura).

A tota la resta de taules d'aquest apèndix, els mètodes apareixen sempre en el mateix ordre que a la Taula A.1, és a dir, ordenats per la mesura global.

Taula A.2

La Taula A.2 conté la pèrdua d'informació corresponent a cada mètode, explicitant les sis mesures següents: PI1, PI2, PI3, PI4, PI5 i PI, tal com han estat definides en aquest mateix capítol.

Taula A.3

Per a cada mètode, a la Taula A.3 apareix ERD-i per a $i = 1$ fins a 7. ERD-i és ERD quan l'intrús coneix un conjunt extern de dades que té i variables comunes amb el conjunt modificat de dades.

L'última columna de la Taula A.3 és la mitjana de les set columnes anteriors.

Taula A.4

Per a cada mètode, a la Taula A.4 apareix ICN-i per a $i = 1$ fins a 10. ICN-i és ICN quan $p = i\%$.

L'última columna de la Taula A.4 és la mitjana de les deu columnes anteriors.

Taula A.5

Per a cada mètode, a la Taula A.5 apareix ICD-i per a $i = 1$ fins a 10. ICD-i és ICD quan $p = i\%$.

L'última columna de la Taula A.5 és la mitjana de les deu columnes anteriors.

6.5.3 Estudi dels mètodes

De cada un dels mètodes pertorbatius de control de la revelació estadística considerats en aquesta experimentació detallarem, en aquest apartat, la pèrdua d'informació, el risc de revelació i la qualitat global, en funció dels paràmetres utilitzats.

Pertorbació additiva aleatòria (`Addt`p)

Pèrdua d'informació

En afegir una pertorbació aleatòria que segueix una distribució N(0, ps), on s és la desviació estandard de la variable original i p és un paràmetre amb valors 0.01, 0.02, 0.04, ..., fins a 0.2, amb increments de 0.02, la pèrdua d'informació, tal com apareix al gràfic 6.5.3, mostra una forta tendència creixent quan el paràmetre p augmenta.

Risc de revelació

La pèrdua de confidencialitat, representada per ERD, ICN i ICD, i com podem observar al mateix gràfic 6.5.3, presenta un clar decreixement en totes tres mesures quan el paràmetre p augmenta. Cal destacar que de les dues mesures ICN i ICD, que representen una mateixa naturalesa respecte el risc de revelació, l'ICD manifesta menys pèrdua de confidencialitat que l'ICN, per a un mateix valor del paràmetre p.

Qualitat del mètode

L'encreuament entre la creixent pèrdua d'informació i el decreixent risc de revelació (combinació de les mesures ERD, ICN i ICD) es produeix prop del valor del paràmetre $p = 0.14$. La mesura global sobre la qualitat del mètode, **MG**, es manté prou estable al voltant del valor 50 a tot el

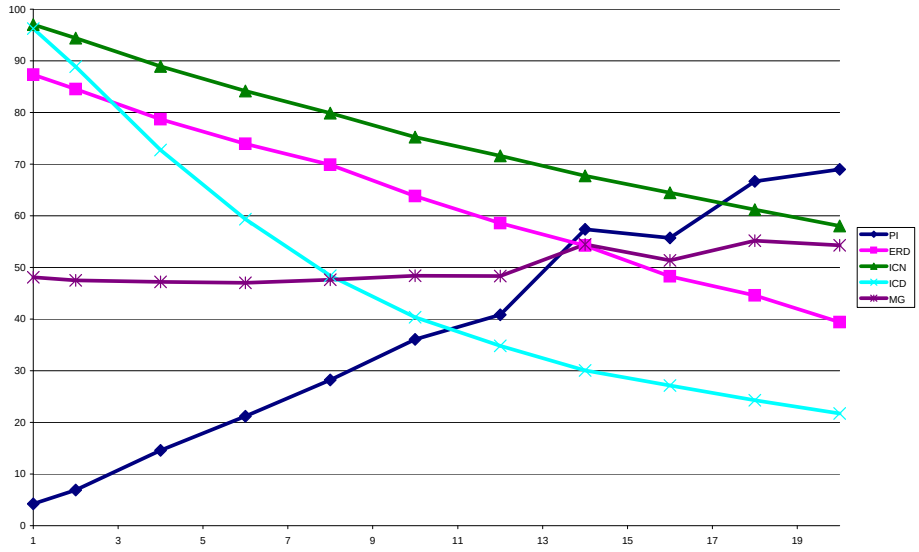


Figura 6.1: Mesures obtingudes en aplicar la pertorbació additiva aleatòria (Addt).

gràfic; per la qual cosa aquest mètode `Addt`p resulta tenir poca utilitat per controlar la revelació estadística.

Pertorbació de les dades segons una distribució de probabilitat (`Distr`)

Pèrdua d'informació

Com que el mètode de la pertorbació de les dades segons una distribució de probabilitat no té paràmetres, només apareix a la Taula A.1 un únic valor (47.6597); la qual cosa suposa aproximadament un 50% de pèrdua d'informació (combinació de les tres naturaleses anteriorment referides).

Risc de revelació

La pèrdua de confidencialitat per a aquest mètode ve representada també per un únic valor de cada una de les tres següents mesures: ERD (71.944), ICN (87.3419) i ICD (59.1838). Respecte les dues mesures ICN i ICD que pertanyen a una mateixa naturalesa sobre risc de revelació cal destacar el valor més petit d'ICD.

Qualitat del mètode

Aquest mètode, `Distr`, presenta una mesura global sobre la qualitat per sobre de 50; per la qual cosa resulta ser un mètode sense cap utilitat per controlar la revelació de dades.

Remostreig (`Remost`)

Pèrdua d'informació

La pèrdua d'informació del mètode de Remostreig ha estat calculada per a dos valors del seu paràmetre $t = 1$ i $t = 3$. Observem a la Taula A.1 un valor lleugerament més petit de la pèrdua d'informació corresponent al valor del paràmetre $t = 3$; tot i que tots dos valors són prou petits (2.9666 i 3.1620 respectivament).

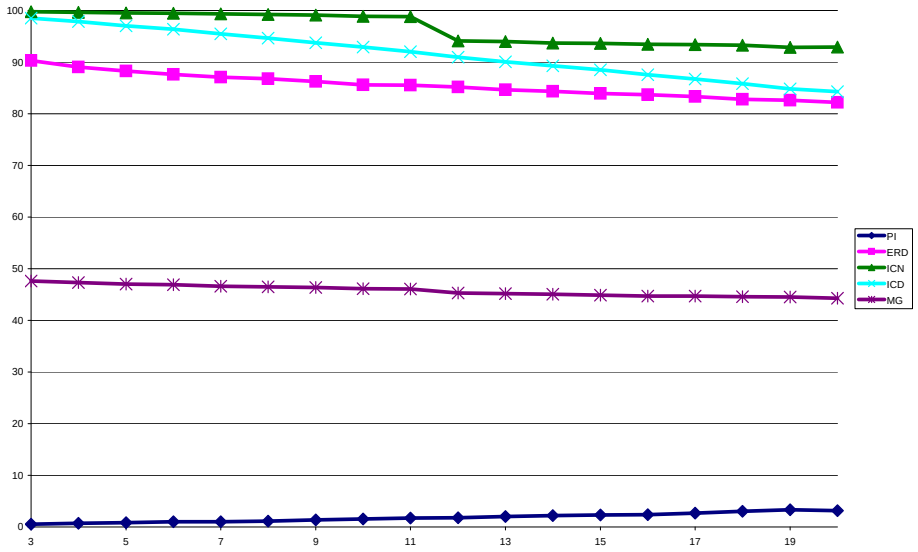


Figura 6.2: Mesures obtingudes en aplicar la microagegació amb ordenació individual (MicOI).

Risc de revelació

Respecte la pèrdua de confidencialitat d'aquest mètode, cal comentar que tot i tenir valors molt alts (per sobre d'un 78% en totes tres mesures i en tots dos valors del paràmetre), podem observar a la Taula A.1 un lleuger augment del risc de revelació per al valor del paràmetre $t = 3$ en totes tres mesures (ERD, ICN i ICD).

Qualitat del mètode

Una bona característica d'aquest mètode de Remostreig com és la seva baixa pèrdua d'informació, queda molt anul·lada per un alt risc de revelació, de manera que la mesura global sobre la seva qualitat per controlar la revelació estadística es troba per sota, però molt propera, al 50% (el millor valor del paràmetre és $t = 1$ amb **MG**= 43.11); per la qual cosa es tracta d'un mètode amb poca efectivitat en aquest context.

Microagregació amb ordenació individual (`MicOI`k)

Pèrdua d'informació

Com podem observar al gràfic 6.5.3, la pèrdua d'informació del mètode de microagregació multivariant amb ordenació individual és extremadament baixa (creixent des d'un 0.54% fins a un 3.15% mentre el paràmetre k augmenta prenent valors enters entre 3 i 20).

Risc de revelació

La pèrdua de confidencialitat d'aquest mètode, contràriament a la seva pèrdua d'informació i com podem observar al gràfic 6.5.3, es manté a tot el gràfic i per a totes tres mesures (ERD, ICN i ICD) molt propera als seus valors màxims (100%); tot i que disminueix molt sensiblement quan el paràmetre k augmenta prenent valors enters entre 3 i 20.

Qualitat del mètode

Respecte la qualitat global del mètode de microagregació multivariant amb ordenació individual, hem de destacar que una molt bona característica com és la seva extraordinàriament baixa

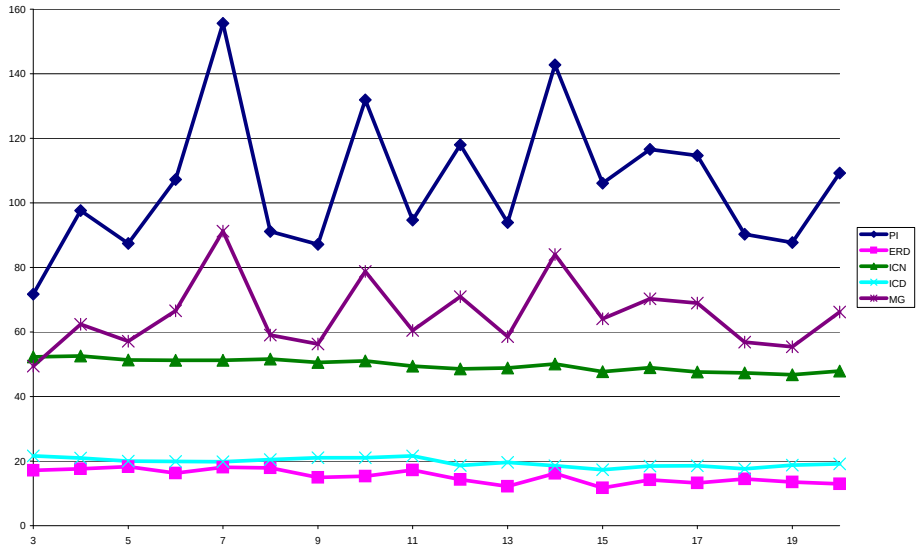


Figura 6.3: Mesures obtingudes en aplicar la microagegació per projecció sobre la suma de les z-puntuacions (MicZ).

pèrdua d'informació, queda pràcticament anul·lada per un risc de revelació molt proper al seu valor màxim (100%). Com a conseqüència d'aquesta interacció contrària entre les dues pèrdues, resulta una mesura global, **MG**, propera a 50; la qual cosa el converteix en un mètode molt poc aconsellable per controlar la revelació estadística.

Microagregació mitjançant les z-puntuacions (`MicZ`k)

Pèrdua d'informació

Com podem observar al gràfic 6.5.3, la pèrdua d'informació de la microagregació multivariant mitjançant les z-puntuacions es manté molt elevada oscil·lant entre un 71% de valor mínim, fins a un quasibé 156%, mentre el paràmetre k va agafant valors enters entre 3 i 20.

Risc de revelació

El risc de revelació d'aquest mètode, com podem veure a la Taula A.1 a través de les tres mesures ERD, ICN i ICD, es manté globalment prou baix; destacant els valors especialment més baixos d'ERD i ICD.

Qualitat del mètode

Respecte la qualitat global d'aquest mètode `MicZ`k , cal dir que la seva extraordinàriament elevada pèrdua d'informació anul·la totalment el seu baix risc de revelació, resultant una mesura global que oscil·la per sobre de 50 (el millor valor del seu paràmetre és $k = 3$ amb **MG**$= 49.40$); la qual cosa el converteix, segons el nostre estudi, en un mètode amb molt poca utilitat per controlar la revelació de dades estadístiques.

Microagregació per projecció sobre la primera component principal (`MicPCP`k)

Pèrdua d'informació

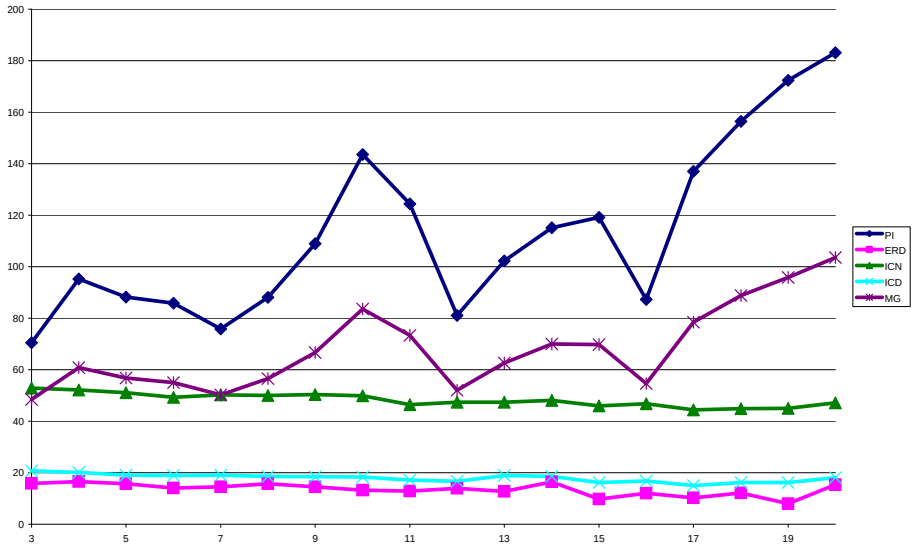


Figura 6.4: Mesures obtingudes en aplicar la microagegació per projecte sobre la primera component principal (MicPCP).

Com podem observar al gràfic 6.5.3, amb el seu comportament aquest mètode `MicPCP`k és molt semblant a `MicZ`k que acabem de comentar. Així doncs, la microagregació mitjançant projecció sobre la primera component principal presenta també una elevadíssima pèrdua d'informació que oscil·la entre un 70.5% de valor mínim (per a $k = 3$), fins un 183.14% de valor màxim (per a $k = 20$).

Risc de revelació

La pèrdua de confidencialitat d'aquest mètode és manté globalment petita; remarcant els valors especialment baixos d'ERD i ICD.

Qualitat del mètode

Pel que fa a la qualitat de `MicPCP`k, observant la Taula A.1 i el gràfic 6.5.3, podem comprovar que la seva mesura global **MG** oscil·la quasibé sempre per sobre de 50 (el valor més petit **MG**= 48.39 s'assoleix per a $k = 3$). Això fa que `MicPCP`k, similarment a `MicZ`k, sigui una opció molt poc útil per controlar la revelació estadística.

Microagregació multivariant amb grups de variables de mida 2 (`Mic2mul`k)

Pèrdua d'informació

La pèrdua d'informació del mètode `Mic2mul`k, com podem observar a la Taula A.1, és molt baixa per a valors petits del paràmetre k i creix lleugerament en augmentar el valor enter del paràmetre, però sense sobrepassar en cap moment el 30%.

Risc de revelació

Com mostra el gràfic 6.5.3, el risc de revelació d'aquesta variant de microagregació multivariant és alt en totes tres mesures ERD, ICN i ICD per a valors petits del paràmetre, i disminueix lleugerament quan el paràmetre k va agafant valors més grans. Cal remarcar també els valors

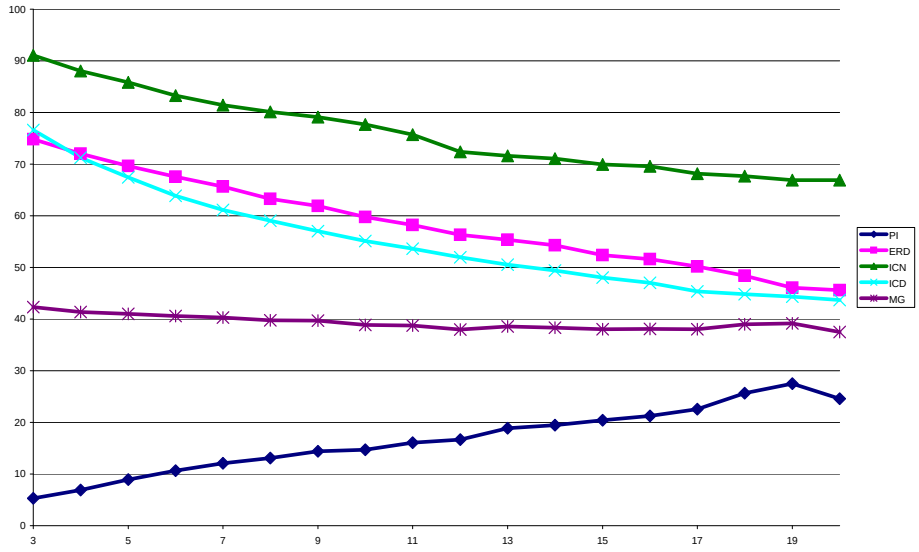


Figura 6.5: Mesures obtingudes en aplicar la microagegació multivariant amb grups de variables de mida 2 (Mic2mul).

més petits d'ICD respecte d'ICN (dues mesures que pertanyen a una mateixa naturalesa pel que fa al risc de revelació).

Qualitat del mètode

Respecte la qualitat global del mètode `Mic2mul`k, podem comprovar a la Taula A.1 que la seva mesura global **MG** es manté molt estable entre 30 i 40. La millor opció d'aquest mètode sembla ser un valor gran del paràmetre k, que produeix riscos de revelació més petits tot i conservant una moderada pèrdua d'informació (per a $k = 20$ tenim una **MG**= 37.51).

Microagregació multivariant amb grups de variables de mida 3 (`Mic3mul`k)

Pèrdua d'informació

La pèrdua d'informació del mètode `Mic3mul`k, com podem observar al gràfic 6.5.3, és molt baixa per a valors petits del paràmetre k i creix lleugerament en augmentar el valor enter del paràmetre, però sense sobrepassar en cap moment el 37%.

Risc de revelació

El risc de revelació d'aquesta variant de microagregació multivariant, com mostra el gràfic 6, és globalment alt per a valors petits del paràmetre, tot disminuint lleugerament a mesura que creix el valor enter del paràmetre k. Cal considerar també els valors més petits d'ICD respecte ICN (dues mesures que pertanyen a una mateixa naturalesa pel que fa al risc de revelació). A més a més, ERD és prou baix per a la majoria de valors grans del paràmetre.

Qualitat del mètode

En el gràfic d'aquest mètode observem que entre els valors del paràmetre $k = 15$ i $k = 20$, la creixent pèrdua d'informació s'encreua amb la decreixent ERD. La mesura global **MG** d'aquest mètode es manté molt estable entre 33 i 38. La millor opció pel que fa al valor del paràmetre

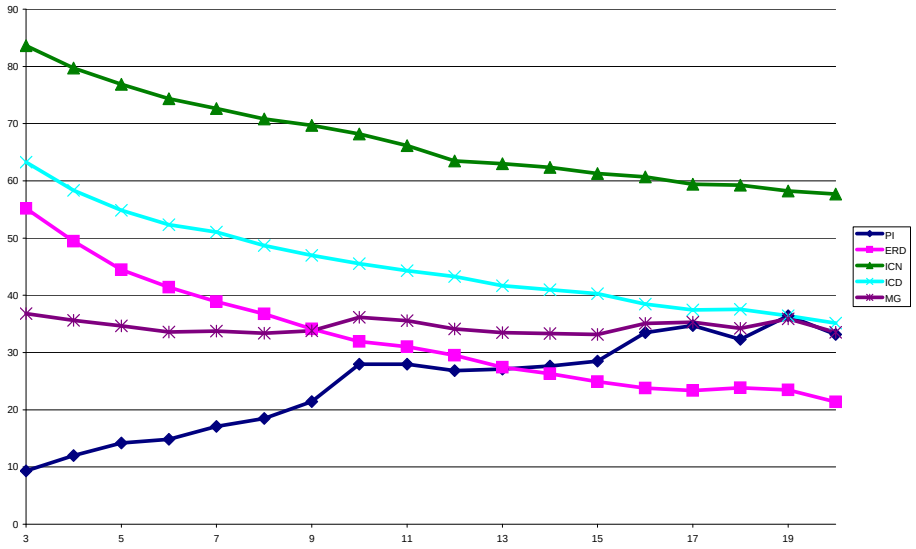


Figura 6.6: Mesures obtingudes en aplicar la microagegació multivariant amb grups de variables de mida 3 (Mic3mul).

k seria $k = 15$ amb una **MG**= 33.18 (zona d'encreuament abans comentada), que equilibraria molt bé els valors petits del risc de revelació amb la creixent pèrdua d'informació.

Microagregació multivariant amb grups de variables de mida 4 (`Mic4mul`k)

Pèrdua d'informació

Tal com apareix al gràfic 6.5.3, el comportament de la pèrdua d'informació del mètode `Mic4mul`k és similar al del mètode `Mic3mul`k darrerament comentat; tot i que, com pot comprovar-se a la Taula A.1, presenta valors moderadament més alts que `Mic3mul`k.

Risc de revelació

Similarment també a `Mic3mul`k, el risc de revelació de `Mic4mul`k és més alt, en totes tres mesures ERD, ICN i ICD, per als valors petits del paràmetre k; tot disminuint a mesura que creix el valor enter del paràmetre. Cal destacar també els valors més petits d'ICD respecte d'ICN; i els valors baixos d'ERD. Globalment, els valors de les tres mesures sobre el risc de revelació a `Mic4mul`k són més petites que a `Mic3mul`k.

Qualitat del mètode

De manera també semblant a `Mic3mul`k, el mètode `Mic4mul`k ofereix una mesura global sobre la qualitat prou estable entre 32 i 38. La millor opció pel que fa al valor del paràmetre k seria $k = 17$ amb una **MG**= 32.41, que equilibraria també de la millor manera els valors petits del risc de revelació amb els valors creixents de la pèrdua d'informació.

Microagregació multivariant amb grups de variables de mida 5 (`Mic5mul`k)

Pèrdua d'informació

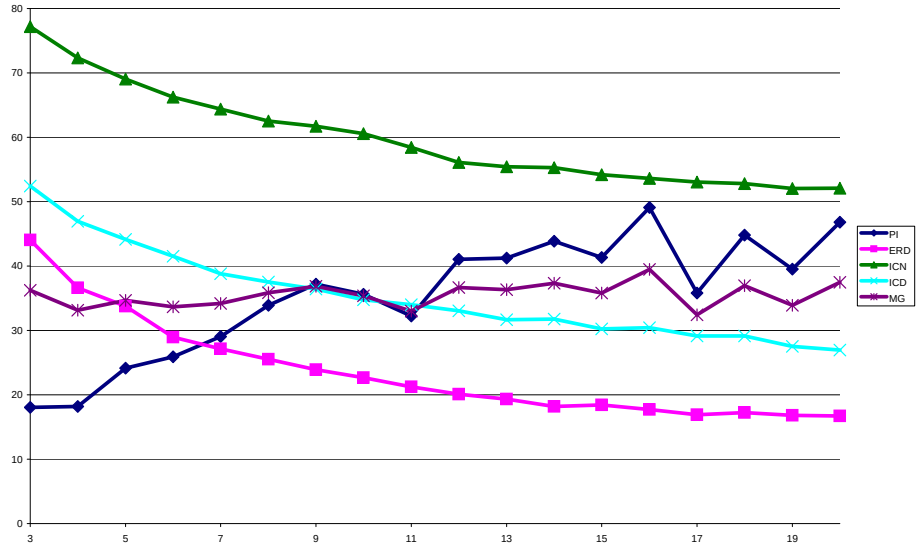


Figura 6.7: Mesures obtingudes en aplicar la microagegació multivariant amb grups de variables de mida 4 (Mic4mul).

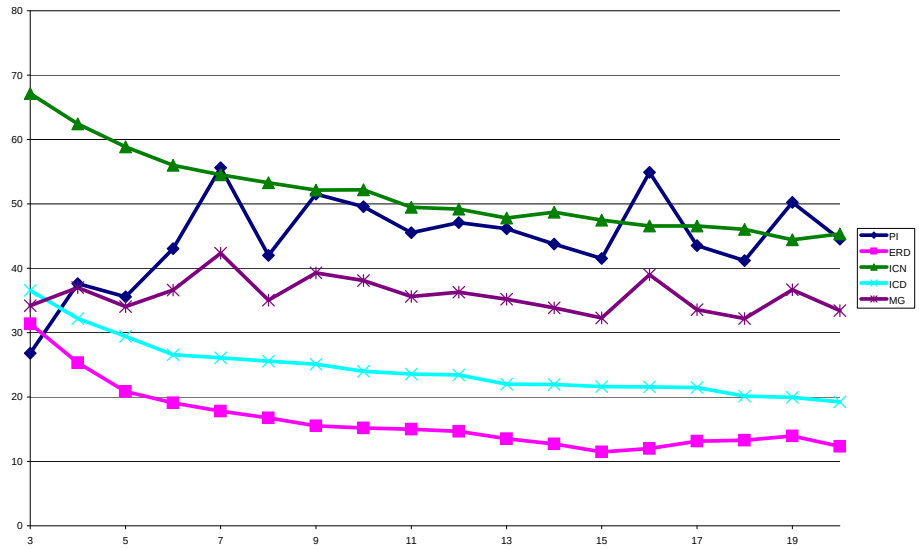


Figura 6.8: Mesures obtingudes en aplicar la microagegació multivariant amb grups de variables de mida 5 (Mic5mul).

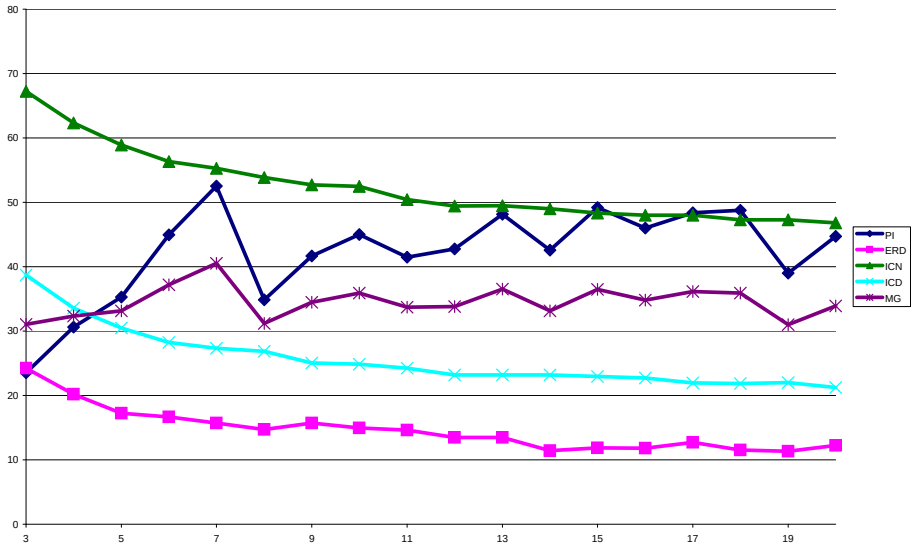


Figura 6.9: Mesures obtingudes en aplicar la microagegació multivariant amb grups de variables de mida 6 (Mic6mul).

Com mostra el gràfic 6.5.3, la pèrdua d'informació en el mètode de microagregació multivariant `Mic5mul`k té valors més alts que en `Mic4mul`k, amb una disminució de la tendència creixent i un augment de les oscil·lacions així que el paràmetre k varia entre 3 i 20.

Risc de revelació

El risc de revelació, tot i presentar la mateixa tendència decreixent que a `Mic4mul`k quan els valors enters del paràmetre k augmenten, es manté més baix que a `Mic4mul`k en les tres mesures ERD, ICN i ICD.

Qualitat del mètode

El comportament de la mesura global **MG** sobre la qualitat del mètode `Mic5mul`k és molt semblant al de `Mic3mul`k i `Mic4mul`k, amb un lleuger creixement de la tendència oscil·lativa, però mantenint-se bàsicament entre 30 i 40. La millor opció es trobaria per al valor del paràmetre $k = 18$ amb una **MG**= 32.18 (molt poca variació respecte `Mic3mul`k i `Mic4mul`k).

Microagregació multivariant amb grups de variables de mida 6 (`Mic6mul`k)

Pèrdua d'informació

Ben bé observant els gràfics 6.5.3 i 6.5.3, notem molt poca diferència entre la pèrdua d'informació a `Mic5mul`k i a `Mic6mul`k respectivament, tot i que a `Mic6mul`k apareix, potser, un molt lleuger creixement de la inestabilitat (major tendència oscil·lativa).

Risc de revelació

Considerant els mateixos dos gràfics 6.5.3 i 6.5.3, també veiem un comportament pràcticament idèntic del risc de revelació a `Mic5mul`k i a `Mic6mul`k respectivament; tot i destacar una lleugera disminució i major estabilitat de l'ERD a `Mic6mul`k.

Qualitat del mètode

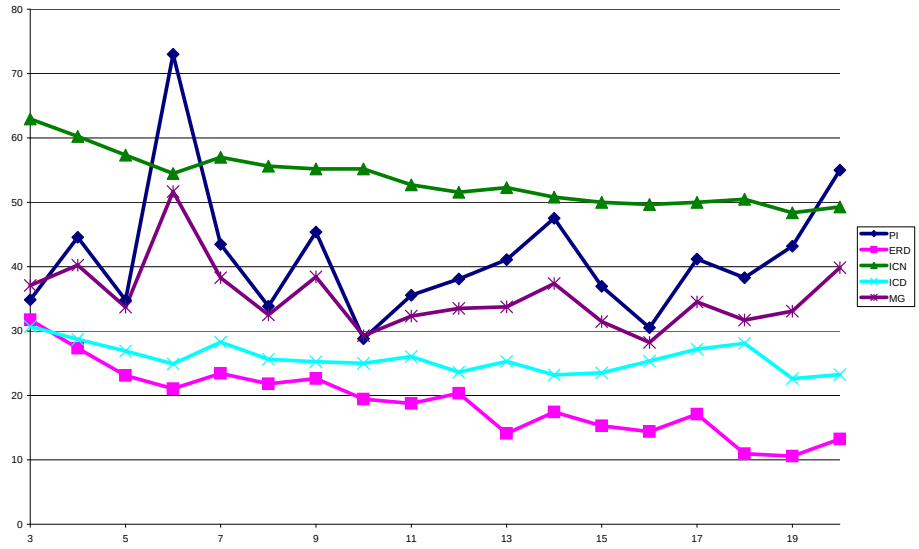


Figura 6.10: Mesures obtingudes en aplicar la microagegació multivariant amb un sol grup de totes les variables (Micmul).

Respecte la qualitat del mètode `Mic6mul`k, ens trobem novament amb una mesura global **MG** bàsicament centrada entre 30 i 40, pràcticament el mateix interval de variació que a `Mic2mul`k, `Mic3mul`k, `Mic4mul`k i `Mic5mul`k, però amb un lleuger creixement de la tendència oscil·lativa a través d'aquests mètodes, la qual ha tingut una evident conseqüència beneficiosa respecte la mesura global **MG** dels mateixos, que ha anat disminuint fins arribar en el mètode `Mic6mul`k a **MG**$= 30.99$ (per a $k = 19$).

Microagregació multivariant amb un sol grup de totes les variables (`Micmul`k)

Pèrdua d'informació

Respecte a la pèrdua d'informació de `Mic5mul`k i de `Mic6mul`k, podem comprovar al gràfic 6.5.3 que la pèrdua d'informació a `Micmul`k manifesta valors més alts i major grau d'inestabilitat (oscil·lacions més grans i més freqüents).

Risc de revelació

Pel que fa al risc de revelació, no notem tanta diferència respecte `Mic5mul`k i `Mic6mul`k; si bé apareix a `Micmul`k una lleugera disminució d'ICD i un petit increment d'inestabilitat en totes tres mesures ERD, ICN i ICD.

Qualitat del mètode

La qualitat del mètode `Micmul`k sí mostra una diferència beneficiosa respecte `Mic6mul`k, puix que el valor mínim de la seva mesura global **MG** ha disminuït fins a 28.24 per al valor del paràmetre $k = 16$. Cal destacar també una marcada inestabilitat de la mesura global **MG** d'aquest mètode `Micmul`k, tot i mantenir-se bàsicament entre 30 i 40.

Pèrdua per compressió (`JPEG`q)

Pèrdua d'informació

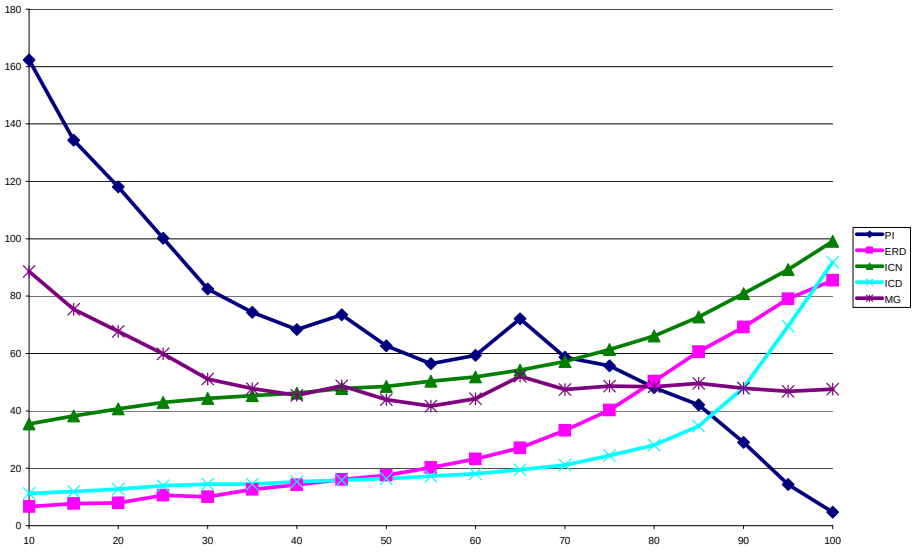


Figura 6.11: Mesures obtingudes en aplicar la pèrdua per compressió (JPEG).

El mètode pertorbatiu JPEGq, com podem observar al gràfic 6.5.3, manifesta una pèrdua d'informació molt elevada per a valors petits del seu paràmetre q (recordem que un valor petit del paràmetre q representa molta compressió de les dades originals). Tot i això, la pèrdua d'informació experimenta un ràpid decreixement quan augmenta el valor del paràmetre.

Risc de revelació

Pel que fa al risc de revelació, observem al mateix gràfic 6.5.3 que les tres mesures ERD, ICN i ICD de pèrdua de confidencialitat tendeixen a augmentar quan creix el valor del paràmetre q. Cal destacar que les tres mesures creixen molt més ràpidament a partir de valors del paràmetre compresos entre 60 i 70.

Qualitat del mètode

Tal com mostra el gràfic 6.5.3, la mesura global **MG** sobre la qualitat del mètode JPEGq presenta inicialment (valors petits del paràmetre) una tendència decreixent fins que assoleix el seu valor mínim en el valor del paràmetre $q = 55\%$ (amb una **MG**$= 41.71$), que millor equilibra la decreixent pèrdua d'informació amb el creixent risc de revelació, tot i que al voltant de $q = 80\%$ es produeixin encreuaments entre totes dues pèrdues. D'altra banda, a partir del valor $q = 70\%$, la mesura global **MG** manifesta una certa estabilitat.

Intercanvi de dades (Rank swapping)(Rank$p$)

Pèrdua d'informació

La pèrdua d'informació del mètode d'intercanvi de dades, com podem observar a la Taula A.1 i al gràfic 6.5.3, presenta una clara tendència creixent a mesura que augmenta el paràmetre p, però mantenint-se a tot el gràfic sota el 54%.

Risc de revelació

El risc de revelació d'aquest mètode per a valors petits del seu paràmetre p és molt gran, però marcadament decreixent quan el paràmetre augmenta. Aquest decreixement és molt ràpid des

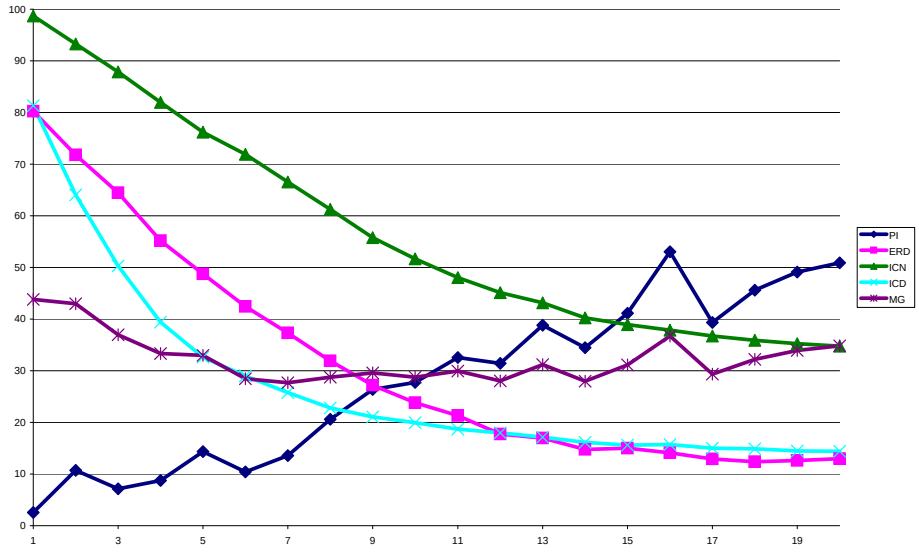


Figura 6.12: Mesures obtingudes en aplicar l'intercanvi de dades (Rank).

dels valors més petits del paràmetre fins a valors de p entre 7 i 10; després, per a valors de p entre 10 i 20, el decreixement és molt més lleuger.

Qualitat del mètode

Respecte la qualitat del mètode Rank swapping, hem de dir primerament que és el mètode amb la millor mesura global sobre la qualitat **MG** (**MG**= 27.65 per al valor del paràmetre $p = 7$) de tots els analitzats en aquest estudi comparatiu. Al gràfic 6.5.3 podem observar un encreuament entre la creixent pèrdua d'informació i les decreixents mesures del risc de revelació ERD i ICD, justament al voltant del valor $p = 7$, millor valor del paràmetre d'aquest mètode pel que fa a la seva qualitat per controlar la revelació estadística.

6.5.4 Conclusions

De tot aquest estudi comparatiu es poden treure conclusions des de dos punts de vista diferents:

1. Comparant els diversos mètodes i variants entre ells mateixos.
2. Buscant, per a un mètode determinat, el valor del paràmetre o la combinació de paràmetres que optimitza la seva qualitat.

Podríem resumir les conclusions més destacables d'aquest estudi comparatiu de la següent manera:

- El mètode d'Intercanvi de dades (*Rank swapping*) és el que té la millor puntuació (més petita) de la mesura global **MG** sobre la qualitat d'un mètode de control de la revelació estadística. En segon lloc es troba la microagregació multivariant considerant totes les variables simultàniament; encara que amb molt poca diferència de puntuació respecte del rank swapping (27.65 per al millor `rank`p i 28.24 per a la millor `Micmul`k).

  Tot i que el rank swapping té una puntuació lleugerament millor, `Micmul`k té l'avantatge sobre el `rank`p de ser un mètode determinista i, per tant, reproductible sense pèrdua de seguretat.

Aquesta característica és important en un context de bases de dades estadístiques "on-line", perquè successives modificacions de les mateixes dades originals, mitjançant `rank`p o qualsevol altre mètode estocàstic, produïrien diferents conjunts modificats de dades, que podrien conduir a la revelació de dades.

- Després de `rank`p i de `Micmul`k, els altres mètodes ordenats segons la seva bona qualitat global serien: `Mic6mul`k, `Mic5mul`k, `Mic4mul`k, `Mic3mul`k, `Mic2mul`k, `JPEG`q, `Remost`t, `MicOI`k, `Addt`p, `MicPCP`k, `MicZ`k i `Distr`.

- Rank swapping: el valor del paràmetre $p = 7$, per al que aquest mètode obté el millor resultat, es troba al voltant de l'encreuament entre la creixent pèrdua d'informació i les decreixents mesures del risc de revelació ERD i ICD.

- Mètodes de microagregació multivariant sense projectar les dades: En aquests mètodes, a mesura que augmenta la fragmentació del conjunt de les variables (cardinal més petit dels conjunts que formen la partició sobre la qual s'aplica l'algorisme DMM), disminueix la pèrdua d'informació. El risc de revelació d'aquests mateixos mètodes manifesta una tendència contrària, puix que en augmentar la fragmentació del conjunt de les variables, augmenta també el risc de revelació (augment especialment manifestat per ERD i ICD).

- Els mètodes `MicZ`k i `MicPCP`k tenen un comportament semblant: obtenen bons resultats sobre la qualitat per a valors molt petits del paràmetre k. Aquests mètodes produeixen una pèrdua d'informació alta.

- La microagregació amb ordenació individual és molt estable respecte dels canvis del paràmetre k, però té una pèrdua de confidencialitat molt alta.

Capítol 7

DMM. Particions del conjunt de variables: nombre de variables

7.1 Microagregació multivariant i nombre de variables

Com ja havíem comentat al capítol 4, els mètodes de microagregació multivariant sense projecció de les dades, si bé disminueixen molt considerablement la pèrdua d'informació, tenen una gran dificultat a l'hora d'elegir el criteri per fer els agrupaments que optimitzi una mesura global sobre la qualitat d'un mètode pertorbatiu que considera tant la pèrdua d'informació com la pèrdua de confidencialitat provocades per la seva aplicació.

En aquest context, els mètodes DM (Distància Màxima) i DMM (Distància Màxima Modificat) per fer microagregació multivariant sense projecció de les dades, tindrien els següents dos casos extrems com a criteris per agrupar registres: microagregació amb ordenació individual (microagregant una sola variable cada cop), i microagregació amb totes les variables simultàniament. Entre aquests dos casos extrems existeixen moltes altres possibilitats. Aquest capítol presenta precisament el treball realitzat per determinar el nombre de variables que ha de tenir cada conjunt de la partició que servirà per executar el mètode DMM de la Distància Màxima Modificat i obtenir una bona mesura global de la qualitat.

Per aconseguir aquest objectiu hem treballat amb els 1080 registres individuals i les 13 variables AFNLWGT, AGI, EMCONTRB, ERNVAL, FEDTAX, FICA, INTVAL, PEARNVAL, POTHVAL, PTOTVAL, STATETAX, TAXINC, WSALVAL, referides al capítol 6 d'aquesta memòria.

Per a tres valors del paràmetre k: $k = 3$, $k = 10$ i $k = 20$, s'ha considerat i comparat la qualitat de les següents variants de la microagregació multivariant sense projecció de les dades:

- `Mic1-12mul`: Un grup amb 1 variable i un altre grup amb les restants 12 variables.
- `Mic2-11mul`: Un grup amb 2 variables i un altre grup amb les restants 11 variables.
- `Mic3-10mul`: Un grup amb 3 variables i un altre grup amb les restants 10 variables.
- `Mic4-9mul`: Un grup amb 4 variables i un altre grup amb les restants 9 variables.
- `Mic5-8mul`: Un grup amb 5 variables i un altre grup amb les restants 8 variables.
- `Mic6-7mul`: Un grup amb 6 variables i un altre grup amb les restants 7 variables.

- `Mic4mul`: Dos grups amb 4 variables cadascun i un tercer grup amb les 5 restants variables.
- `Mic3mul`: Tres grups amb 3 variables cadascun i un quart grup amb les 4 restants variables.
- `Mic2mul`: Cinc grups amb 2 variables cadascun i un sisè grup amb les 3 restants variables.
- `Micmul`: Un sol grup amb totes 13 variables.

Observem que les variants `Mic4mul`, `Mic3mul` i `Mic2mul` han estat variants-extensió de `Mic4-9mul`, `Mic3-10mul` i `Mic2-11mul` respectivament, per valorar l'efecte sobre la qualitat de les possibles reproduccions de grups de variables de mida 4, 3 i 2 respectivament.

A la següent secció desenvoluparem maneres de comptar quantes possibles particions del conjunt original de variables, $\mathbf{V}$, constituït per les p variables observades, existeixen, formades per $h-1$ conjunts amb s elements cadascun i un únic conjunt amb $s+r$ elements (essent $p = h \cdot s + r$).

7.2 Particions del conjunt V de variables

Considerem un conjunt de microdades amb p variables mètriques i n vectors de dades (és a dir, el resultat d'observar p variables en n individus). Així doncs, un vector de dades particular es pot veure com una fila $\mathbf{X} = (x_1, x_2, \ldots, x_p)$.

Sigui s un nombre natural fixat tal que $s \leq p$. Dividint p entre s, tindríem:

$$p = hs + r \tag{7.1}$$

essent $0 \leq r < s$.

Si $\mathbf{V}$ és el conjunt constituït per les p variables observades, i $\mathcal{F}(\mathbf{V})$ és el conjunt de totes les possibles particions de $\mathbf{V}$ formades per $h-1$ conjunts amb s elements cadascun i un únic conjunt amb $s+r$ elements (en el cas que $r = 0$ totes les particions tindrien h grups, tots de mida s), llavors deduïrem el número d'elements μ del conjunt $\mathcal{F}(\mathbf{V})$ de la següent manera:

1r.Cas: $\mathbf{r = 0}$ Si el residu r de la igualtat (4.1) és igual a zero, llavors totes les particions considerades tenen h grups de mida s, i obtenim:

$$\mu = \binom{p-1}{s-1} \cdot \binom{p-s-1}{s-1} \cdot \binom{p-2s-1}{s-1} \cdots \binom{p-(h-2)s-1}{s-1} \tag{7.2}$$

puix que, si considerem numerades les variables del conjunt $\mathbf{V}$ des d'1 fins a p, podem escollir de $\binom{p-1}{s-1}$ maneres diferents les $s-1$ variables que estaran en el mateix grup que la primera variable. A més a més, si considerem, per a cadascuna de les anteriors $\binom{p-1}{s-1}$ maneres diferents d'escollir, la primera variable que no ha estat asignada al primer grup de s variables; llavors podem escollir, ara, de $\binom{p-s-1}{s-1}$ maneres diferents, les $s-1$ variables que estaran en el mateix grup que la abans anomenada primera variable no asignada al primer grup. Semblantment al pas anterior, en el conjunt $\mathbf{V}$ de variables numerades, tornaríem a buscar la primera variable que no ha estat asignada encara a cap dels dos grups ja formats; i escolliríem, ara, de $\binom{p-2s-1}{s-1}$ maneres diferents, les $s-1$ variables que estaran en el mateix grup que ella. I així successivament ....

2n.Cas: r $\neq$ 0 Si el residu r de la igualtat (4.1) és diferent de zero, llavors totes les particions considerades tenen $h-1$ grups de mida s, i un únic grup de mida $s+r$. D'aquesta manera obtenim:

$$\mu = \left(\begin{array}{c} p \\ s+r \end{array} \right) \cdot \left(\begin{array}{c} p-s-r-1 \\ s-1 \end{array} \right) \cdot \left(\begin{array}{c} p-2s-r-1 \\ s-1 \end{array} \right) \cdots \left(\begin{array}{c} p-(h-2)s-r-1 \\ s-1 \end{array} \right) \quad (7.3)$$

puix que, escollint, primerament, les $s+r$ variables que formaran el grup de mida diferent dels altres, tindríem $\left(\begin{array}{c} p \\ s+r \end{array} \right)$ possibilitats d'elecció diferents. Després, distribuïríem les restants variables en $h-1$ grups tots de mida s, de la mateixa manera que hem fet en l'anterior cas $r=0$.

Podem simplificar les expressions (7.2) i (7.3) de la següent manera:

1r.Cas: r = 0 Si el residu r de la igualtat (4.1) és igual a zero, operem d'aquesta manera:

$$\mu = \left(\begin{array}{c} p-1 \\ s-1 \end{array} \right) \cdot \left(\begin{array}{c} p-s-1 \\ s-1 \end{array} \right) \cdot \left(\begin{array}{c} p-2s-1 \\ s-1 \end{array} \right) \cdots \left(\begin{array}{c} p-(h-2)s-1 \\ s-1 \end{array} \right) =$$

$$= \frac{(p-1)(p-2)\cdots(p-s+1)}{(s-1)!} \cdot \frac{(p-s-1)(p-s-2)\cdots(p-2s+1)}{(s-1)!} \cdot$$

$$\cdot \frac{(p-2s-1)(p-2s-2)\cdots(p-3s+1)}{(s-1)!} \cdots$$

$$\cdots \frac{[p-(h-2)s-1][p-(h-2)s-2]\cdots[p-(h-1)s+1]}{(s-1)!} =$$

$$= \frac{p(p-s)(p-2s)\cdots[p-(h-1)s]!}{p(p-s)(p-2s)\cdots[p-(h-1)s]!} \cdot \frac{(p-1)(p-2)\cdots(p-s+1)}{(s-1)!} \cdot$$

$$\cdot \frac{(p-s-1)(p-s-2)\cdots(p-2s+1)}{(s-1)!} \cdot \frac{(p-2s-1)(p-2s-2)\cdots(p-3s+1)}{(s-1)!} \cdots$$

$$\cdots \frac{[p-(h-2)s-1][p-(h-2)s-2]\cdots[p-(h-1)s+1]}{(s-1)!} =$$

$$= \frac{p!}{hs \cdot (h-1)s \cdot (h-2)s \cdots 3s \cdot 2s \cdot s \cdot (s-1)! \cdot [(s-1)!]^{h-1}} =$$

$$= \frac{p!}{h! \cdot (s!)^h}$$

puix que $p = hs$ en aquest cas.

2n.Cas: r $\neq$ 0 Si el residu r de la igualtat (4.1) és diferent de zero, operem d'aquesta altra manera:

$$\mu = \left(\begin{array}{c} p \\ s+r \end{array} \right) \cdot \left(\begin{array}{c} p-s-r-1 \\ s-1 \end{array} \right) \cdot \left(\begin{array}{c} p-2s-r-1 \\ s-1 \end{array} \right) \cdots \left(\begin{array}{c} p-(h-2)s-r-1 \\ s-1 \end{array} \right) =$$

$$= \frac{p(p-1)(p-2)\cdots(p-s-r+1)}{(s+r)!} \cdot \frac{(p-s-r-1)(p-s-r-2)\cdots(p-2s-r+1)}{(s-1)!} \cdot$$

$$\cdot \frac{(p-2s-r-1)(p-2s-r-2)\cdots(p-3s-r+1)}{(s-1)!} \cdots$$

$$\cdots \frac{[p-(h-2)s-r-1][p-(h-2)s-r-2]\cdots[p-(h-1)s-r+1]}{(s-1)!} =$$

$$= \frac{(p-s-r)(p-2s-r)\cdots[p-(h-1)s-r]!}{(p-s-r)(p-2s-r)\cdots[p-(h-1)s-r]!} \cdot \frac{p(p-1)(p-2)\cdots(p-s-r+1)}{(s+r)!} \cdot$$

$$\cdot \frac{(p-s-r-1)(p-s-r-2)\cdots(p-2s-r+1)}{(s-1)!} \cdot$$

$$\cdot \frac{(p-2s-r-1)(p-2s-r-2)\cdots(p-3s-r+1)}{(s-1)!} \cdots$$

$$\cdots \frac{[p-(h-2)s-r-1][p-(h-2)s-r-2]\cdots[p-(h-1)s-r+1]}{(s-1)!} =$$

$$= \frac{p!}{(h-1)s\cdot(h-2)s\cdots 3s\cdot 2s\cdot s\cdot(s-1)!\cdot(s+r)!\cdot[(s-1)!]^{h-2}} =$$

$$= \frac{p!}{(h-1)!\cdot(s+r)!\cdot(s!)^{h-1}} \tag{7.4}$$

puix que $p = hs + r$ en aquest cas.

7.2.1 Nombre de particions per al cas de 13 variables

Tot seguit detallem els diferents números de particions del conjunt constituït per les 13 variables AFNLWGT, AGI, EMCONTRB, ERNVAL, FEDTAX, FICA, INTVAL, PEARNVAL, POTHVAL, PTOTVAL, STATETAX, TAXINC i WSALVAL, objecte del nostre estudi, d'acord amb les 10 variants de microagregació multivariant referides a l'anterior secció:

— `Mic1-12mul`

En el cas d'un grup amb 1 variable i un altre grup amb les restants 12 variables, tindrem el següent número, μ_1, de particions possibles:

$$\mu_1 = \left(\begin{array}{c} 13 \\ 1 \end{array} \right) = 13$$

— `Mic2-11mul`

En el cas d'un grup amb 2 variables i un altre grup amb les restants 11 variables, tindrem el següent número, μ_2, de particions possibles:

$$\mu_2 = \left(\begin{array}{c} 13 \\ 2 \end{array} \right) = \frac{13!}{2!\cdot 11!} = 78$$

— `Mic3-10mul`

En el cas d'un grup amb 3 variables i un altre grup amb les restants 10 variables, tindrem el següent número, μ_3, de particions possibles:

$$\mu_3 = \left(\begin{array}{c} 13 \\ 3 \end{array} \right) = \frac{13!}{3!\cdot 10!} = 286$$

— `Mic4-9mul`

En el cas d'un grup amb 4 variables i un altre grup amb les restants 9 variables, tindrem el següent número, μ_4, de particions possibles:

$$\mu_4 = \left(\begin{array}{c} 13 \\ 4 \end{array} \right) = \frac{13!}{4! \cdot 9!} = 715$$

— `Mic5-8mul`

En el cas d'un grup amb 5 variables i un altre grup amb les restants 8 variables, tindrem el següent número, μ_5, de particions possibles:

$$\mu_5 = \left(\begin{array}{c} 13 \\ 5 \end{array} \right) = \frac{13!}{5! \cdot 8!} = 1287$$

— `Mic6-7mul`

En el cas d'un grup amb 6 variables i un altre grup amb les restants 7 variables, tindrem el següent número, μ_6, de particions possibles:

$$\mu_6 = \left(\begin{array}{c} 13 \\ 6 \end{array} \right) = \frac{13!}{6! \cdot 7!} = 1716$$

— `Mic4mul`

En el cas de dos grups amb 4 variables cadascun i un tercer grup amb les 5 restants variables, mitjançant l'expressió (7.4), calcularem el número, μ_7, de particions possibles de la següent manera:

Com que $13 = 3 \cdot 4 + 1$, llavors:

$$\mu_7 = \frac{p!}{(h-1)! \cdot (s+r)! \cdot (s!)^{h-1}} = \frac{13!}{2! \cdot 5! \cdot (4!)^2} = 45045$$

— `Mic3mul`

En el cas de tres grups amb 3 variables cadascun i un quart grup amb les 4 restants variables, mitjançant l'expressió (7.4), calcularem el número, μ_8, de particions possibles de la següent manera:

Com que $13 = 4 \cdot 3 + 1$, llavors:

$$\mu_8 = \frac{p!}{(h-1)! \cdot (s+r)! \cdot (s!)^{h-1}} = \frac{13!}{3! \cdot 4! \cdot (3!)^3} = 200200$$

— `Mic2mul`

En el cas de cinc grups amb 2 variables cadascun i un sisè grup amb les 3 restants variables, mitjançant l'expressió (7.4), calcularem el número, μ_9, de particions possibles de la següent manera:

Com que $13 = 6 \cdot 2 + 1$, llavors:

$$\mu_9 = \frac{p!}{(h-1)! \cdot (s+r)! \cdot (s!)^{h-1}} = \frac{13!}{5! \cdot 3! \cdot (2!)^5} = 270270$$

— `Micmul`

En el cas d'un sol grup amb totes 13 variables, el número, μ_{10}, de particions possibles, òbviament és $\mu_{10} = 1$.

7.3 Estudi sobre el nombre de variables

En aquesta secció estudiarem com la pèrdua d'informació i la pèrdua de confidencialitat de dades tractades pel mètode de microagregació multivariant DMM (Distància Màxima Modificat) estan influenciades pel nombre de variables que formen els conjunts de la partició sobre la que s'executa el mètode DMM.

La pèrdua d'informació **PI** ha estat calculada i ponderada de la mateixa manera que al capítol 6 (diferenciant tres naturaleses de discrepància entre el conjunt modificat de dades i el conjunt original). Així doncs, seguint també la mateixa nomenclatura que al capítol 6, la pèrdua d'informació calculada ha estat la següent:

$$\mathbf{PI} = \frac{1}{3} \cdot PI1 + \frac{1}{6} \cdot PI2 + \frac{1}{6} \cdot PI4 + \frac{1}{6} \cdot PI3 + \frac{1}{6} \cdot PI5$$

Pel que fa a la pèrdua de confidencialitat **PC**, hem realitzat un estudi diferenciat per a les dues naturaleses de risc de revelació, tal i com han estat definides al capítol 6:

1. Pèrdua de confidencialitat basada en enllaç de registres (ERD).
2. Pèrdua de confidencialitat basada en intervals de confidencialitat (distingint també entre intervals de confidencialitat sobre el número de registres ICN, i sobre la desviació típica ICD).

Finalment, hem estudiat la influència del nombre de variables dels conjunts de la partició sobre la mesura global **MG**, que avalua la qualitat d'un mètode pertorbatiu de control de la revelació considerant a la vegada la pèrdua d'informació i la pèrdua de confidencialitat:

$$\mathbf{MG} = 0.5 \cdot \mathrm{PI} + 0.25 \cdot \mathrm{ERD} + 0.125 \cdot \mathrm{ICN} + 0.125 \cdot \mathrm{ICD}$$

Perquè l'estudi resultés el més complet possible, tots els anteriors mesuraments han estat efectuats, mitjançant implementació dels respectius programes de computació, per a cada un dels tres valors del paràmetre k ($k = 3$, $k = 10$ i $k = 20$), per a cadascuna de les variants de microagregació multivariant `Mic1-12mul`, `Mic2-11mul`, `Mic3-10mul`, `Mic4-9mul`, `Mic5-8mul`, `Mic6-7mul`, `Mic4mul`, `Mic3mul`, `Mic2mul` i `Micmul`, i per a cadascuna de les $\mu_1 = 13$, $\mu_2 = 78$, $\mu_3 = 286$, $\mu_4 = 715$, $\mu_5 = 1287$, $\mu_6 = 1716$, $\mu_7 = 45045$, $\mu_8 = 200200$, $\mu_9 = 270270$ i $\mu_{10} = 1$ respectives possibles particions.

7.3.1 Nombre de variables i pèrdua d'informació

En aquest apartat estudiem la influència del nombre de variables dels conjunts que formen la partició sobre la que s'executa el mètode DMM, respecte la pèrdua d'informació de les dades modificades.

Les files de les Taules 7.1, 7.2 i 7.3 es corresponen amb cada una de les 10 variants de microagregació multivariant, col·locades en el mateix ordre com han estat anteriorment referides. Les columnes MITJANA, MÍNIM, MÀXIM i DESVEST corresponen a la mitjana, el valor mínim, el valor màxim i la desviació típica de la pèrdua d'informació **PI**, calculades i calculats sobre totes les respectives μ_i, $1 \leq i \leq 10$, possibles particions de cada una de les 10 variants de microagregació.

Concretament la columna MÍNIM, després d'una molt breu oscil·lació entre la primera i segona fila, assoleix el valor més petit de les sis primeres files, a la tercera fila (`Mic3-10mul`). D'altra banda,

	MITJANA	MÍNIM	MÀXIM	DESVEST
Mic1-12mul	30,13	20,94	39,66	4,95
Mic2-11mul	30,02	16,70	40,25	4,79
Mic3-10mul	29,60	14,73	44,29	5,27
Mic4-9mul	29,87	15,44	47,50	4,70
Mic5-8mul	29,69	18,26	48,25	4,05
Mic6-7mul	29,26	17,82	48,07	3,44
Mic4mul	21,13	9,91	35,22	3,25
Mic3mul	15,02	6,35	25,89	3,00
Mic2mul	7,08	2,59	14,71	2,01
Micmul	34,85			

Taula 7.1: Mesures de pèrdua d'informació obtingudes en aplicar les diverses versions de microagregació multivariant amb $k = 3$.

	MITJANA	MÍNIM	MÀXIM	DESVEST
Mic1-12mul	39,05	26,51	59,00	10,70
Mic2-11mul	42,06	28,23	72,24	9,42
Mic3-10mul	45,18	25,85	83,37	8,76
Mic4-9mul	47,08	28,78	79,99	8,65
Mic5-8mul	48,36	28,47	94,95	8,59
Mic6-7mul	48,87	30,15	91,47	8,54
Mic4mul	41,49	23,31	77,69	6,27
Mic3mul	32,77	16,10	60,69	5,30
Mic2mul	17,47	7,05	35,60	4,21
Micmul	28,80			

Taula 7.2: Mesures de pèrdua d'informació obtingudes en aplicar les diverses versions de microagregació multivariant amb $k = 10$.

	MITJANA	MÍNIM	MÀXIM	DESVEST
Mic1-12mul	47,85	24,80	76,15	14,60
Mic2-11mul	46,72	26,47	86,80	10,94
Mic3-10mul	48,71	24,57	91,68	11,70
Mic4-9mul	51,29	27,97	115,48	11,91
Mic5-8mul	52,07	26,15	110,98	11,17
Mic6-7mul	53,32	30,27	115,38	11,09
Mic4mul	51,63	28,66	106,79	8,49
Mic3mul	43,74	23,08	87,72	7,15
Mic2mul	26,59	11,78	48,27	4,97
Micmul	54,99			

Taula 7.3: Mesures de pèrdua d'informació obtingudes en aplicar les diverses versions de microagregació multivariant amb $k = 20$.

	MITJANA	MÍNIM	MÀXIM	DESVEST
Mic1-12mul	31,98	26,03	49,31	6,43
Mic2-11mul	34,45	25,58	60,26	7,67
Mic3-10mul	36,88	24,34	58,61	7,91
Mic4-9mul	38,89	23,51	54,89	6,97
Mic5-8mul	40,22	22,57	54,42	5,63
Mic6-7mul	40,92	22,02	53,84	4,49
Mic4mul	54,16	32,21	66,80	4,55
Mic3mul	63,64	42,92	75,87	4,01
Mic2mul	75,10	56,85	81,84	3,30
Micmul	31,79			

Taula 7.4: Mesures de pèrdua de confidencialitat per enllaç de registres obtingudes en aplicar les diverses versions de microagregació multivariant amb $k = 3$.

la columna MITJANA, tot i que a les sis primeres files i per a $k = 3$ sembla presentar la mateixa tendència que la columna MÍNIM, per a valors més grans de k ($k = 10$ i $k = 20$) presenta una clara tendència creixent a les sis primeres files.

Clarament observem una molt ràpida disminució de la pèrdua d'informació per a les variants `Mic4mul`, `Mic3mul` i `Mic2mul`, reflectida a les respectives tres columnes i destacant el valor mínim corresponent a `Mic2mul`.

A la columna DESVEST apareix una disminució continuada, i especialment ràpida a les files setena, vuitena i novena, de la desviació estàndard des de la primera fila fins a la novena, reduint-se a menys de la meitat del seu valor inicial a la novena fila (`Mic2mul`).

Així doncs, destacaríem tres conclusions finals respecte la pèrdua d'informació i el nombre de variables:

- Com cal esperar, en augmentar el valor del paràmetre k, la pèrdua d'informació en general és més alta. Tot i això, excepcionalment, els valors de la columna MÍNIM corresponents a les cinc primeres files són més petits per a $k = 20$ que per a $k = 10$.
- Considerant el número de conjunts que formen la partició sobre la que s'executa el mètode DMM, com més gran és el número de conjunts que formen la partició (major fragmentació del conjunt de variables), menys pèrdua d'informació.
- Suposant particions formades per un mateix número de conjunts ordenades per la diferència (en valor absolut) entre els cardinals dels seus conjunts en ordre descendent, generalment els valors mínims més petits de la pèrdua d'informació s'aconseguiran en les particions que ocupen les posicions centrals.

7.3.2 Nombre de variables i ERD-pèrdua de confidencialitat

En aquest apartat estudiem la influència del nombre de variables dels conjunts que formen la partició sobre la que s'executa el mètode DMM respecte la pèrdua de confidencialitat mitjançant l'enllaç de registres basat en distàncies (ERD) aplicat a les dades modificades.

Les files de les Taules 7.4, 7.5 i 7.6 es corresponen també amb cada una de les 10 variants de microagregació multivariant, col·locades en el mateix ordre com han estat anteriorment referides. Les

	MITJANA	MÍNIM	MÀXIM	DESVEST
Mic1-12mul	19,81	16,92	26,89	3,02
Mic2-11mul	21,28	16,52	35,32	3,69
Mic3-10mul	21,30	14,19	33,48	3,56
Mic4-9mul	21,41	13,77	30,62	3,01
Mic5-8mul	21,55	13,13	30,37	2,48
Mic6-7mul	21,61	13,23	30,75	2,21
Mic4mul	29,76	15,01	43,60	3,64
Mic3mul	40,40	20,60	58,23	4,77
Mic2mul	59,10	35,98	70,91	5,16
Micmul	19,44			

Taula 7.5: Mesures de pèrdua de confidencialitat per enllaç de registres obtingudes en aplicar les diverses versions de microagregació multivariant amb $k = 10$.

	MITJANA	MÍNIM	MÀXIM	DESVEST
Mic1-12mul	15,92	10,60	21,63	3,18
Mic2-11mul	16,87	9,51	24,81	3,00
Mic3-10mul	17,22	10,79	26,03	2,58
Mic4-9mul	17,30	9,47	23,36	2,36
Mic5-8mul	17,42	10,67	23,89	2,07
Mic6-7mul	17,46	10,81	24,15	1,90
Mic4mul	21,60	11,71	32,63	2,50
Mic3mul	28,79	14,47	45,50	3,75
Mic2mul	46,46	26,35	61,81	5,30
Micmul	13,23			

Taula 7.6: Mesures de pèrdua de confidencialitat per enllaç de registres obtingudes en aplicar les diverses versions de microagregació multivariant amb $k = 20$.

	MITJANA	MÍNIM	MÀXIM	DESVEST
Mic1-12mul	63,70	62,13	66,97	1,36
Mic2-11mul	64,77	61,88	72,02	2,22
Mic3-10mul	65,35	61,15	73,05	2,16
Mic4-9mul	65,74	62,21	72,84	1,91
Mic5-8mul	65,97	62,82	74,00	1,66
Mic6-7mul	66,09	63,33	75,05	1,49
Mic4mul	73,11	69,07	82,08	1,95
Mic3mul	79,31	74,09	87,07	1,95
Mic2mul	89,01	86,13	95,29	1,48
Micmul	62,93			

Taula 7.7: Mesures de pèrdua de confidencialitat per intervals de confiança, basat en rangs, obtingudes en aplicar les diverses versions de microagregació multivariant amb $k = 3$.

columnes MITJANA, MÍNIM, MÀXIM i DESVEST corresponen a la mitjana, el valor mínim, el valor màxim i la desviació típica de la ERD-pèrdua de confidencialitat, calculades i calculats sobre totes les respectives μ_i, $1 \leq i \leq 10$, possibles particions de cada una de les 10 variants de microagregació.

A la columna MÍNIM apareix una disminució dels valors mínims de la ERD-pèrdua de confidencialitat, des de la primera fila fins a les files cinquena i sisena (`Mic5-8mul` i `Mic6-7mul`), on s'assoleixen els dos valors més petits; disminució que, per altra banda, tendeix a estabilitzar-se per a valors grans del paràmetre k ($k = 20$). Tot i això, a la columna MITJANA de les mateixes Taules observem un suau increment de la ERD-pèrdua de confidencialitat mitjana a les sis primeres files, però amb tendència a estabilitzar-se per als valors centrals del paràmetre k ($k = 10$).

D'altra banda, a les files setena, vuitena i novena (`Mic4mul`, `Mic3mul`, `Mic2mul`) de les columnes MITJANA i MÍNIM podem observar clarament un augment molt ràpid de la ERD-pèrdua de confidencialitat.

A la columna DESVEST també apareixen les dues desviacions estàndard més petites a les mateixes files cinquena i sisena.

La columna MÀXIM presenta un comportament molt similar a la columna MÍNIM.

Tres conclusions finals respecte del nombre de variables i la ERD-pèrdua de confidencialitat:

- Com caldria esperar, en augmentar el valor del paràmetre k, la ERD-pèrdua de confidencialitat disminueix.
- Considerant el número de conjunts que formen la partició sobre la que s'executa el mètode DMM, com més petit és el número de conjunts que formen la partició (menor fragmentació del conjunt de variables), menys ERD-pèrdua de confidencialitat.
- Suposant particions formades per un mateix número de conjunts ordenades per la diferència (en valor absolut) entre els cardinals dels seus conjunts en ordre descendent, generalment els valors mínims més petits de la ERD-pèrdua de confidencialitat s'aconseguiran a partir de les particions amb menor diferència entre els seus cardinals, però endarrerint progressivament la seva posició a mesura que augmenta el valor del paràmetre k.

7.3.3 Nombre de variables i ICN-pèrdua de confidencialitat

En aquest apartat estudiem la influència del nombre de variables dels conjunts que formen la

	MITJANA	MÍNIM	MÀXIM	DESVEST
Mic1-12mul	53,15	51,15	55,83	1,38
Mic2-11mul	52,80	50,08	57,93	1,60
Mic3-10mul	52,10	48,16	58,04	1,80
Mic4-9mul	51,54	48,00	57,79	1,63
Mic5-8mul	51,20	47,38	59,30	1,49
Mic6-7mul	51,01	48,42	58,69	1,40
Mic4mul	55,96	51,26	68,95	2,21
Mic3mul	62,46	56,39	74,03	2,54
Mic2mul	74,81	69,40	86,55	2,27
Micmul	55,21			

Taula 7.8: Mesures de pèrdua de confidencialitat per intervals de confiança, basats en rangs, obtingudes en aplicar les diverses versions de microagregació multivariant amb $k = 10$.

	MITJANA	MÍNIM	MÀXIM	DESVEST
Mic1-12mul	47,90	44,35	50,85	1,93
Mic2-11mul	47,08	43,31	51,82	1,75
Mic3-10mul	46,31	42,70	51,97	1,65
Mic4-9mul	45,55	41,75	51,87	1,57
Mic5-8mul	45,07	41,59	52,01	1,48
Mic6-7mul	44,75	41,65	52,82	1,37
Mic4mul	47,63	42,71	60,45	2,00
Mic3mul	52,69	46,35	65,10	2,48
Mic2mul	64,03	58,14	76,90	2,52
Micmul	49,30			

Taula 7.9: Mesures de pèrdua de confidencialitat per intervals de confiança, basats en rangs, obtingudes en aplicar les diverses versions de microagregació multivariant amb $k = 20$.

	MITJANA	MÍNIM	MÀXIM	DESVEST
Mic1-12mul	33,14	31,62	38,63	1,93
Mic2-11mul	34,78	31,65	43,37	2,35
Mic3-10mul	35,20	31,68	44,23	2,17
Mic4-9mul	35,26	31,60	44,47	2,18
Mic5-8mul	35,30	31,42	46,72	2,21
Mic6-7mul	35,35	31,75	49,44	2,22
Mic4mul	44,53	38,27	63,52	3,35
Mic3mul	54,27	45,96	72,00	3,68
Mic2mul	72,05	65,33	85,77	2,98
Micmul	30,72			

Taula 7.10: Mesures de pèrdua de confidencialitat per intervals de confiança, basats en desviacions típiques, obtingudes en aplicar les diverses versions de microagregació multivariant amb $k = 3$.

partició sobre la que s'executa el mètode DMM respecte la pèrdua de confidencialitat a través d'intervals de confidencialitat sobre el número de registres (ICN) aplicats a les dades modificades.

Les files de les Taules 7.7, 7.8 i 7.9 es corresponen igualment amb cada una de les 10 variants de microagregació multivariant, col·locades en el mateix ordre com han estat anteriorment referides. Les columnes MITJANA, MÍNIM, MÀXIM i DESVEST corresponen a la mitjana, el valor mínim, el valor màxim i la desviació típica de la ICN-pèrdua de confidencialitat, calculades i calculats sobre totes les respectives μ_i, $1 \leq i \leq 10$, possibles particions de cada una de les 10 variants de microagregació.

Observem a les columnes MITJANA i MÍNIM una disminució dels respectius valors mitjans i mínims de la ICN-pèrdua de confidencialitat, des de la primera fila fins a les files cinquena i sisena (`Mic5-8mul` i `Mic6-7mul`), on s'assoleixen els valors més petits; tanmateix, disminució que tendeix a estabilitzar-se en totes dues columnes per a valors petits del paràmetre k ($k = 3$).

D'altra banda, podem observar un ràpid augment dels també respectius valors mitjans i mínims a les files setena, vuitena i novena (`Mic4mul`, `Mic3mul`, `Mic2mul`).

A la columna DESVEST també apareixen desviacions estàndard petites a les mateixes files cinquena i sisena.

Tres conclusions finals respecte del nombre de variables i la ICN-pèrdua de confidencialitat:

- Com caldria esperar, en augmentar el valor del paràmetre k, la ICN-pèrdua de confidencialitat disminueix.
- Considerant el número de conjunts que formen la partició sobre la que s'executa el mètode DMM, com més petit és el número de conjunts que formen la partició (menor fragmentació del conjunt de variables), menys ICN-pèrdua de confidencialitat.
- Suposant particions formades per un mateix número de conjunts ordenades per la diferència (en valor absolut) entre els cardinals dels seus conjunts en ordre descendent, generalment els valors mínims més petits de la ICN-pèrdua de confidencialitat s'aconseguiran a partir de les particions centrals, però avançant progressivament la seva posició a mesura que augmenta el valor del paràmetre k.

7.3.4 Nombre de variables i ICD-pèrdua de confidencialitat

En aquest apartat estudiem la influència del nombre de variables dels conjunts que formen la

	MITJANA	MÍNIM	MÀXIM	DESVEST
Mic1-12mul	24,63	22,99	26,38	1,07
Mic2-11mul	24,41	21,98	28,32	1,27
Mic3-10mul	23,85	20,42	28,84	1,35
Mic4-9mul	23,36	20,36	28,58	1,34
Mic5-8mul	23,06	20,08	30,19	1,29
Mic6-7mul	22,91	20,62	32,05	1,24
Mic4mul	27,36	22,74	43,95	2,48
Mic3mul	33,94	27,02	54,24	3,45
Mic2mul	49,61	41,91	69,23	3,96
Micmul	24,98			

Taula 7.11: Mesures de pèrdua de confidencialitat per intervals de confiança, basats en desviacions típiques, obtingudes en aplicar les diverses versions de microagregació multivariant amb $k = 10$.

	MITJANA	MÍNIM	MÀXIM	DESVEST
Mic1-12mul	22,70	20,50	25,77	1,48
Mic2-11mul	21,56	18,47	25,00	1,41
Mic3-10mul	20,72	16,99	25,43	1,38
Mic4-9mul	20,15	17,24	24,80	1,29
Mic5-8mul	19,78	16,85	26,56	1,17
Mic6-7mul	19,57	17,14	26,63	1,10
Mic4mul	21,79	18,17	35,70	1,75
Mic3mul	26,10	20,88	43,25	2,78
Mic2mul	38,18	30,61	56,97	3,88
Micmul	23,22			

Taula 7.12: Mesures de pèrdua de confidencialitat per intervals de confiança, basats en desviacions típiques, obtingudes en aplicar les diverses versions de microagregació multivariant amb $k = 20$.

	MITJANA	MÍNIM	MÀXIM	DESVEST
Mic1-12mul	35,17	29,48	43,06	3,72
Mic2-11mul	36,07	27,53	43,66	3,66
Mic3-10mul	36,59	27,62	43,81	3,58
Mic4-9mul	37,28	28,63	46,19	3,01
Mic5-8mul	37,56	28,21	45,79	2,54
Mic6-7mul	37,54	29,95	45,88	2,18
Mic4mul	38,81	30,20	45,30	2,03
Mic3mul	40,12	31,94	46,37	1,82
Mic2mul	42,45	35,86	46,78	1,40
Micmul	37,08			

Taula 7.13: Mesures generals obtingudes en aplicar les diverses versions de microagregació multivariant amb $k = 3$.

partició sobre la que s'executa el mètode DMM respecte la pèrdua de confidencialitat a través d'intervals de confidencialitat sobre la desviació estandard (ICD) aplicats a les dades modificades.

Les files de les Taules 7.10, 7.11 i 7.12 es corresponen igualment amb cada una de les 10 variants de microagregació multivariant, col·locades en el mateix ordre com han estat anteriorment referides. Les columnes MITJANA, MÍNIM, MÀXIM i DESVEST corresponen a la mitjana, el valor mínim, el valor màxim i la desviació típica de la ICD-pèrdua de confidencialitat, calculades i calculats sobre totes les respectives μ_i, $1 \leq i \leq 10$, possibles particions de cada una de les 10 variants de microagregació.

Observem a les columnes MITJANA i MÍNIM una disminució dels respectius valors mitjans i mínims de la ICD-pèrdua de confidencialitat, des de la primera fila fins a les files cinquena i sisena (`Mic5-8mul` i `Mic6-7mul`), on s'assoleixen els valors més petits; tanmateix, disminució que tendeix a estabilitzar-se en totes dues columnes per a valors petits del paràmetre k ($k = 3$).

D'altra banda, podem observar un ràpid augment dels també respectius valors mitjans i mínims a les files setena, vuitena i novena (`Mic4mul`, `Mic3mul`, `Mic2mul`).

La columna MÀXIM presenta un augment continuat dels valors màxims de la ICD-pèrdua de confidencialitat, lleugerament evidenciat fins la fila sisena, i molt remarcat a les files setena, vuitena i novena.

Tres conclusions finals respecte del nombre de variables i la ICD-pèrdua de confidencialitat:

- Com caldria esperar, en augmentar el valor del paràmetre k, la ICD-pèrdua de confidencialitat disminueix. A més a més, com podem observar a les respectives taules, els valors de la ICD-pèrdua de confidencialitat es redueixen pràcticament a la meitat dels valors corresponents a la ICN-pèrdua de confidencialitat.

- Considerant el número de conjunts que formen la partició sobre la que s'executa el mètode DMM, com més petit és el número de conjunts que formen la partició (menor fragmentació del conjunt de variables), menys ICD-pèrdua de confidencialitat.

- Suposant particions formades per un mateix número de conjunts ordenades per la diferència (en valor absolut) entre els cardinals dels seus conjunts en ordre descendent, generalment els valors mínims més petits de la ICD-pèrdua de confidencialitat s'aconseguiran en les posicions immediatament anteriors a les darreres.

	MITJANA	MÍNIM	MÀXIM	DESVEST
Mic1-12mul	34,20	27,14	46,36	5,90
Mic2-11mul	36,00	28,22	52,14	5,50
Mic3-10mul	37,41	26,40	57,35	4,90
Mic4-9mul	38,26	28,00	54,93	4,72
Mic5-8mul	38,85	28,21	61,64	4,46
Mic6-7mul	39,08	29,01	60,63	4,35
Mic4mul	38,60	27,65	53,94	3,20
Mic3mul	38,54	28,35	51,63	2,91
Mic2mul	39,06	30,06	47,85	2,54
Micmul	29,29			

Taula 7.14: Mesures generals obtingudes en aplicar les diverses versions de microagregació multivariant amb $k = 10$.

	MITJANA	MÍNIM	MÀXIM	DESVEST
Mic1-12mul	36,73	26,42	52,79	7,54
Mic2-11mul	36,15	25,08	54,61	5,84
Mic3-10mul	37,04	24,47	59,72	6,07
Mic4-9mul	38,18	25,77	70,83	6,08
Mic5-8mul	38,50	24,94	68,52	5,65
Mic6-7mul	39,06	26,63	71,05	5,59
Mic4mul	39,89	28,56	67,34	4,26
Mic3mul	38,91	28,76	59,39	3,64
Mic2mul	37,69	28,19	48,93	2,89
Micmul	39,87			

Taula 7.15: Mesures generals obtingudes en aplicar les diverses versions de microagregació multivariant amb $k = 20$.

7.3.5 Nombre de variables i mesura global MG sobre la qualitat

En aquest darrer apartat estudiem la influència del nombre de variables dels conjunts que formen la partició sobre la que s'executa el mètode DMM respecte la mesura global **MG** sobre la qualitat d'un mètode de control de la revelació estadística pertorbatiu.

Les files de les Taules 7.13, 7.14 i 7.15 es corresponen també amb cada una de les 10 variants de microagregació multivariant, col·locades en el mateix ordre com han estat anteriorment referides. Les columnes MITJANA, MÍNIM, MÀXIM i DESVEST corresponen a la mitjana, el valor mínim, el valor màxim i la desviació típica de la mesura global **MG**, calculades i calculats sobre totes les respectives μ_i, $1 \leq i \leq 10$, possibles particions de cada una de les 10 variants de microagregació.

La columna MITJANA presenta bàsicament un lleuger increment dels valors mitjos de la mesura global **MG**, en augmentar el número de conjunts que formen la partició sobre la que s'executa el mètode DMM.

Tot i això, la columna MÍNIM presenta oscil·lacions de creixement i decreixement, però augmentant molt lleugerament en el mateix sentit que la columna MITJANA. Cal destacar el valor més petit de la columna MÍNIM assolit al voltant de la tercera fila (`Mic3-10mul`).

A la columna MÀXIM apareix bàsicament un augment dels valors màxims des de la primera fila fins la fila central (`Mic5-8mul`), i, a continuació, una disminució dels valors màxims fins la darrera fila.

Finalment, a la columna DESVEST apareix un continuat decreixement de la desviació típica de la mesura global **MG**, en augmentar el número de conjunts que formen la partició sobre la que s'executa el mètode DMM.

Conclusions pel que fa a la relació entre el nombre de variables i la mesura global **MG** sobre la qualitat d'un mètode de control de la revelació estadística pertorbatiu:

- Respecte dels valors del paràmetre k, tot i que l'interval de variació de la mesura global **MG** es manté bastant estable, cal destacar la general disminució dels valors de la columna MÍNIM quan augmenta el valor del paràmetre k.
- Tots els anteriors comentaris semblen evidenciar que els millors valors de la mesura global **MG** (valors més petits) s'aconsegueixin per a les particions del conjunt de variables amb poca fragmentació (valors petits del número de conjunts que les formen).
- Suposant particions formades per un mateix número de conjunts ordenades per la diferència (en valor absolut) entre els cardinals dels seus conjunts en ordre descendent, molt probablement els valors mínims més petits de la mesura global **MG** s'aconseguiran en les particions que ocupen les posicions centrals (`Mic3-10mul`).

Capítol 8

Estudi de la combinació de variables

8.1 Necessitat de mesures prèvies per a cada combinació de variables

Com ha estat ja referit a l'anterior capítol 7, els dos mètodes DM (Distància Màxima) i DMM (Distància Màxima Modificat) per fer microagregació multivariant sense projectar les dades, tindrien els següents dos casos extrems com a criteris per agrupar registres: microagregació amb ordenació individual (microagregant una sola variable cada cop), i microagregació amb totes les variables simultàniament. Entre aquests dos casos extrems existeixen moltes altres possibilitats, que estan determinades per dues característiques bàsiques:

1. Nombre de variables que ha de tenir cada conjunt de la partició sobre la qual s'executaran els mètodes DM i DMM. En aquest aspecte, un primer objectiu serà esbrinar quin és el nombre de variables més apropiat per obtenir unes bones mesures de qualitat.
2. Combinació de les variables dintre els conjunts que formen la partició. Una vegada determinat el nombre de variables més apropiat, un segon objectiu serà cercar, per a aquest nombre, les combinacions de les variables que optimitzen les diferents mesures sobre la qualitat d'aquests mètodes.

L'anterior capítol 7 ha estat dedicat precisament al treball de recerca realitzat respecte al primer objectiu (nombre de variables) en el mètode DMM (Distància Màxima Modificat), mètode amb menys complexitat computacional que el mètode DM (Distància Màxima), com ha estat provat al capítol 4.

En aquest capítol presentem el treball realitzat respecte al segon objectiu, que consisteix a trobar, per al nombre de variables més apropiat, les combinacions de variables dels conjunts de la partició que donaran bones mesures de qualitat mitjançant l'execució del mètode DMM. Per determinar aquest segon objectiu, s'han creat diferents estadístics, lligats a cada combinació de variables i a la vegada molt ràpids de càlcul, per tal d'observar si algun o alguns d'ells estan correlacionats amb la pèrdua d'informació, el risc de revelació (mesures ERD, ICN i ICD) i la mesura global de qualitat **MG**.

L'objectiu final del treball es proporcionar a l'usuari de la microagregació multivariant una guia per saber quantes i quines variables haurien de formar cada grup de la partició sobre la qual s'executarà el mètode DMM, per obtenir conjunts de dades modificats amb bones mesures de qualitat.

Per aconseguir aquest objectiu hem treballat amb els mateixos 1080 registres individuals i les mateixes 13 variables AFNLWGT, AGI, EMCONTRB, ERNVAL, FEDTAX, FICA, INTVAL,

PEARNVAL, POTHVAL, PTOTVAL, STATETAX, TAXINC, WSALVAL, dels capítols anteriors.

Aquest estudi ha estat desenvolupat per a cada un dels tres aspectes que determinen la qualitat del mètode DMM: pèrdua d'informació, risc de revelació i mesura global.

Les variants de microagregació multivariant analitzades en aquest estudi, amb el corresponent número de possibles combinacions de variables per a cadascuna d'elles, han estat les següents:

- `Mic1-12mul`: Un grup amb 1 variable i un altre grup amb les restants 12 variables. Aquest mètode presenta 13 possibles combinacions de variables.
- `Mic2-11mul`: Un grup amb 2 variables i un altre grup amb les restants 11 variables. Aquest mètode presenta 78 possibles combinacions de variables.
- `Mic3-10mul`: Un grup amb 3 variables i un altre grup amb les restants 10 variables. Aquest mètode presenta 286 possibles combinacions de variables.
- `Mic4-9mul`: Un grup amb 4 variables i un altre grup amb les restants 9 variables. Aquest mètode presenta 715 possibles combinacions de variables.
- `Mic5-8mul`: Un grup amb 5 variables i un altre grup amb les restants 8 variables. Aquest mètode presenta 1287 possibles combinacions de variables.
- `Mic6-7mul`: Un grup amb 6 variables i un altre grup amb les restants 7 variables. Aquest mètode presenta 1716 possibles combinacions de variables.

D'acord amb els resultats obtinguts al capítol anterior, a través d'aquest estudi hem analitzat especialment les millors variants de microagregació pel que fa als cardinals dels conjunts que formen la partició.

8.2 Estadístics per al control de la revelació

Per evitar l'efecte del tipus d'unitats en què cada variable està mesurada, hem estandarditzat les dades de les 13 variables corresponents als 1080 individus. Podríem diferenciar, en principi, cinc maneres d'estandarditzar:

Estandardització SX. Per a cada variable, es calcula el màxim valor absolut entre els seus respectius 1080 valors absoluts; i es divideixen els 1080 valors de cada variable pel seu corresponent màxim valor absolut.

Estandardització SN. Per a cada variable, es calcula el mínim valor absolut entre els seus respectius 1080 valors absoluts; i es divideixen els 1080 valors de cada variable pel seu corresponent mínim valor absolut.

Estandardització SM. Per a cada variable, es calcula la respectiva mitjana aritmètica, per la qual es divideixen tots els seus 1080 corresponents valors.

Estandardització SD. Per a cada variable, es calcula la respectiva desviació típica, per la qual es divideixen tots els seus 1080 corresponents valors.

Estandardització ST. Per a cada variable, es calculen les respectives mitjana aritmètica i desviació típica. Després es divideixen, per la desviació típica calculada, totes les diferències entre cadascun dels seus 1080 valors i la mitjana aritmètica.

En el nostre estudi hem emprat l'Estandardització ST.

Seguint la mateixa nomenclatura utilitzada a capítols anteriors, per a cada una de les particions $\mathbf{G} \in \mathcal{F}(\mathbf{V})$, $\mathbf{G} = \{G_1, G_2, G_3, \ldots, G_h\}$, on

$$G_l = \{V_{l_1}, V_{l_2}, \ldots, V_{l_s}\} \qquad \text{per a} \qquad 1 \leq l \leq h-1$$

$$G_h = \{V_{h_1}, V_{h_2}, \ldots, V_{h_{s+r}}\}$$

hem calculat dos tipus d'estadístics:

Estadístics intra-grup. Anomenats d'aquesta manera perquè els càlculs vectorials només es realitzen dintre de cada un dels conjunts de cada partició; tot i que posteriorment es comparen els resultats obtinguts a cada conjunt.

Estadístics inter-grup. Anomenats d'aquesta manera perquè generalment els càlculs vectorials es realitzen en dues fases: en una primera fase, s'efectuen dintre de cada un dels conjunts de cada partició; i, en una segona fase, es realitzen entre els resultats obtinguts a la primera fase.

8.2.1 Estadístics Intra-grup

Hem classificat els estadístics intra-grup estudiats de la següent manera:

1.Mòduls de variables intra-grup

Dintre de cada grup $G_1, G_2, G_3, \ldots, G_h$ de la partició $\mathbf{G}$, es calculen els mòduls al quadrat de cadascuna de les variables que conté:

$$\|V_{l_1}\|^2, \|V_{l_2}\|^2, \ldots, \|V_{l_s}\|^2 \qquad \text{per a} \qquad 1 \leq l \leq h-1$$

$$\|V_{h_1}\|^2, \|V_{h_2}\|^2, \ldots, \|V_{h_{s+r}}\|^2$$

2.Diferències entre mòduls de variables intra-grup

Dintre de cada grup $G_1, G_2, G_3, \ldots, G_h$ de la partició $\mathbf{G}$, i per a cada variable que conté, es calcula la diferència, en valor absolut, entre el mòdul al quadrat d'aquesta i el mòdul al quadrat de cadascuna de les altres variables que conté:

$$|\, \|V_{l_i}\|^2 - \|V_{l_k}\|^2 \,| \qquad \text{per a} \qquad 1 \leq i < k \leq s \qquad 1 \leq l \leq h-1$$

$$|\, \|V_{h_i}\|^2 - \|V_{h_k}\|^2 \,| \qquad \text{per a} \qquad 1 \leq i < k \leq s+r$$

3.Distàncies entre variables intra-grup

Dintre de cada grup $G_1, G_2, G_3, \ldots, G_h$ de la partició $\mathbf{G}$, es calcula el quadrat de la distància euclidiana de cadascuna de les variables que conté a totes les altres:

$$\|V_{l_i} - V_{l_k}\|^2 \qquad \text{per a} \qquad 1 \leq i < k \leq s \qquad 1 \leq l \leq h-1$$

$$\|V_{h_i} - V_{h_k}\|^2 \qquad \text{per a} \qquad 1 \leq i < k \leq s+r$$

4.Desviacions modulars intra-grup

Primerament, per als grups $G_1, G_2, G_3, \ldots, G_h$ de la partició $\mathbf{G}$, es calculen els seus respectius vectors mitjana-individu, $M_1, M_2, M_3, \ldots, M_h$, les n coordenades de cadascun dels quals es corresponen amb les mitjanes aritmètiques dels valors estandarditzats de les variables que pertanyen a cada grup, per a cadascun dels n individus:

$$M_l = \frac{1}{s}(V_{l_1} + V_{l_2} + \cdots + V_{l_s}) \qquad \text{per a} \qquad 1 \leq l \leq h-1$$

$$M_h = \frac{1}{s+r}(V_{h_1} + V_{h_2} + \cdots + V_{h_{s+r}})$$

Posteriorment, dintre de cada grup $G_1, G_2, G_3, \ldots, G_h$ de la partició $\mathbf{G}$, es calculen les diferències, en valor absolut, entre el mòdul al quadrat de cadascuna de les seves variables i el mòdul al quadrat del vector mitjana-individu del grup:

$$|\,\|V_{l_i}\|^2 - \|M_l\|^2\,| \qquad \text{per a} \qquad 1 \leq i \leq s \qquad 1 \leq l \leq h-1$$

$$|\,\|V_{h_i}\|^2 - \|M_h\|^2\,| \qquad \text{per a} \qquad 1 \leq i \leq s+r$$

5.Desviacions vectorials intra-grup

Primerament, es calculen els vectors mitjana-individu $M_1, M_2, M_3, \ldots, M_h$, de la mateixa manera que a l'anterior estadístic **4**.

Posteriorment, dintre de cada grup $G_1, G_2, G_3, \ldots, G_h$ de la partició $\mathbf{G}$, es calculen els quadrats de les distàncies euclidianes entre cadascuna de les seves variables i el punt que defineix el vector mitjana-individu del grup:

$$\|V_{l_i} - M_l\|^2 \qquad \text{per a} \qquad 1 \leq i \leq s \qquad 1 \leq l \leq h-1$$

$$\|V_{h_i} - M_h\|^2 \qquad \text{per a} \qquad 1 \leq i \leq s+r$$

6.Productes escalars intra-grup

Dintre de cada grup $G_1, G_2, G_3, \ldots, G_h$ de la partició $\mathbf{G}$, es calcula el producte escalar entre cada un dels vectors-variable que conté i tots els altres:

$$V_{l_i} \cdot V_{l_k} \qquad \text{per a} \qquad 1 \leq i < k \leq s \qquad 1 \leq l \leq h-1$$

$$V_{h_i} \cdot V_{h_k} \qquad \text{per a} \qquad 1 \leq i < k \leq s+r$$

7.Coeficients angulars intra-grup

Dintre de cada grup $G_1, G_2, G_3, \ldots, G_h$ de la partició $\mathbf{G}$, es calculen, entre cada un dels vectors-variable que conté i tots els altres, el següents coeficients angulars:

$$\frac{V_{l_i} \cdot V_{l_k}}{\|V_{l_i}\|\,\|V_{l_k}\|} \qquad \text{per a} \qquad 1 \leq i < k \leq s \qquad 1 \leq l \leq h-1$$

$$\frac{V_{h_i} \cdot V_{h_k}}{\|V_{h_i}\|\,\|V_{h_k}\|} \qquad \text{per a} \qquad 1 \leq i < k \leq s+r$$

8.Covariàncies intra-grup

Siguin $\nu_1, \nu_2, \nu_3, \ldots, \nu_p$ les corresponents mitjanes aritmètiques de les variables $V_1, V_2, V_3, \ldots, V_p$ calculades sobre els n individus. Sigui $\mathbf{1}$ el vector fila de n columnes amb totes les seves coordenades iguals a la unitat.

Dintre de cada grup $G_1, G_2, G_3, \ldots, G_h$ de la partició $\mathbf{G}$, es calcula la covariància entre cada un dels vectors-variable que conté i tots els altres:

$$\frac{1}{n}(V_{l_i} - \mathbf{1}\nu_{l_i}) \cdot (V_{l_k} - \mathbf{1}\nu_{l_k}) \qquad \text{per a} \qquad 1 \leq i < k \leq s \qquad 1 \leq l \leq h-1$$

$$\frac{1}{n}(V_{h_i} - \mathbf{1}\nu_{h_i}) \cdot (V_{h_k} - \mathbf{1}\nu_{h_k}) \qquad \text{per a} \qquad 1 \leq i < k \leq s+r$$

on el signe $\cdot$ indica el producte escalar ordinari entre vectors.

9.Coeficients de correlació intra-grup

Dintre de cada grup $G_1, G_2, G_3, \ldots, G_h$ de la partició $\mathbf{G}$, es calcula el coeficient de correlació entre cada un dels vectors-variable que conté i tots els altres.

Amb la mateixa nomenclatura utilitzada a l'estadístic **8**, tindríem:

$$\frac{(V_{l_i} - \mathbf{1}\nu_{l_i}) \cdot (V_{l_k} - \mathbf{1}\nu_{l_k})}{\|V_{l_i} - \mathbf{1}\nu_{l_i}\| \, \|V_{l_k} - \mathbf{1}\nu_{l_k}\|} \qquad \text{per a} \qquad 1 \leq i < k \leq s \qquad 1 \leq l \leq h-1$$

$$\frac{(V_{h_i} - \mathbf{1}\nu_{h_i}) \cdot (V_{h_k} - \mathbf{1}\nu_{h_k})}{\|V_{h_i} - \mathbf{1}\nu_{h_i}\| \, \|V_{h_k} - \mathbf{1}\nu_{h_k}\|} \qquad \text{per a} \qquad 1 \leq i < k \leq s+r$$

10.Productes escalars i mòduls intra-grup

De igual manera que l'estadístic **6**, dintre de cada grup $G_1, G_2, G_3, \ldots, G_h$ de la partició $\mathbf{G}$, es calcula el producte escalar entre cada un dels vectors-variable que conté i tots els altres. A més a més, s'afegeix el mòdul al quadrat de cada vector-variable (producte escalar entre cada vector-variable i ell mateix) dintre de cada grup.

11.Covariàncies i variàncies intra-grup

De igual manera que l'estadístic **8**, dintre de cada grup $G_1, G_2, G_3, \ldots, G_h$ de la partició $\mathbf{G}$, es calcula la covariància entre cada un dels vectors-variable que conté i tots els altres. A més a més, s'afegeix la variància de cada vector-variable (covariància entre cada vector-variable i ell mateix) dintre de cada grup.

12.Variàncies intra-grup

Dintre de cada grup $G_1, G_2, G_3, \ldots, G_h$ de la partició $\mathbf{G}$, es calcula la variància de cada un dels vectors-variable que conté:

$$\frac{1}{n}(V_{l_i} - \mathbf{1}\nu_{l_i}) \cdot (V_{l_i} - \mathbf{1}\nu_{l_i}) \qquad \text{per a} \qquad 1 \leq i \leq s \qquad 1 \leq l \leq h-1$$

$$\frac{1}{n}(V_{h_i} - \mathbf{1}\nu_{h_i}) \cdot (V_{h_i} - \mathbf{1}\nu_{h_i}) \qquad \text{per a} \qquad 1 \leq i \leq s+r$$

on el signe $\cdot$ indica el producte escalar entre vectors ordinari.

Per a cadascuna de les particions del conjunt $\mathcal{F}(\mathbf{V})$, i per a cadascun d'aquests estadístics intra-grup s'han computat, mitjançant implementació dels corresponents programes, els següents indicadors:

Indicadors intra-grup

1. El valor mínim que assoleix cada estadístic dintre de cada grup.
2. El valor màxim que assoleix cada estadístic dintre de cada grup.
3. La diferència entre els valors màxim i mínim que assoleix cada estadístic dintre de cada grup.
4. La suma dels valors màxim i mínim que assoleix cada estadístic dintre de cada grup.
5. La mitjana aritmètica de tots els valors de cada estadístic dintre de cada grup.
6. La mediana de tots els valors de cada estadístic dintre de cada grup.
7. La desviació típica de tots els valors de cada estadístic dintre de cada grup.

Indicadors inter-grup

a. Mitjana aritmètica, mediana i desviació típica de tots els valors mínims corresponents a cadascun dels grups i calculats com indicadors intra-grup.

b. Mitjana aritmètica, mediana i desviació típica de tots els valors màxims corresponents a cadascun dels grups i calculats com indicadors intra-grup.

c. Mitjana aritmètica, mediana i desviació típica de totes les diferències entre els valors màxim i mínim, corresponents a cadascun dels grups i calculades com indicadors intra-grup.

d. Mitjana aritmètica, mediana i desviació típica de totes les sumes dels valors màxim i mínim, corresponents a cadascun dels grups i calculades com indicadors intra-grup.

e. Mitjana aritmètica, mediana i desviació típica de totes les mitjanes aritmètiques corresponents a cadascun dels grups i calculades com indicadors intra-grup.

f. Mitjana aritmètica, mediana i desviació típica de totes les medianes corresponents a cadascun dels grups i calculades com indicadors intra-grup.

g. Mitjana aritmètica, mediana i desviació típica de totes les desviacions típiques corresponents a cadascun dels grups i calculades com indicadors intra-grup.

8.2.2 Estadístics Inter-grup

Tots aquests estadístics inter-grup han estat calculats sobre els vectors mitjana-individu, $M_1, M_2, M_3, \ldots, M_h$ dels grups $G_1, G_2, G_3, \ldots, G_h$ de la partició $\mathbf{G}$, les n coordenades de cadascun dels quals es corresponen amb les mitjanes aritmètiques dels valors estandarditzats de les variables que pertanyen a cada grup, per a cadascun dels n individus:

$$M_l = \frac{1}{s}(V_{l_1} + V_{l_2} + \cdots + V_{l_s}) \qquad \text{per a} \qquad 1 \leq l \leq h-1$$

$$M_h = \frac{1}{s+r}(V_{h_1} + V_{h_2} + \cdots + V_{h_{s+r}})$$

Hem classificat els estadístics inter-grup estudiats de la següent manera:

13.Mòduls dels vectors mitjana-individu

Es calculen els mòduls al quadrat dels vectors mitjana-individu $M_1, M_2, M_3, \ldots, M_h$:

$$\|M_1\|^2, \|M_2\|^2, \|M_3\|^2 \ldots, \|M_h\|^2$$

14.Diferències entre mòduls de vectors mitjana-individu

Per a cada vector mitjana-individu, es calcula la diferència, en valor absolut, entre el mòdul al quadrat d'aquest i el mòdul al quadrat de cadascun dels altres:

$$| \|M_i\|^2 - \|M_k\|^2 | \qquad \text{per a} \qquad 1 \leq i < k \leq h$$

15.Distàncies entre vectors mitjana-individu

Es calcula el quadrat de la distància euclidiana de cadascun dels vectors mitjana-individu a tots els altres:

$$\|M_i - M_k\|^2 \qquad \text{per a} \qquad 1 \leq i < k \leq h$$

16.Desviacions modulars dels vectors mitjana-individu

Primerament, es calcula el vector Ψ, mitjana dels vectors mitjana-individu, $M_1, M_2, M_3, \ldots, M_h$:

$$\Psi = \frac{1}{h}(M_1 + M_2 + M_3 + \cdots + M_h)$$

Posteriorment, es calculen les diferències, en valor absolut, entre el mòdul al quadrat de cadascun dels vectors mitjana-individu i el mòdul al quadrat del vector mitjana Ψ:

$$| \|M_i\|^2 - \|\Psi\|^2 | \qquad \text{per a} \qquad 1 \leq i \leq h$$

17.Desviacions vectorials dels vectors mitjana-individu

Primerament, es calcula el vector Ψ, mitjana dels vectors mitjana-individu, $M_1, M_2, M_3, \ldots, M_h$, de la mateixa manera que a l'anterior estadístic **16**.

Posteriorment, es calculen els quadrats de les distàncies euclidianes entre cadascun dels vectors mitjana-individu i el punt que defineix el vector Ψ:

$$\|M_i - \Psi\|^2 \qquad \text{per a} \qquad 1 \leq i \leq h$$

18.Productes escalars dels vectors mitjana-individu

Es calcula el producte escalar entre cada un dels vectors mitjana-individu i tots els altres:

$$M_i \cdot M_k \qquad \text{per a} \qquad 1 \leq i < k \leq h$$

19.Coeficients angulars dels vectors mitjana-individu

Es calculen, entre cada un dels vectors mitjana-individu i tots els altres, el següents coeficients angulars:

$$\frac{M_i \cdot M_k}{\|M_i\| \, \|M_k\|} \qquad \text{per a} \qquad 1 \leq i < k \leq h$$

20.Covariàncies entre vectors mitjana-individu

Siguin els vectors mitjana-individu:

$$M_i = (m_{i1}, m_{i2}, m_{i3}, \ldots, m_{in}) \qquad \text{per a} \qquad 1 \leq i \leq h$$

I siguin:

$$\lambda_i = \frac{m_{i1} + m_{i2} + m_{i3} + \cdots + m_{in}}{n} \qquad \text{per a} \qquad 1 \leq i \leq h$$

les seves corresponents mitjanes.

Sigui també $\mathbf{1}$ el vector fila de n columnes amb totes les seves coordenades iguals a la unitat.

Es calcula la covariància entre cada un dels vectors mitjana-individu i tots els altres:

$$\frac{1}{n}(M_i - \mathbf{1}\lambda_i) \cdot (M_k - \mathbf{1}\lambda_k) \qquad \text{per a} \qquad 1 \leq i < k \leq h$$

on el signe $\cdot$ indica el producte escalar entre vectors ordinari.

21.Coeficients de correlació entre vectors mitjana-individu

Es calcula el coeficient de correlació entre cada un dels vectors mitjana-individu i tots els altres.

Amb la mateixa nomenclatura utilitzada a l'estadístic **20** anterior, tindríem:

$$\frac{(M_i - \mathbf{1}\lambda_i) \cdot (M_k - \mathbf{1}\lambda_k)}{\|M_i - \mathbf{1}\lambda_i\| \, \|M_k - \mathbf{1}\lambda_k\|} \qquad \text{per a} \qquad 1 \leq i < k \leq h$$

22.Productes escalars i mòduls entre vectors mitjana-individu

De igual manera que amb l'estadístic **18**, es calcula el producte escalar entre cada un dels vectors mitjana-individu i tots els altres. A més a més, s'afegeix el mòdul al quadrat de cada vector mitjana-individu (producte escalar entre cada vector mitjana-individu i ell mateix).

23.Covariàncies i variàncies entre vectors mitjana-individu

De igual manera que amb l'estadístic **20**, es calcula la covariància entre cada un dels vectors mitjana-individu i tots els altres. A més a més, s'afegeix la variància de cada vector mitjana-individu (covariància entre cada vector mitjana-individu i ell mateix).

24.Variàncies dels vectors mitjana-individu

Es calcula la variància de cada un dels vectors mitjana-individu:

$$\frac{1}{n}(M_i - \mathbf{1}\lambda_i) \cdot (M_i - \mathbf{1}\lambda_i) \qquad \text{per a} \qquad 1 \leq i \leq h$$

on el signe $\cdot$ indica el producte escalar entre vectors ordinari.

25.Desviacions modulars respecte el vector mitjana-variable

Primerament, es calcula el vector Ω, mitjana de tots els vectors variable, $V_1, V_2, V_3, \ldots, V_p$:

$$\Omega = \frac{1}{p}(V_1 + V_2 + V_3 + \cdots + V_p)$$

Posteriorment, es calculen les diferències, en valor absolut, entre el mòdul al quadrat de cadascun dels vectors mitjana-individu i el mòdul al quadrat del vector mitjana-variable Ω:

$$| \, \|M_i\|^2 - \|\Omega\|^2 \, | \qquad \text{per a} \qquad 1 \leq i \leq h$$

26.Desviacions vectorials respecte el vector mitjana-variable

Primerament, es calcula el vector Ω, mitjana de tots els vectors variable, $V_1, V_2, V_3, \ldots, V_p$, de la mateixa manera que a l'anterior estadístic **25**.

Posteriorment, es calculen els quadrats de les distàncies euclidianes entre cadascun dels vectors mitjana-individu i el punt que defineix el vector Ω:

$$\|M_i - \Omega\|^2 \qquad \text{per a} \qquad 1 \leq i \leq h$$

27.Diferència modular Ψ-Ω

Es calcula la diferència, en valor absolut, entre el mòdul al quadrat del vector Ψ, mitjana dels vectors mitjana-individu, $M_1, M_2, M_3, \ldots, M_h$, i el mòdul al quadrat del vector mitjana-variable Ω:

$$| \, \|\Psi\|^2 - \|\Omega\|^2 \, |$$

28.Diferència vectorial Ψ-Ω

Es calcula el quadrat de la distància euclidiana entre els vectors Ψ, mitjana dels vectors mitjana-individu, $M_1, M_2, M_3, \ldots, M_h$, i Ω, vector mitjana-variable:

$$\|\Psi - \Omega\|^2$$

29.Mòdul del vector Ψ

Es calcula el mòdul al quadrat del vector Ψ, mitjana dels vectors mitjana-individu, $M_1, M_2, M_3, \ldots, M_h$:

$$\|\Psi\|^2$$

Per a cadascuna de les particions del conjunt $\mathcal{F}(\mathbf{V})$, i per a cadascun d'aquests estadístics inter-grup (excepte per als tres estadístics 27, 28 i 29) **s'han computat, mitjançant implementació dels corresponents programes, els següents indicadors**:

1. El valor mínim que assoleix cada estadístic.

2. El valor màxim que assoleix cada estadístic.

3. La diferència entre els valors màxim i mínim que assoleix cada estadístic.

4. La suma dels valors màxim i mínim que assoleix cada estadístic.

5. La mitjana aritmètica de tots els valors de cada estadístic.

6. La mediana de tots els valors de cada estadístic.

7. La desviació típica de tots els valors de cada estadístic.

8.2.3 Notació i Simbologia

En aquest apartat introduïm la notació que usarem per representar cadascun dels resultats obtinguts pels estadístics referits en aquesta mateixa secció, perquè resultin més entenedores les conclusions del nostre estudi sobre la combinació de variables dels conjunts de la partició sobre la que s'executa el mètode DMM.

Cada estadístic es representarà pel número que precedeix la seva descripció.

Respecte dels estadístics intra-grup, representarem cada un dels indicadors intra-grup referits a la subsecció 8.2.1 pel número que encapçala la seva descripció.

Els tres indicadors inter-grup: mitjana aritmètica, mediana i desviació típica, corresponents a cada indicador intra-grup i referits també a la mateixa subsecció 8.2.1, es referenciaran pels números 0, 1 i 2 respectivament.

Pel que fa als estadístics inter-grup, els indicadors referits al final de la subsecció 8.2.2 també es representaran pel número que encapçala la seva descripció.

D'aquesta manera, simbolitzarem els indicadors de cadascun dels estadístics mitjançant el següent format: A-B , on A representa el número de l'estadístic, i B representa el número que identificarà el tipus d'indicador de la manera que tot seguit expliquem a través d'un esquema diversificat.

1r.Cas: $1 \leq A \leq 12$ Sempre que el número A es trobi entre 1 i 12 (ambdós inclosos), es tractarà d'un estadístic intra-grup i distingirem tres situacions diferents:

1. $1 \leq$ B ≤ 7. En aquest cas, el format A-B referencia l'indicador intra-grup B de l'estadístic A, corresponent al primer grup de variables de la partició (recordem que aquest estudi sempre considera particions de variables amb dos grups).
2. $8 \leq$ B ≤ 14. En aquest cas, el format A-B referencia l'indicador intra-grup B $- 7$ de l'estadístic A, corresponent al segon grup de variables de la partició.
3. $15 \leq$ B ≤ 35. En aquest cas, si B $= 3(4 + m) + r$, on $1 \leq m \leq 7$ i $0 \leq r \leq 2$, llavors, el format A-B referencia l'indicador inter-grup r de l'indicador intra-grup m de l'estadístic A.

2n.Cas: $13 \leq A \leq 26$ Cas que el número A es trobi entre 13 i 26 (ambdós inclosos), es tractarà d'un estadístic inter-grup i B només podrà ser un número entre 1 i 7 (ambdós inclosos també). Per la qual cosa, el format A-B referenciarà l'indicador B de l'estadístic A.

3r.Cas: $27 \leq A \leq 29$ Cas que el número A es trobi entre 27 i 29 (ambdós inclosos), es tractarà també d'un estadístic inter-grup i B forçosament serà el número 1 perquè aquests tres estadístics només tenen un únic indicador.

8.3 Resultats i conclusions de l'estudi sobre la combinació de variables

El nostre treball sobre les combinacions de variables de la partició sobre la que s'executa el mètode DMM, que optimitzen les diferents mesures de qualitat del referit mètode, ha estat basat en l'anàlisi de les correlacions entre cada un dels diferents indicadors dels estadístics descrits a l'anterior secció i cada una de les cinc mesures de qualitat definides al capítol 6 d'aquesta memòria:

- Pèrdua d'informació **PI**.

- Pèrdua de confidencialitat basada en enllaç de registres **ERD**.
- Pèrdua de confidencialitat basada en intervals sobre el número de registres **ICN**.
- Pèrdua de confidencialitat basada en intervals sobre la desviació típica **ICD**.
- Mesura global sobre la qualitat **MG**.

Les variants de microagregació multivariant analitzades han estat les sis variants referides a la secció 8.1 d'aquest capítol i representades com `Mic1-12mul`, `Mic2-11mul`, `Mic3-10mul`, `Mic4-9mul`, `Mic5-8mul` i `Mic6-7mul`.

Per tal de comprovar que una bona correlació entre un determinat indicador d'un estadístic i una determinada mesura de qualitat no depèn de la variant de microagregació aplicada, sinó que realment relaciona l'indicador de l'estadístic amb la mesura de qualitat referida, en una primera fase, per a cada mesura de qualitat i per a cada variant de microagregació, hem dividit totes les correlacions dels diferents indicadors pel seu valor absolut màxim, estandarditzant-les així respecte del seu valor absolut màxim, de manera que la màxima correlació de cada variant de microagregació es converteixi en un 1 o en -1 depenent de si la màxima correlació és positiva o negativa. Posteriorment, en una segona fase, hem sumat, per a cada mesura de qualitat i per a cada indicador, les correlacions estandarditzades de les sis variants de microagregació. La Taula 1 mostra aquestes sumes de correlacions estandarditzades que relacionen els indicadors estadístics amb les mesures de qualitat.

Les cinc següents subseccions estaran dedicades a l'anàlisi de cada una de les cinc mesures de qualitat anteriorment referides.

Nota1. Com que en el nostre estudi hem emprat l'Estandardització ST, cal observar que els resultats dels estadístics 1.*Mòduls de variables intra-grup*, 2.*Diferències entre mòduls de variables intra-grup* i 12.*Variàncies intra-grup*, per a qualsevol de les sis variants de microagregació considerades i per a qualsevol combinació del conjunt de variables, són constants; per la qual cosa, no té cap sentit considerar les seves correlacions amb les cinc mesures de qualitat considerades.

Nota2. Degut també a l'ús de l'Estandardització ST, cal observar que tots els estadístics de cada un dels agrupaments que tot seguit detallem tenen la mateixa correlació amb cadascuna de les cinc mesures de qualitat considerades, perquè els seus resultats només es diferencien en una constant que multiplica. Així doncs, només cal considerar un sol estadístic de cadascun dels següents agrupaments:

1. Estadístics 6.Productes escalars intra-grup, 7.Coeficients angulars intra-grup, 8.Covariàncies intra-grup i 9.Coeficients de correlació intra-grup.
2. Estadístics 10.Productes escalars i mòduls intra-grup i 11.Covariàncies i variàncies intra-grup.
3. Estadístics 13.Mòduls dels vectors mitjana-individu i 24.Variàncies dels vectors mitjana-individu.
4. Estadístics 18.Productes escalars dels vectors mitjana-individu i 20.Covariàncies entre vectors mitjana-individu.
5. Estadístics 19.Coeficients angulars dels vectors mitjana-individu i 21.Coeficients de correlació entre vectors mitjana-individu.
6. Estadístics 22.Productes escalars i mòduls entre vectors mitjana-individu i 23.Covariàncies i variàncies entre vectors mitjana-individu.

8.3.1 Combinació de variables i pèrdua d'informació

En l'anàlisi de la relació entre la combinació de variables i la pèrdua d'informació, cal diferenciar, en principi, les correlacions positives i les correlacions negatives dels diferents indicadors.

Correlacions negatives

Com podem observar a la Taula 1, els indicadors més ben correlacionats negativament amb la pèrdua d'informació són 5-11, 5-14 i 5-9, que tot seguit descrivim:

5-11. Correspon a l'indicador intra-grup "suma dels valors màxim i mínim" de l'estadístic **Desviacions vectorials intra-grup** del segon grup de la partició, i apareix a la Taula 1 amb una puntuació de -4.3080. A més a més, les seves correlacions amb les variants de microagregació `Mic3-10mul`, `Mic2-11mul` i `Mic1-12mul`, on, segons l'estudi del capítol anterior sobre el nombre de variables, s'assoleixen els valors més petits respecte a la pèrdua d'informació, són -0.3030, -0.3393 i -0.3334 respectivament.

5-14. Correspon a l'indicador intra-grup "desviació típica" de l'estadístic **Desviacions vectorials intra-grup** del segon grup de la partició, i apareix a la Taula 1 amb una puntuació de -4.0972. A més a més, les seves correlacions amb les variants de microagregació `Mic3-10mul`, `Mic2-11mul` i `Mic1-12mul` són -0.2347, -0.1141 i -0.2849 respectivament.

5-9. Correspon a l'indicador intra-grup "valor màxim" de l'estadístic **Desviacions vectorials intra-grup** del segon grup de la partició, i apareix a la Taula 1 amb una puntuació de -3.8057. Les seves correlacions amb les variants de microagregació `Mic3-10mul`, `Mic2-11mul` i `Mic1-12mul` són -0.2435, -0.1412 i -0.2574 respectivament.

Correlacions positives

Els indicadors més ben correlacionats positivament amb la pèrdua d'informació són 3-23, 6-23, 10-17, 10-23 i 10-26.

3-23, 6-23. Corresponen a l'indicador inter-grup "desviació típica" de l'indicador intra-grup "diferència entre els valors màxim i mínim" dels estadístics **Distàncies entre variables intra-grup** i **Productes escalars intra-grup** respectivament. Donades les característiques d'aquests indicadors, respecte la puntuació reflectida per la Taula 1, només té sentit sumar les cinc correlacions estandarditzades corresponents a les variants de microagregació `Mic2-11mul`, `Mic3-10mul`, `Mic4-9mul`, `Mic5-8mul` i `Mic6-7mul`, acumulant un total de 3.2966. Les seves correlacions amb les variants de microagregació `Mic3-10mul` i `Mic2-11mul` són 0.1718 i 0.1938 respectivament.

10-17, 10-23 i 10-26. Corresponen tots tres a l'indicador inter-grup "desviació típica" dels indicadors intra-grup "valor mínim", "diferència entre els valors màxim i mínim" i "suma dels valors màxim i mínim" de l'estadístic **Productes escalars i mòduls intra-grup**, ja que $17 = 3(4+1)+2$, $23 = 3(4+3)+2$ i $26 = 3(4+4)+2$; i apareixen a la Taula 1 amb una puntuació de 3.0809. Les seves correlacions amb les variants de microagregació `Mic3-10mul` i `Mic1-12mul` són 0.1301 i 0.2961 respectivament.

Considerant aquests resultats, podem concloure que l'indicador que millor correlaciona la pèrdua d'informació és el 5-11, que correspon a l'indicador intra-grup "suma dels valors màxim i mínim" de l'estadístic **Desviacions vectorials intra-grup** del segon grup de la partició, i té una correlació negativa.

8.3.2 Combinació de variables i ERD-pèrdua de confidencialitat

En l'anàlisi de la relació entre la combinació de variables i la ERD-pèrdua de confidencialitat caldrà diferenciar també les correlacions positives i les correlacions negatives dels diferents indicadors.

Correlacions negatives

Com podem observar també a la Taula 1, els indicadors més ben correlacionats negativament amb la ERD-pèrdua de confidencialitat són 5-8, 18-1, 18-2, 18-4, 18-5 i 18-6, que tot seguit descrivim:

5-8. Correspon a l'indicador intra-grup "valor mínim" de l'estadístic **Desviacions vectorials intra-grup** del segon grup de la partició, i apareix a la Taula 1 amb una puntuació de -5.9999. A més a més, les seves correlacions amb les variants de microagregació `Mic5-8mul` i `Mic6-7mul`, on, segons l'estudi del capítol anterior sobre el nombre de variables, s'assoleixen els dos valors més petits respecte a la ERD-pèrdua de confidencialitat, són -0.4130 i -0.2097 respectivament.

18-1, 18-2, 18-4, 18-5 i 18-6. Corresponen als indicadors "valor mínim", "valor màxim", "suma dels valors màxim i mínim", "mitjana" i "mediana" respectivament de l'estadístic **Productes escalars dels vectors mitjana-individu**, i apareixen tots ells a la Taula 1 amb una puntuació de -4.6582. Les seves correlacions amb les variants de microagregació `Mic5-8mul` i `Mic6-7mul` són -0.3084 i -0.1768 respectivament.

Correlacions positives

Com mostra la Taula 1, els indicadors més ben correlacionats positivament amb la ERD-pèrdua de confidencialitat són 10-13, 10-12 i 6-12, que tot seguit descrivim:

10-13. Correspon a l'indicador intra-grup "mediana" de l'estadístic **Productes escalars i mòduls intra-grup** del segon grup de la partició, i apareix a la Taula 1 amb una puntuació de 4.5991. Les seves correlacions amb les variants de microagregació `Mic5-8mul` i `Mic6-7mul` són 0.3215 i 0.0786 respectivament.

10-12, 6-12. Corresponen a l'indicador intra-grup "mitjana" dels estadístics **Productes escalars i mòduls intra-grup** i **Productes escalars intra-grup**, respectivament, del segon grup de la partició, i apareixen a la Taula 1 amb una puntuació de 4.3804. Les seves correlacions amb les variants de microagregació `Mic5-8mul` i `Mic6-7mul` són 0.3026 i 0.1229 respectivament.

Considerant tots aquests resultats, podem concloure que l'indicador que millor correlaciona la ERD-pèrdua de confidencialitat és el 5-8, que correspon a l'indicador intra-grup "valor mínim" de l'estadístic **Desviacions vectorials intra-grup** del segon grup de la partició, i té una correlació negativa.

8.3.3 Combinació de variables i ICN-pèrdua de confidencialitat

En l'anàlisi de la relació entre la combinació de variables i la ICN-pèrdua de confidencialitat diferenciarem també les correlacions positives i les correlacions negatives dels diferents indicadors.

Correlacions negatives

Com podem observar també a la Taula 1, els indicadors més ben correlacionats negativament amb la ICN-pèrdua de confidencialitat són 19-1, 19-2, 19-4, 19-5, 19-6, 5-8, 18-1, 18-2, 18-4, 18-5 i 18-6, que tot seguit descrivim:

19-1, 19-2, 19-4, 19-5 i 19-6. Corresponen als indicadors "valor mínim", "valor màxim", "suma dels valors màxim i mínim", "mitjana" i "mediana" respectivament de l'estadístic **Coeficients angulars dels vectors mitjana-individu**, i apareixen tots ells a la Taula 1 amb una mateixa puntuació de -5.6422. Les seves correlacions amb les variants de microagregació `Mic3-10mul`, `Mic4-9mul` i `Mic5-8mul`, on, segons l'estudi del capítol anterior sobre el nombre de variables, s'assoleixen els valors més petits respecte a la ICN-pèrdua de confidencialitat, són -0.6556, -0.6339 i -0.5904 respectivament.

5-8. Correspon a l'indicador intra-grup "valor mínim" de l'estadístic **Desviacions vectorials intra-grup** del segon grup de la partició, i apareix a la Taula 1 amb una puntuació de -5.2933. Les seves correlacions amb les variants de microagregació `Mic3-10mul`, `Mic4-9mul` i `Mic5-8mul` són -0.6913, -0.6453 i -0.5167 respectivament.

18-1, 18-2, 18-4, 18-5 i 18-6. Corresponen als indicadors "valor mínim", "valor màxim", "suma dels valors màxim i mínim", "mitjana" i "mediana" respectivament de l'estadístic **Productes escalars dels vectors mitjana-individu**, i apareixen tots ells a la Taula 1 amb una mateixa puntuació de -5.2645. Les seves correlacions amb les variants de microagregació `Mic3-10mul`, `Mic4-9mul` i `Mic5-8mul` són -0.6222, -0.5960 i -0.5286 respectivament.

Correlacions positives

Com observem també a la Taula 1, els indicadors més ben correlacionats positivament amb la ICN-pèrdua de confidencialitat són 26-1, 26-2, 26-3, 26-4, 26-5, 26-6, 26-7, 15-1, 15-2, 15-4, 15-5 i 15-6.

26-1, 26-2, 26-3, 26-4, 26-5, 26-6 i 26-7. Corresponen a tots set indicadors de l'estadístic **Desviacions vectorials respecte el vector mitjana-variable**, i apareixen tots ells a la Taula 1 amb una puntuació de 5.0973. Les seves correlacions amb les variants de microagregació `Mic3-10mul`, `Mic4-9mul` i `Mic5-8mul` són 0.5795, 0.5370 i 0.5182 respectivament.

15-1, 15-2, 15-4, 15-5 i 15-6. Corresponen als indicadors "valor mínim", "valor màxim", "suma dels valors màxim i mínim", "mitjana" i "mediana" respectivament de l'estadístic **Distàncies entre vectors mitjana-individu**, i s'obtenen els mateixos resultats que amb l'anterior estadístic **26**.

Considerant tots aquests resultats, podem concloure que els indicadors que millor correlacionen la ICN-pèrdua de confidencialitat són 19-1, 19-2, 19-4, 19-5 i 19-6, que es corresponen amb el "valor mínim", "valor màxim", "suma dels valors màxim i mínim", "mitjana" i "mediana" respectivament de l'estadístic **Coeficients angulars dels vectors mitjana-individu** i tenen correlació negativa.

8.3.4 Combinació de variables i ICD-pèrdua de confidencialitat

En l'anàlisi de la relació entre la combinació de variables i la ICD-pèrdua de confidencialitat diferenciarem també les correlacions positives i les correlacions negatives dels diferents indicadors.

Correlacions negatives

Com podem observar també a la Taula 1, els indicadors més ben correlacionats negativament amb la ICD-pèrdua de confidencialitat són 19-1, 19-2, 19-4, 19-5, 19-6, 18-1, 18-2, 18-4, 18-5, 18-6 i 5-8, que tot seguit descrivim:

19-1, 19-2, 19-4, 19-5 i 19-6. Corresponen als indicadors "valor mínim", "valor màxim", "suma dels valors màxim i mínim", "mitjana" i "mediana" respectivament de l'estadístic **Coeficients angulars dels vectors mitjana-individu**, i apareixen tots ells a la Taula 1 amb una mateixa puntuació de -5.9153. Les seves correlacions amb les variants de microagregació `Mic3-10mul`, `Mic4-9mul` i `Mic5-8mul`, on, segons l'estudi del capítol anterior sobre el nombre de variables, s'assoleixen els valors més petits respecte a la ICD-pèrdua de confidencialitat, són -0.7215, -0.6534 i -0.5833 respectivament.

18-1, 18-2, 18-4, 18-5 i 18-6. Corresponen als indicadors "valor mínim", "valor màxim", "suma dels valors màxim i mínim", "mitjana" i "mediana" respectivament de l'estadístic **Productes escalars dels vectors mitjana-individu**, i apareixen tots ells a la Taula 1 amb una mateixa puntuació de -5.5353. Les seves correlacions amb les variants de microagregació `Mic3-10mul`, `Mic4-9mul` i `Mic5-8mul` són -0.6848, -0.6153 i -0.5259 respectivament.

5-8. Correspon a l'indicador intra-grup "valor mínim" de l'estadístic **Desviacions vectorials intra-grup** del segon grup de la partició, i apareix a la Taula 1 amb una puntuació de -4.9112. Les seves correlacions amb les variants de microagregació `Mic3-10mul`, `Mic4-9mul` i `Mic5-8mul` són -0.6667, -0.6040 i -0.4768 respectivament.

Correlacions positives

Com observem també a la Taula 1, els indicadors més ben correlacionats positivament amb la ICD-pèrdua de confidencialitat són 26-1, 26-2, 26-3, 26-4, 26-5, 26-6, 26-7, 15-1, 15-2, 15-4, 15-5 i 15-6.

26-1, 26-2, 26-3, 26-4, 26-5, 26-6 i 26-7. Corresponen a tots set indicadors de l'estadístic **Desviacions vectorials respecte el vector mitjana-variable**, i apareixen tots ells a la Taula 1 amb una mateixa puntuació de 5.3951. Les seves correlacions amb les variants de microagregació `Mic3-10mul`, `Mic4-9mul` i `Mic5-8mul` són 0.6571, 0.5537 i 0.5097 respectivament.

15-1, 15-2, 15-4, 15-5 i 15-6. Corresponen als indicadors "valor mínim", "valor màxim", "suma dels valors màxim i mínim", "mitjana" i "mediana" respectivament de l'estadístic **Distàncies entre vectors mitjana-individu**, i s'obtenen els mateixos resultats que amb l'anterior estadístic **26**.

Considerant aquests resultats, podem concloure que els indicadors que millor correlacionen la ICD-pèrdua de confidencialitat són 19-1, 19-2, 19-4, 19-5 i 19-6, que es corresponen amb el "valor mínim", "valor màxim", "suma dels valors màxim i mínim", "mitjana" i "mediana" respectivament de l'estadístic **Coeficients angulars dels vectors mitjana-individu** i tenen correlació negativa.

8.3.5 Combinació de variables i mesura global MG sobre la qualitat

En l'anàlisi de la relació entre la combinació de variables i la mesura global **MG**, cal diferenciar també les correlacions positives i les correlacions negatives dels diferents indicadors.

Correlacions negatives

Com podem observar a la Taula 1, els indicadors més ben correlacionats negativament amb la mesura global **MG** són 5-11, 5-14, 22-1 i 5-9, que tot seguit descrivim:

5-11. Correspon a l'indicador intra-grup "suma dels valors màxim i mínim" de l'estadístic **Desviacions vectorials intra-grup** del segon grup de la partició, i apareix a la Taula

1 amb una puntuació de -4.6111. A més a més, les seves correlacions amb les variants de microagregació `Mic3-10mul`, `Mic2-11mul` i `Mic1-12mul`, on, segons l'estudi del capítol anterior sobre el nombre de variables, s'assoleixen els valors més petits respecte a la mesura global **MG**, són -0.3347, -0.3514 i -0.3354 respectivament.

5-14. Correspon a l'indicador intra-grup "desviació típica" de l'estadístic **Desviacions vectorials intra-grup** del segon grup de la partició, i apareix a la Taula 1 amb una puntuació de -3.7656. A més a més, les seves correlacions amb les variants de microagregació `Mic3-10mul`, `Mic2-11mul` i `Mic1-12mul` són -0.2200, -0.1013 i -0.2582 respectivament.

22-1. Correspon a l'indicador "valor mínim" de l'estadístic **Productes escalars i mòduls entre vectors mitjana-individu**, i apareix a la Taula 1 amb una puntuació de -3.5142. Les seves correlacions amb les variants de microagregació `Mic3-10mul`, `Mic2-11mul` i `Mic1-12mul` són -0.0956, -0.1443 i -0.1927 respectivament.

5-9. Correspon a l'indicador intra-grup "valor màxim" de l'estadístic **Desviacions vectorials intra-grup** del segon grup de la partició, i apareix a la Taula 1 amb una puntuació de -3.3174. Les seves correlacions amb les variants de microagregació `Mic3-10mul`, `Mic2-11mul` i `Mic1-12mul` són -0.2045, -0.0956 i -0.1875 respectivament.

Correlacions positives

Els indicadors més ben correlacionats positivament amb la mesura global **MG** són 3-32 i 6-32.

3-32, 6-32. Corresponen a l'indicador inter-grup "desviació típica" de l'indicador intra-grup "mediana" dels estadístics **Distàncies entre variables intra-grup** i **Productes escalars intra-grup** respectivament, ja que $32 = 3(4+6)+2$. Donades les característiques d'aquests indicadors, respecte la puntuació reflectida per la Taula 1, només té sentit sumar les cinc correlacions estandarditzades corresponents a les variants de microagregació `Mic2-11mul`, `Mic3-10mul`, `Mic4-9mul`, `Mic5-8mul` i `Mic6-7mul`, acumulant un total de 3.1373. La seva correlació amb la variant de microagregació `Mic3-10mul` és 0.1791.

Considerant aquests resultats, podem concloure que l'indicador que millor correlaciona la mesura global **MG** sobre la qualitat és el 5-11, que correspon a l'indicador intra-grup "suma dels valors màxim i mínim" de l'estadístic **Desviacions vectorials intra-grup** del segon grup de la partició, i té una correlació negativa.

Capítol 9

Conclusions i recerca futura

En aquest darrer capítol i com a cloenda d'aquesta memòria de tesi, exposarem les conclusions més significatives de tot el treball realitzat i proposarem possibles línies d'ampliació i futura recerca.

9.1 Conclusions

Podem resumir de la següent manera les conclusions més significatives de tot el treball realitzat:

- Hem provat que l'atac mitjançant estadístics d'ordre contra la microagregació amb ordenació individual, és bastant efectiu per a dades uniformement distribuïdes, per a dades normals i per a dades esbiaixades (Weibull). Tot i això, hem vist que els intervals de probabilitat posterior obtinguts per a dades normals i dades Weibull produeixen més reducció de l'amplitud sobre els intervals trivials ($[a_{-1}, a_1]$,$[a_{-1}, a_0]$,$[a_0, a_1]$) que els obtinguts per a dades uniformes. Per tant, l'atac és encara més efectiu quan les dades originals són normals, o bé, Weibull; les quals, a la vegada, són distribucions que s'ajusten molt bé a les dades financeres contínues ordinàries.

- Els intervals de probabilitat posterior obtinguts per als estadístics d'ordre extrems (els valors més petits o bé els valors més grans) són especialment més estrets que els seus corresponents intervals trivials. Això implica que l'atac és molt més efectiu quan s'aplica per estimar els valors extrems; la qual cosa és especialment preocupant dins del marc de la confidencialitat estadística, on habitualment s'aplica la microagregació amb ordenació individual, puix que una estimació prou precisa d'un valor extrem, fa que sigui molt fàcil la identificació de l'individu a qui pertany.

- Hem provat que el número de distàncies calculades i la complexitat computacional de l'algorisme DM (Mètode de microagregació de la "Distància Màxima"), sense emmagatzemar la matriu de distàncies, són funcions polinòmiques de tercer grau respecte del número de registres n que es microagreguen. Tanmateix, hem vist que el número de distàncies calculades i la complexitat computacional de l'algorisme DMM (Mètode de microagregació de la "Distància Màxima Modificat"), sense emmagatzemar la matriu de distàncies, són funcions polinòmiques quadràtiques respecte del número de registres n que es microagreguen. Per tant, en el nou mètode de microagregació DMM hi ha una reducció de la complexitat computacional molt considerable respecte del mètode de microagregació DM existent.

- De l'estudi comparatiu sobre la qualitat entre els sis mètodes de control de la revelació pertorbatius més rellevants, actualment existents, per a la protecció de microdades contínues,

destaquem que el mètode d'Intercanvi de dades (*Rank swapping*) és el que presenta la millor puntuació (més petita) de la mesura global **MG** sobre la qualitat. En segon lloc es troba la microagregació multivariant considerant totes les variables simultàniament, encara que amb molt poca diferència de puntuació respecte del rank swapping (27.65 per al millor `rank`p i 28.24 per a la millor `Micmul`k). Tot i que el rank swapping té una puntuació lleugerament millor, `Micmul`k té l'avantatge sobre el `rank`p de ser un mètode determinista i, per tant, reproductible sense pèrdua de seguretat. Aquesta característica és important en un context de bases de dades estadístiques "on-line" perquè successives modificacions de les mateixes dades originals, mitjançant `rank`p o qualsevol altre mètode estocàstic, produïrien diferents conjunts modificats de dades, que podrien conduir a la revelació de dades.

Respecte dels mètodes de microagregació multivariant sense projectar les dades, cal destacar que a mesura que augmenta la fragmentació del conjunt de les variables (cardinal més petit dels conjunts que formen la partició sobre la qual s'aplica l'algorisme DMM), disminueix la pèrdua d'informació d'aquests mètodes; però el risc de revelació manifesta una tendència contrària, puix que en augmentar la fragmentació del conjunt de les variables, augmenta també el risc de revelació (augment especialment manifestat per ERD i ICD).

Els mètodes de microagregació multivariant mitjançant projecció de les dades `MicZ`k i `MicPCP`k produeixen una pèrdua d'informació alta.

- Després del càlcul del número de particions i la implementació del programa mitjançant el qual ha estat possible l'anàlisi exhaustiva de la qualitat de les deu variants de microagregació multivariant `Mic1-12mul`, `Mic2-11mul`, `Mic3-10mul`, `Mic4-9mul`, `Mic5-8mul`, `Mic6-7mul`, `Mic4mul`, `Mic3mul`, `Mic2mul` i `Micmul` sobre totes les corresponents particions del conjunt de variables, hem observat el següent comportament de la qualitat d'aquests mètodes respecte del número de conjunts que formen la partició sobre la que s'executa l'algorisme de microagregació DMM:
  - L'augment del valor del paràmetre k i la menor fragmentació del conjunt de variables són dos factors que disminueixen la pèrdua de confidencialitat. A més a més, suposant particions formades per un mateix número de conjunts i ordenant-les per la diferència (en valor absolut) entre els cardinals dels seus conjunts en ordre descendent, en disminuir el valor absolut d'aquesta diferència també van decreixent els valors mínims de pèrdua de confidencialitat (tendència molt marcada en la ERD-pèrdua de confidencialitat i només perceptible per a valors grans del paràmetre k en la ICN i la ICD pèrdues de confidencialitat).
  - La disminució del valor del paràmetre k i la major fragmentació del conjunt de variables són dos factors que disminueixen la pèrdua d'informació. A més a més, suposant particions formades per un mateix número de conjunts, ordenades per la diferència (en valor absolut) entre els cardinals dels seus conjunts, generalment tindran menys pèrdua d'informació les particions que ocupen les posicions centrals.
  - Cal destacar que l'augment dels valors del paràmetre k i la menor fragmentació del conjunt de variables són dos factors que generalment milloren els valors de la mesura global **MG** sobre la qualitat (valors més petits). A més a més, suposant particions formades per un mateix número de conjunts, ordenades per la diferència (en valor absolut) entre els cardinals dels seus conjunts, els millors valors de la mesura global **MG** s'obtenen per a les particions que ocupen les posicions centrals (`Mic3-10mul`).
- Respecte la recerca de les combinacions de variables dintre els dos conjunts de la partició sobre la que s'executa l'algorisme DMM, que produeixen conjunts modificats de dades amb una bona qualitat, destacarem el comportament de l'estadístic **5.Desviacions vectorials intra-grup**, ja que els valors del seu indicador intra-grup 5-11 ("suma dels valors màxim i mínim" del segon grup de la partició) són els que millor correlacionen amb la mesura global **MG** sobre la qualitat

i té una correlació negativa. A més a més, el mateix indicador intra-grup 5-11 resulta ser també el que millor correlaciona negativament amb la pèrdua d'informació. I cal afegir que l'indicador intra-grup 5-8 ("valor mínim" del segon grup de la partició) del mateix estadístic **5** es troba entre els que millor correlacionen també negativament amb la pèrdua de confidencialitat.

9.2 Ampliacions i futura recerca

Si bé tot aquest treball realitzat millora alguns aspectes del mètode de microagregació, no pretén en absolut acabar la recerca sobre aquesta mateixa temàtica, sinó més bé obrir noves vies de futures millores que potser es nodriran d'algun resultat aconseguit i referit en aquesta memòria. En aquest sentit, a continuació oferim algunes orientacions:

- Una primera línia de possible ampliació de la tasca desenvolupada seria estudiar el comportament dels 29 estadístics descrits al capítol 8 d'aquesta memòria respecte a la relació entre combinació de variables dintre els conjunts de la partició sobre la que s'executa l'algorisme DMM i bona qualitat del conjunt modificat de dades obtingut, a partir d'estandarditzacions de les dades originals diferents a l'estandardització ST (tal com apareix referenciada al mateix capítol 8) emprada en el nostre estudi; altres estandarditzacions que podrien ser, per exemple, les quatre allí referides: estandardització SX, estandardització SN, estandardització SM i estandardització SD.
- Hem desenvolupat l'estudi sobre la combinació de variables més adequada dels conjunts de la partició base de l'algorisme DMM, per a particions del conjunt de variables en dos grups. Es pot ampliar aquest estudi amb les mateixes 13 variables i usant els mateixos 29 estadístics descrits al capítol 8, per a particions del conjunt de variables amb grups de 4, 3 o 2 variables (`Mic4mul`, `Mic3mul` o `Mic2mul` respectivament).
- També es pot treballar en posteriors modificacions de mètodes de microagregació multivariant existents perquè puguin incorporar en els seus respectius algorismes nous aspectes, potser evidenciats per l'actual recerca, que els condueixin més eficientment a resultats amb més qualitat.
- Utilitzant el mètode de Txolesky de descomposició d'una matriu de covariància, es pot treballar en la creació d'un mètode de control de la revelació estadística pertorbatiu per a microdades contínues de manera que el conjunt modificat de dades conservi les covariàncies del conjunt original.

Apèndix A

Taules (comparació de mètodes pertorbatius)

Mètode	PI	ERD	ICN	ICD	MG	PI Rank	ERD Rank	ICN Rank	ICD Rank
rank07	13,56	37,33	66,55	25,76	27,65	37	147	145	102
rank14	34,47	14,75	40,23	16,11	27,97	88	58	9	16
rank12	31,40	17,75	45,12	17,98	28,03	80	93	18	28
mict_16	30,54	14,39	49,68	25,31	28,24	78	51	61	99
rank06	10,42	42,43	71,88	28,99	28,43	31	152	162	118
rank08	20,57	31,88	61,22	22,81	28,76	54	140	129	77
rank10	27,74	23,80	51,64	19,97	28,77	70	121	81	53
mict_10	28,80	19,44	55,21	24,98	29,29	75	103	105	94
rank17	39,34	12,92	36,71	14,98	29,36	107	31	5	11
rank09	26,39	27,22	55,79	21,09	29,61	64	132	110	61
rank11	32,55	21,32	48,04	18,72	29,95	83	111	44	39
mic6_19	39,00	11,32	47,30	21,98	30,99	106	11	33	73
mic6_03	23,51	24,25	67,26	38,70	31,06	59	125	149	145
rank15	41,15	15,03	38,90	15,59	31,14	112	62	8	13
rank13	38,84	16,96	43,13	17,14	31,19	105	83	12	23
mic6_08	34,88	14,74	53,83	26,85	31,21	92	57	99	106
mict_15	36,94	15,28	50,02	23,55	31,49	100	64	64	85
mict_18	38,30	10,95	50,49	28,08	31,71	104	10	71	114
mic5_18	41,18	13,31	46,03	20,13	32,19	114	38	22	54
rank18	45,58	12,35	35,88	14,85	32,22	138	24	4	9
mic5_15	41,52	11,49	47,48	21,64	32,27	118	13	36	67
mict_11	35,55	18,77	52,73	26,07	32,32	94	100	93	103
mic6_04	30,63	20,17	62,33	33,55	32,35	79	105	133	132
mic4_17	35,83	16,90	53,04	29,16	32,41	97	82	96	120
mict_08	33,87	21,80	55,61	25,64	32,54	86	113	109	101
rank05	14,32	48,77	76,22	32,59	32,95	39	160	170	130
mic4_11	32,23	21,24	58,45	33,99	32,98	81	110	120	133
mict_19	43,20	10,56	48,40	22,61	33,11	125	8	47	75
mic6_05	35,28	17,24	58,91	30,49	33,12	93	88	122	125
mic4_04	18,19	36,64	72,33	46,94	33,16	49	145	163	161
mic6_14	42,57	11,43	48,98	23,19	33,17	122	12	53	81
mic3_15	28,52	24,91	61,28	40,28	33,18	74	126	130	148
mic3_14	27,64	26,31	62,35	40,97	33,31	69	129	134	150
rank04	8,73	55,19	81,94	39,40	33,33	28	169	180	147
mic3_08	18,50	36,75	70,85	48,70	33,38	50	146	158	167
mic5_20	44,50	12,34	45,31	19,25	33,41	130	23	20	46

Taula A.1: Mesures generals obtingudes en aplicar els diversos mètodes (1/5).

A. Taules (comparació de mètodes pertorbatius)

Mètode	PI	ERD	ICN	ICD	MG	PI Rank	ERD Rank	ICN Rank	ICD Rank
mic3_13	27,09	27,42	63,01	41,66	33,49	67	134	138	152
mict_12	38,09	20,32	51,56	23,61	33,52	103	107	79	87
mic3_20	33,16	21,36	57,69	35,17	33,53	84	112	117	137
mic5_17	43,52	13,15	46,55	21,45	33,55	127	34	25	64
mic3_06	14,86	41,39	74,36	52,36	33,62	44	151	167	173
mic4_06	25,89	28,96	66,26	41,51	33,66	63	135	144	151
mict_05	34,81	23,08	57,35	26,92	33,71	90	116	116	107
mic6_11	41,48	14,62	50,44	24,24	33,73	117	55	70	89
mic3_07	17,11	38,92	72,66	51,07	33,75	47	148	165	171
mict_13	41,08	14,11	52,28	25,29	33,76	111	47	89	98
mic3_09	21,42	34,14	69,67	46,99	33,83	57	144	156	162
mic6_12	42,78	13,47	49,41	23,19	33,83	123	40	58	82
mic5_14	43,74	12,70	48,70	21,98	33,88	128	27	50	72
mic6_20	44,71	12,22	46,81	21,22	33,92	132	21	28	63
mic4_19	39,53	16,83	52,07	27,52	33,92	108	81	83	112
rank19	49,13	12,62	35,23	14,46	33,93	150	26	2	7
mic5_05	35,58	20,85	58,86	29,45	34,04	95	108	121	121
mic3_12	26,81	29,54	63,48	43,30	34,14	66	136	139	153
mic5_03	26,80	31,36	67,13	36,56	34,20	65	138	148	140
mic4_07	29,04	27,14	64,40	38,79	34,21	77	131	140	146
mic3_18	32,32	23,85	59,26	37,55	34,23	82	123	123	143
mic6_09	41,64	15,73	52,73	25,07	34,48	119	70	92	95
mict_17	41,17	17,12	49,98	27,17	34,51	113	84	63	110
mic4_05	24,13	33,77	69,03	44,16	34,66	60	143	154	155
mic3_05	14,18	44,48	76,85	54,84	34,67	38	154	171	176
mic6_16	45,99	11,83	47,98	22,73	34,79	139	16	42	76
rank20	50,89	12,96	34,76	14,41	34,83	154	32	1	5
mic5_08	41,99	16,79	53,29	25,55	35,05	120	80	97	100
mic3_16	33,50	23,81	60,69	38,46	35,10	85	122	127	144
mic5_13	46,15	13,51	47,80	22,00	35,17	140	41	40	74
mic3_17	34,72	23,36	59,41	37,46	35,31	89	118	124	141
mic4_10	35,62	22,67	60,58	34,75	35,40	96	115	126	135
mic3_11	27,98	31,03	66,18	44,30	35,56	72	137	143	156
mic3_04	12,01	49,43	79,70	58,31	35,62	34	161	175	179
mic5_11	45,54	14,99	49,45	23,55	35,64	137	60	59	86
mic4_15	41,33	18,41	54,19	30,22	35,82	116	99	101	123
mic4_08	33,91	25,52	62,53	37,53	35,84	87	128	136	142
mic6_18	48,75	11,52	47,28	21,88	35,90	148	14	31	70
mic3_19	36,41	23,48	58,23	36,41	35,90	99	120	119	138
mic6_10	45,01	14,96	52,48	24,85	35,91	135	59	90	92
mic6_17	48,39	12,72	47,99	21,95	36,12	147	28	43	71
mic3_10	27,98	31,90	68,19	45,54	36,18	71	141	153	160
mic4_03	18,05	44,05	77,18	52,41	36,24	48	153	172	174
mic5_12	47,12	14,68	49,20	23,43	36,31	142	56	54	84
mic4_13	41,23	19,31	55,42	31,65	36,33	115	102	108	127
mic6_15	49,20	11,88	48,32	22,93	36,48	151	17	46	78
mic6_13	48,13	13,45	49,48	23,19	36,51	146	39	60	79
mic5_06	43,04	19,10	56,02	26,59	36,62	124	101	111	105
mic5_19	50,26	13,93	44,43	19,96	36,66	153	43	15	51
mic4_12	41,04	20,08	56,09	33,03	36,68	110	104	112	131
rank16	53,03	14,09	37,88	15,74	36,74	157	46	6	14
mic3_03	9,33	55,16	83,61	63,26	36,81	30	168	182	184
mic4_09	37,19	23,90	61,72	36,45	36,84	101	124	132	139
rank03	7,13	64,44	87,88	50,30	36,95	27	179	186	169
mic4_18	44,82	17,22	52,80	29,16	36,96	133	86	94	119

Taula A.1: Mesures generals obtingudes en aplicar els diversos mètodes (2/5).

Mètode	PI	ERD	ICN	ICD	MG	PI Rank	ERD Rank	ICN Rank	ICD Rank
mic5_04	37,63	25,36	62,42	32,21	36,98	102	127	135	129
mict_03	34,85	31,79	62,93	30,72	37,08	91	139	137	126
mic6_06	44,96	16,65	56,34	28,24	37,21	134	78	113	115
mic4_14	43,85	18,17	55,26	31,78	37,35	129	97	106	128
mict_14	47,54	17,41	50,79	23,19	37,37	143	89	73	80
mic4_20	46,81	16,71	52,10	26,95	37,46	141	79	84	108
mic2_20	24,60	45,58	66,92	43,67	37,52	61	156	147	154
mic2_12	16,66	56,28	72,39	51,98	37,95	46	171	164	172
mic2_17	22,55	50,20	68,13	45,37	38,01	58	162	152	159
mic2_15	20,41	52,37	69,93	48,06	38,04	53	165	157	164
mic5_10	49,56	15,19	52,17	24,01	38,10	152	63	87	88
mic2_16	21,25	51,63	69,59	47,03	38,11	56	164	155	163
mict_07	43,50	23,41	57,00	28,29	38,26	126	119	114	116
mic2_14	19,45	54,26	71,08	49,39	38,35	52	167	159	168
mict_09	45,40	22,62	55,19	25,23	38,41	136	114	104	97
mic2_13	18,90	55,36	71,60	50,53	38,55	51	170	160	170
mic2_11	16,05	58,19	75,72	53,61	38,74	45	172	169	175
mic2_10	14,70	59,78	77,69	55,10	38,90	43	174	173	177
mic5_16	54,92	12,02	46,56	21,59	38,98	158	18	26	66
mic2_18	25,68	48,39	67,65	44,80	38,99	62	159	150	158
mic2_19	27,50	46,06	66,90	44,35	39,17	68	157	146	157
mic5_09	51,53	15,52	52,15	25,10	39,30	155	67	85	96
mic4_16	49,11	17,70	53,62	30,41	39,49	149	92	98	124
mic2_09	14,40	61,93	79,10	57,04	39,70	41	176	174	178
mic2_08	13,11	63,27	80,13	59,04	39,77	36	177	177	180
mict_20	54,99	13,23	49,30	23,22	39,87	159	35	55	83
mict_04	44,55	27,35	60,24	28,71	40,23	131	133	125	117
mic2_07	12,11	65,67	81,45	61,11	40,29	35	180	179	183
mic6_07	52,53	15,73	55,27	27,35	40,52	156	71	107	111
mic2_06	10,66	67,58	83,27	63,90	40,62	32	181	181	185
mic2_05	8,91	69,63	85,83	67,42	41,02	29	183	184	187
mic2_04	6,91	72,05	88,05	71,27	41,38	26	187	187	189
jpg055	56,43	20,20	50,35	17,27	41,71	163	106	69	25
mic2_03	5,31	74,80	91,06	76,60	42,31	24	189	190	191
mic5_07	55,62	17,79	54,51	26,08	42,33	160	94	103	104
rank02	10,73	71,77	93,29	64,08	42,98	33	185	193	186
mostreig1	3,16	78,36	96,76	78,81	43,12	20	190	202	192
rank01	2,57	80,21	98,72	81,28	43,84	15	193	205	193
jpg050	62,72	17,62	48,54	16,40	43,88	167	90	49	20
jpg060	59,38	23,27	51,90	18,10	44,26	166	117	82	30
micir_20	3,16	82,20	92,94	84,26	44,28	19	195	192	194
micir_19	3,35	82,61	92,85	84,81	44,53	21	196	191	195
micir_18	3,04	82,80	93,29	85,84	44,61	18	197	194	197
micir_17	2,71	83,31	93,38	86,74	44,70	16	198	195	198
micir_16	2,35	83,72	93,43	87,57	44,73	14	199	196	199
micir_15	2,31	83,92	93,62	88,50	44,90	13	200	197	200
mostreig3	2,97	82,13	98,36	85,13	44,95	17	194	204	196
micir_14	2,19	84,35	93,69	89,27	45,05	12	201	198	202
micir_13	2,00	84,64	93,97	90,08	45,17	11	203	199	203
micir_12	1,76	85,20	94,09	90,96	45,31	10	204	200	204
jpg040	68,39	14,23	46,23	15,29	45,44	169	49	23	12
micir_11	1,70	85,53	98,79	92,00	46,08	9	206	206	206
micir_10	1,56	85,60	98,85	92,91	46,15	8	207	207	207
micir_09	1,38	86,23	99,09	93,73	46,35	7	208	208	208
micir_08	1,14	86,80	99,24	94,63	46,50	6	209	210	209

Taula A.1: Mesures generals obtingudes en aplicar els diversos mètodes (3/5).

A. Taules (comparació de mètodes pertorbatius)

Mètode	PI	ERD	ICN	ICD	MG	PI Rank	ERD Rank	ICN Rank	ICD Rank
micir_07	0,99	87,06	99,33	95,48	46,61	4	210	211	210
jpg095	14,36	79,03	89,22	69,55	46,78	40	192	189	188
micir_06	1,01	87,62	99,46	96,40	46,89	5	212	212	212
soroll06	21,17	73,90	84,16	59,34	47,00	55	188	183	182
micir_05	0,82	88,27	99,53	97,05	47,05	3	213	213	213
soroll04	14,61	78,72	88,95	72,73	47,19	42	191	188	190
micir_04	0,70	89,07	99,67	97,85	47,31	2	214	214	214
jpg070	58,74	33,21	57,24	21,11	47,47	165	142	115	62
soroll02	6,89	84,50	94,42	88,87	47,48	25	202	201	201
soroll08	28,19	69,85	79,86	48,42	47,59	73	184	176	166
jpg100	4,70	85,49	99,14	91,82	47,59	23	205	209	205
micir_03	0,55	90,30	99,79	98,49	47,63	1	215	215	215
jpg035	74,37	12,61	45,29	14,44	47,80	176	25	19	6
jpg090	29,00	69,17	80,87	48,15	47,92	76	182	178	165
soroll01	4,21	87,30	97,03	96,24	48,09	22	211	203	211
soroll12	40,83	58,55	71,63	34,80	48,35	109	173	161	136
jpg080	48,08	50,30	66,15	28,03	48,39	145	163	142	113
micp_03	70,51	15,79	52,84	20,73	48,40	171	72	95	57
soroll10	36,04	63,80	75,22	40,34	48,42	98	178	168	149
jpg075	55,76	40,30	61,38	24,40	48,68	162	150	131	91
jpg045	73,46	16,08	47,75	15,88	48,71	175	73	39	15
micz_03	71,73	17,18	52,26	21,66	49,40	172	85	88	68
jpg085	42,08	60,65	72,78	34,69	49,64	121	175	166	134
micp_07	75,87	14,48	50,23	19,10	50,22	177	52	67	44
jpg030	82,48	10,07	44,39	14,47	51,12	179	6	14	8
soroll16	55,71	48,25	64,45	27,17	51,37	161	158	141	109
mict_06	73,01	21,03	54,48	24,90	51,68	174	109	102	93
micp_12	81,04	13,97	47,34	16,64	52,01	178	44	35	21
jpg065	72,14	27,13	54,17	19,50	52,06	173	130	100	47
soroll20	68,98	39,38	58,04	21,75	54,31	170	149	118	69
soroll14	57,36	54,25	67,71	30,05	54,46	164	166	151	122
micp_16	87,26	12,06	46,83	16,76	54,59	182	19	29	22
micp_06	85,85	14,05	49,32	18,88	54,96	180	45	56	41
soroll18	66,66	44,58	61,17	24,26	55,15	168	155	128	90
micz_19	87,75	13,57	46,73	18,74	55,45	184	42	27	40
micz_09	87,18	15,00	50,58	21,08	56,30	181	61	72	60
micp_08	88,09	15,66	50,03	18,56	56,53	185	68	65	34
micp_05	88,27	15,67	51,05	19,09	56,82	186	69	75	43
micz_18	90,27	14,52	47,29	17,58	56,87	187	53	32	27
micz_05	87,43	18,25	51,32	19,96	57,19	183	98	78	52
micz_13	93,88	12,24	48,83	19,66	58,56	189	22	51	48
micz_08	91,16	17,88	51,61	20,44	59,06	188	95	80	56
jpg025	100,20	10,60	42,96	13,98	59,87	193	9	11	4
distrib	47,66	71,94	87,34	59,18	60,13	144	186	185	181
micz_11	94,65	17,22	49,40	21,58	60,50	190	87	57	65
micp_04	95,27	16,56	52,16	20,17	60,82	191	77	86	55
micz_04	97,60	17,62	52,54	20,91	62,38	192	91	91	58
micp_13	102,24	12,75	47,33	18,90	62,58	194	29	34	42
micz_15	106,10	11,71	47,68	17,33	64,10	195	15	38	26
micz_20	109,21	12,96	47,93	19,12	66,23	198	33	41	45
micz_06	107,26	16,27	51,21	19,94	66,59	196	75	76	50
micp_09	108,97	14,52	50,30	18,45	66,71	197	54	68	32
jpg020	118,07	7,96	40,71	12,72	67,71	203	3	10	3
micz_17	114,64	13,27	47,66	18,62	68,92	199	37	37	37
micp_15	119,21	9,72	45,91	16,25	69,80	204	5	21	19

Taula A.1: Mesures generals obtingudes en aplicar els diversos mètodes (4/5).

Mètode	PI	ERD	ICN	ICD	MG	PI Rank	ERD Rank	ICN Rank	ICD Rank
micp_14	115,16	16,40	48,09	18,60	70,02	200	76	45	35
micz_16	116,54	14,18	48,96	18,49	70,24	201	48	52	33
micz_12	117,99	14,29	48,53	18,69	70,97	202	50	48	38
micp_11	124,36	12,91	46,42	17,20	73,36	205	30	24	24
jpg015	134,34	7,76	38,28	11,85	75,38	207	2	7	2
micp_17	137,05	10,25	44,37	14,95	78,50	208	7	13	10
micz_10	131,91	15,34	51,03	21,03	78,80	206	65	74	59
micp_10	143,56	13,23	49,87	18,38	83,62	210	36	62	31
micz_14	142,77	16,19	50,09	18,60	84,02	209	74	66	36
jpg010	162,28	6,68	35,51	11,22	88,65	213	1	3	1
micp_18	156,39	12,14	44,85	16,15	88,85	212	20	16	17
micz_07	155,66	18,08	51,26	19,77	91,23	211	96	77	49
micp_19	172,42	8,00	45,01	16,16	95,86	214	4	17	18
micp_20	183,15	15,34	47,11	18,09	103,56	215	66	30	29

Taula A.1: Mesures generals obtingudes en aplicar els diversos mètodes (5/5).

Mètode	PI1	PI2	PI3	PI4	PI5	PI
rank07	30,32	0,00	0,00	17,88	2,84	13,56
rank14	82,61	0,00	0,00	35,28	6,32	34,47
rank12	79,53	0,00	0,00	24,37	5,00	31,40
mict_16	73,19	0,00	18,27	13,04	5,56	30,54
rank06	23,83	0,00	0,00	12,95	1,91	10,42
rank08	52,12	0,00	0,00	16,62	2,55	20,57
rank10	69,34	0,00	0,00	23,66	4,10	27,74
mict_10	72,21	0,00	14,22	10,02	4,17	28,80
rank17	93,08	0,00	0,00	41,73	8,17	39,34
rank09	69,39	0,00	0,00	16,53	3,04	26,39
rank11	84,88	0,00	0,00	21,36	4,21	32,55
mic6_19	99,25	0,00	13,17	19,07	3,27	39,00
mic6_03	65,07	0,00	2,97	6,91	1,06	23,51
rank15	105,48	0,00	0,00	29,42	6,52	41,15
rank13	102,92	0,00	0,00	22,11	5,09	38,84
mic6_08	94,55	0,00	7,33	10,84	1,99	34,88
mict_15	92,40	0,00	17,12	14,72	5,00	36,94
mict_18	95,26	0,00	18,80	15,13	5,36	38,30
mic5_18	106,30	0,00	13,21	18,01	3,25	41,18
rank18	110,02	0,00	0,00	44,24	9,22	45,58
mic5_15	109,18	0,00	11,93	15,77	3,06	41,52
mict_11	91,93	0,00	14,90	10,06	4,48	35,55
mic6_04	84,15	0,00	4,21	9,94	1,36	30,63
mic4_17	93,91	0,00	8,52	16,44	2,16	35,83
mict_08	88,21	0,00	12,39	10,72	3,66	33,87
rank05	34,23	0,00	0,00	15,70	1,75	14,32
mic4_11	85,96	0,00	6,09	13,62	1,73	32,23
mict_19	111,52	0,00	19,44	11,12	5,60	43,20
mic6_05	97,91	0,00	5,26	9,02	1,58	35,28
mic4_04	50,45	0,00	2,51	4,83	0,89	18,19
mic6_14	111,27	0,00	10,61	19,35	2,94	42,57
mic3_15	75,27	0,00	5,93	12,92	1,72	28,52
mic3_14	72,52	0,00	5,63	13,61	1,55	27,64

Taula A.2: Mesures de pèrdua d'informació obtingudes en aplicar els diversos mètodes (1/5).

Mètode	PI1	PI2	PI3	PI4	PI5	PI
rank04	20,34	0,00	0,00	10,50	1,18	8,73
mic3_08	47,58	0,00	3,44	11,33	1,06	18,50
mic5_20	115,21	0,00	14,25	18,82	3,53	44,50
mic3_13	72,70	0,00	5,36	10,42	1,38	27,09
mict_12	99,80	0,00	15,17	9,17	4,62	38,09
mic3_20	88,54	0,00	7,57	12,19	2,13	33,16
mic5_17	115,71	0,00	12,62	13,85	3,25	43,52
mic3_06	39,63	0,00	2,67	6,43	0,79	14,86
mic4_06	71,14	0,00	3,68	8,19	1,20	25,89
mict_05	94,91	0,00	9,09	7,36	2,61	34,81
mic6_11	110,65	0,00	9,29	15,69	2,61	41,48
mic3_07	45,54	0,00	3,03	7,66	0,92	17,11
mict_13	107,50	0,00	16,21	10,53	4,71	41,08
mic3_09	55,28	0,00	3,81	12,96	1,16	21,42
mic6_12	113,36	0,00	9,92	17,42	2,64	42,78
mic5_14	115,48	0,00	11,43	16,98	3,07	43,74
mic6_20	117,25	0,00	13,37	17,16	3,25	44,71
mic4_19	106,11	0,00	9,21	13,48	2,30	39,53
rank19	126,23	0,00	0,00	33,21	9,12	49,13
mic5_05	97,51	0,00	5,50	11,27	1,72	35,58
mic3_12	71,49	0,00	5,02	11,63	1,26	26,81
mic5_03	75,13	0,00	3,23	6,34	1,01	26,80
mic4_07	79,77	0,00	4,22	9,17	1,31	29,04
mic3_18	86,15	0,00	6,82	12,90	1,93	32,32
mic6_09	112,90	0,00	8,00	13,81	2,26	41,64
mict_17	105,73	0,00	18,40	11,78	5,37	41,17
mic4_05	66,74	0,00	3,12	7,08	1,10	24,13
mic3_05	37,66	0,00	2,22	6,89	0,64	14,18
mic6_16	122,63	0,00	11,72	15,87	3,08	45,99
rank20	128,15	0,00	0,00	39,24	9,80	50,89
mic5_08	115,63	0,00	7,65	10,69	2,32	41,99
mic3_16	89,24	0,00	6,24	14,48	1,82	33,50
mic5_13	125,76	0,00	10,89	11,40	3,09	46,15
mic3_17	93,16	0,00	6,51	13,65	1,82	34,72
mic4_10	96,68	0,00	5,75	12,95	1,68	35,62
mic3_11	75,05	0,00	4,71	11,81	1,28	27,98
mic3_04	32,21	0,00	1,68	5,43	0,54	12,01
mic5_11	121,76	0,00	9,76	17,03	2,94	45,54
mic4_15	112,46	0,00	7,86	13,20	1,98	41,33
mic4_08	94,44	0,00	4,71	8,45	1,45	33,91
mic6_18	129,11	0,00	12,63	18,58	3,08	48,75
mic3_19	98,32	0,00	7,18	12,71	1,94	36,41
mic6_10	122,43	0,00	8,56	14,12	2,48	45,01
mic6_17	129,07	0,00	12,18	16,94	3,11	48,39
mic3_10	74,70	0,00	4,25	13,01	1,20	27,98
mic4_03	50,93	0,00	1,63	4,18	0,64	18,05
mic5_12	129,49	0,00	10,25	10,62	2,85	47,12
mic4_13	114,05	0,00	7,09	10,39	1,82	41,23
mic6_15	131,60	0,00	11,20	17,85	2,95	49,20
mic6_13	129,82	0,00	10,23	16,11	2,78	48,13
mic5_06	120,13	0,00	6,27	9,82	1,88	43,04
mic5_19	133,31	0,00	13,58	18,09	3,26	50,26
mic4_12	111,58	0,00	6,51	14,67	1,87	41,04
rank16	139,02	0,00	0,00	32,96	7,17	53,03
mic3_03	24,87	0,00	1,13	4,64	0,48	9,33

Taula A.2: Mesures de pèrdua d'informació obtingudes en aplicar els diversos mètodes (2/5).

Mètode	PI1	PI2	PI3	PI4	PI5	PI
mic4_09	102,04	0,00	5,26	12,17	1,62	37,19
rank03	16,27	0,00	0,00	9,18	1,05	7,13
mic4_18	122,51	0,00	8,75	12,89	2,27	44,82
mic5_04	106,92	0,00	4,37	6,12	1,44	37,63
mict_03	99,01	0,00	5,69	3,77	1,63	34,85
mic6_06	126,47	0,00	6,03	9,02	1,75	44,96
mic4_14	120,88	0,00	7,51	11,89	1,93	43,85
mict_14	125,54	0,00	16,60	12,52	5,06	47,54
mic4_20	125,97	0,00	9,54	16,95	2,47	46,81
mic2_20	65,66	0,00	4,72	10,44	1,11	24,60
mic2_12	45,37	0,00	2,96	5,49	0,76	16,66
mic2_17	60,64	0,00	4,12	8,85	1,02	22,55
mic2_15	55,32	0,00	3,66	7,21	0,92	20,41
mic5_10	135,32	0,00	9,07	15,14	2,53	49,56
mic2_16	56,75	0,00	3,91	9,10	0,98	21,25
mict_07	119,49	0,00	11,60	7,15	3,26	43,50
mic2_14	52,07	0,00	3,42	8,21	0,95	19,45
mict_09	121,85	0,00	13,29	11,49	3,94	45,40
mic2_13	51,45	0,00	3,22	6,44	0,83	18,90
mic2_11	43,32	0,00	2,72	6,24	0,71	16,05
mic2_10	39,92	0,00	2,46	5,20	0,73	14,70
mic5_16	149,43	0,00	12,24	15,27	3,12	54,92
mic2_18	69,77	0,00	4,33	9,19	1,02	25,68
mic2_19	73,95	0,00	4,52	11,54	1,06	27,50
mic5_09	142,47	0,00	8,43	13,43	2,39	51,53
mic4_16	134,29	0,00	8,18	15,88	2,05	49,11
mic2_09	38,41	0,00	2,21	6,72	0,66	14,40
mic2_08	35,51	0,00	1,95	5,18	0,53	13,11
mict_20	144,44	0,00	19,54	15,87	5,65	54,99
mict_04	125,82	0,00	7,51	5,84	2,35	44,55
mic2_07	32,74	0,00	1,63	5,14	0,43	12,11
mic6_07	146,70	0,00	6,73	13,04	1,99	52,53
mic2_06	28,00	0,00	1,40	6,06	0,48	10,66
mic2_05	23,23	0,00	1,17	5,37	0,44	8,91
mic2_04	18,35	0,00	0,83	3,62	0,31	6,91
jpg055	128,78	6,92	7,91	60,82	5,33	56,43
mic2_03	13,25	0,00	0,52	4,55	0,27	5,31
mic5_07	158,06	0,00	7,11	8,39	2,11	55,62
rank02	25,23	0,00	0,00	13,21	0,70	10,73
mostreig1	3,23	1,06	5,57	5,53	0,36	3,16
rank01	4,91	0,00	0,00	5,16	0,42	2,57
jpg050	142,04	7,76	8,57	69,67	6,23	62,72
jpg060	146,80	5,96	6,98	45,07	4,69	59,38
micir_20	6,13	0,00	1,95	4,38	0,35	3,16
micir_19	5,85	0,00	1,85	6,21	0,33	3,35
micir_18	4,98	0,00	1,75	6,20	0,34	3,04
micir_17	4,71	0,00	1,65	4,86	0,31	2,71
micir_16	4,43	0,00	1,55	3,45	0,27	2,35
micir_15	4,12	0,00	1,45	3,84	0,31	2,31
mostreig3	2,32	1,59	6,01	5,36	0,20	2,97
micir_14	3,84	0,00	1,33	3,82	0,33	2,19
micir_13	3,56	0,00	1,24	3,41	0,23	2,00
micir_12	3,18	0,00	1,13	2,90	0,17	1,76
jpg040	147,79	9,43	9,02	88,59	7,70	68,39
micir_11	2,82	0,00	1,02	3,27	0,25	1,70

Taula A.2: Mesures de pèrdua d'informació obtingudes en aplicar els diversos mètodes (3/5).

Mètode	PI1	PI2	PI3	PI4	PI5	PI
micir_10	2,61	0,00	0,90	3,00	0,24	1,56
micir_09	2,27	0,00	0,79	2,79	0,17	1,38
micir_08	1,98	0,00	0,66	2,04	0,18	1,14
micir_07	1,76	0,00	0,53	1,74	0,14	0,99
jpg095	40,23	0,83	0,21	4,36	0,30	14,36
micir_06	1,51	0,00	0,43	2,50	0,14	1,01
soroll06	61,43	0,15	0,33	3,43	0,25	21,17
micir_05	1,30	0,00	0,34	1,88	0,13	0,82
soroll04	42,76	0,08	0,20	1,71	0,13	14,61
micir_04	1,04	0,00	0,24	1,72	0,13	0,70
jpg070	154,95	4,42	6,09	28,75	3,28	58,74
soroll02	19,97	0,05	0,09	1,18	0,06	6,89
soroll08	81,20	0,20	0,62	5,47	0,42	28,19
jpg100	12,72	1,21	0,07	1,41	0,08	4,70
micir_03	0,76	0,00	0,11	1,57	0,08	0,55
jpg035	152,74	10,61	10,90	110,00	9,22	74,37
jpg090	84,41	0,64	0,59	3,39	0,57	29,00
soroll01	12,37	0,01	0,05	0,44	0,03	4,21
soroll12	118,64	0,31	1,37	5,22	0,77	40,83
jpg080	131,96	2,59	2,32	18,19	1,46	48,08
micp_03	172,08	0,00	27,90	39,69	11,31	70,51
soroll10	104,43	0,19	0,64	5,97	0,60	36,04
jpg075	144,31	3,52	4,03	36,11	2,28	55,76
jpg045	163,01	8,14	9,30	90,28	7,03	73,46
micz_03	143,41	0,00	26,64	102,89	14,03	71,73
jpg085	120,27	1,58	1,12	8,29	0,98	42,08
micp_07	172,27	0,00	36,03	56,94	17,72	75,87
jpg030	141,21	13,08	13,02	175,21	11,16	82,48
soroll16	160,62	0,55	2,67	8,37	1,43	55,71
mict_06	208,29	0,00	10,38	7,91	3,17	73,01
micp_12	185,77	0,00	38,25	55,52	20,94	81,04
jpg065	190,11	4,91	7,67	36,22	3,85	72,14
soroll20	199,62	0,29	3,47	8,89	1,99	68,98
soroll14	167,20	0,25	2,18	6,36	0,98	57,36
micp_16	202,32	0,00	39,26	56,61	23,06	87,26
micp_06	209,22	0,00	34,52	45,92	16,22	85,85
soroll18	193,76	0,45	3,13	7,02	1,82	66,66
micz_19	153,14	0,00	39,50	148,24	32,49	87,75
micz_09	162,43	0,00	36,11	137,43	24,66	87,18
micp_08	212,84	0,00	36,28	48,50	18,08	88,09
micp_05	211,22	0,00	33,50	58,27	15,39	88,27
micz_18	160,98	0,00	39,24	148,46	31,95	90,27
micz_05	173,79	0,00	32,35	125,10	19,56	87,43
micz_13	176,99	0,00	37,65	143,97	27,66	93,88
micz_08	177,68	0,00	34,87	133,86	22,86	91,16
jpg025	143,75	14,33	14,34	271,18	13,83	100,20
distrib	11,09	9,54	197,19	54,89	2,17	47,66
micz_11	182,48	0,00	37,26	138,93	26,74	94,65
micp_04	244,88	0,00	31,26	37,02	13,60	95,27
micz_04	210,44	0,00	30,50	116,61	17,61	97,60
micp_13	249,29	0,00	38,85	53,91	22,08	102,24
micz_15	211,03	0,00	38,63	145,71	30,21	106,10
micz_20	218,81	0,00	39,64	145,49	32,50	109,21
micz_06	233,66	0,00	32,72	123,36	20,13	107,26

Taula A.2: Mesures de pèrdua d'informació obtingudes en aplicar els diversos mètodes (4/5).

Mètode	PI1	PI2	PI3	PI4	PI5	PI
micp_09	274,22	0,00	37,13	49,31	18,95	108,97
jpg020	136,10	17,12	18,48	384,31	16,33	118,07
micz_17	236,99	0,00	38,97	144,44	30,48	114,64
micp_15	297,24	0,00	39,11	58,71	22,96	119,21
micp_14	284,54	0,00	38,83	60,85	22,23	115,16
micz_16	241,83	0,00	38,76	146,81	29,99	116,54
micz_12	250,53	0,00	37,66	141,54	27,66	117,99
micp_11	315,89	0,00	38,13	55,53	20,71	124,36
jpg015	125,68	23,26	22,99	489,65	18,80	134,34
micp_17	349,39	0,00	39,37	60,58	23,59	137,05
micz_10	295,77	0,00	36,45	138,14	25,34	131,91
micp_10	372,04	0,00	37,56	59,77	19,98	143,56
micz_14	322,94	0,00	38,14	143,83	28,78	142,77
jpg010	112,13	34,15	29,62	662,40	23,23	162,28
micp_18	410,51	0,00	39,61	53,55	24,14	156,39
micz_07	373,22	0,00	34,14	131,49	21,90	155,66
micp_19	456,92	0,00	39,54	57,28	23,86	172,42
micp_20	487,78	0,00	39,81	58,73	24,80	183,15

Taula A.2: Mesures de pèrdua d'informació obtingudes en aplicar els diversos mètodes (5/5).

Mètode	ERD-1	ERD-2	ERD-3	ERD-4	ERD-5	ERD-6	ERD-7	ERD
rank07	0,09	3,61	20,56	34,63	57,31	71,30	73,80	37,33
rank14	0,19	1,94	5,93	9,91	20,19	29,54	35,56	14,75
rank12	0,37	2,41	7,13	11,57	25,19	35,83	41,76	17,75
mict_16	0,00	1,48	5,93	8,89	25,19	28,15	31,11	14,39
rank06	0,00	4,35	28,06	44,63	66,11	75,56	78,33	42,43
rank08	0,00	3,06	16,94	27,50	47,69	61,85	66,11	31,88
rank10	0,56	2,69	11,57	19,17	33,43	46,76	52,41	23,80
mict_10	0,00	2,78	11,11	16,67	27,78	38,89	38,89	19,44
rank17	0,09	1,57	4,81	8,06	17,31	27,41	31,20	12,92
rank09	0,19	2,78	11,94	23,61	39,81	54,54	57,69	27,22
rank11	0,37	1,94	8,06	15,28	31,48	43,61	48,52	21,32
mic6_19	0,37	1,85	7,69	10,65	15,00	18,70	25,00	11,32
mic6_03	1,02	6,11	20,19	23,80	32,31	36,48	49,81	24,25
rank15	0,28	1,02	5,56	10,19	20,65	31,57	35,93	15,03
rank13	0,19	2,22	6,57	11,76	24,63	33,52	39,81	16,96
mic6_08	1,02	4,81	11,85	13,80	18,80	24,17	28,70	14,74
mict_15	0,00	5,56	9,72	8,33	22,22	29,17	31,94	15,28
mict_18	0,00	1,67	8,33	6,67	15,00	21,67	23,33	10,95
mic5_18	0,19	2,78	8,24	11,67	16,30	26,57	27,41	13,31
rank18	0,46	1,20	5,65	10,09	16,30	24,07	28,70	12,35
mic5_15	0,00	1,48	7,87	10,00	13,52	21,76	25,83	11,49
mict_11	0,00	3,06	16,30	17,31	28,52	31,57	34,63	18,77
mic6_04	0,74	6,39	15,00	20,09	25,09	31,57	42,31	20,17
mic4_17	0,83	3,15	9,17	10,46	22,78	33,43	38,52	16,90
mict_08	0,74	5,93	9,63	17,78	31,85	43,70	42,96	21,80
rank05	0,19	5,83	38,98	57,31	73,43	82,22	83,43	48,77
mic4_11	0,56	4,35	10,19	11,57	29,17	42,59	50,28	21,24
mict_19	0,00	1,76	1,76	7,04	10,56	21,11	31,67	10,56

Taula A.3: Mesures de pèrdua de confidencialitat per enllaç de registres obtingudes en aplicar els diversos mètodes (1/5).

A. Taules (comparació de mètodes pertorbatius)

Mètode	ERD-1	ERD-2	ERD-3	ERD-4	ERD-5	ERD-6	ERD-7	ERD
mic6_05	1,20	3,61	13,15	15,19	23,33	27,04	37,13	17,24
mic4_04	1,02	6,11	19,91	24,07	55,00	72,04	78,33	36,64
mic6_14	0,19	2,41	9,26	9,17	15,00	19,17	24,81	11,43
mic3_15	0,28	2,59	7,78	17,04	36,11	51,85	58,70	24,91
mic3_14	0,83	2,87	8,33	19,63	37,87	54,81	59,81	26,31
rank04	0,00	8,61	49,26	67,78	83,24	88,43	88,98	55,19
mic3_08	0,83	4,35	12,96	31,48	56,76	73,24	77,59	36,75
mic5_20	0,28	1,67	7,78	10,83	14,91	22,31	28,61	12,34
mic3_13	0,65	3,15	9,35	19,81	40,00	55,93	63,06	27,42
mict_12	1,11	5,56	11,11	15,56	31,11	35,56	42,22	20,32
mic3_20	0,37	2,69	6,94	12,96	30,46	44,44	51,67	21,36
mic5_17	0,56	3,06	9,44	9,44	14,17	24,63	30,74	13,15
mic3_06	0,83	5,00	16,57	38,15	66,48	79,54	83,15	41,39
mic4_06	0,93	4,81	14,35	16,76	41,76	58,89	65,19	28,96
mict_05	1,39	5,09	14,81	21,76	34,72	42,59	41,20	23,08
mic6_11	0,09	3,33	12,59	13,80	19,72	23,61	29,17	14,62
mic3_07	0,56	4,91	14,44	32,87	61,85	76,39	81,39	38,92
mict_13	0,00	1,20	8,43	13,24	20,46	26,57	28,89	14,11
mic3_09	0,56	3,24	12,04	27,87	51,57	68,89	74,81	34,14
mic6_12	0,37	4,26	10,65	11,30	16,48	22,78	28,43	13,47
mic5_14	0,09	2,78	8,43	10,00	15,37	23,33	28,89	12,70
mic6_20	0,83	2,78	9,35	12,69	16,02	21,02	22,87	12,22
mic4_19	0,37	2,50	9,07	9,63	21,94	34,35	39,91	16,83
rank19	0,09	1,57	5,19	10,09	16,67	25,65	29,07	12,62
mic5_05	0,83	3,52	13,33	16,11	24,44	40,19	47,50	20,85
mic3_12	0,46	3,52	9,81	21,67	44,07	60,00	67,22	29,54
mic5_03	1,02	7,59	23,70	27,59	34,72	58,06	66,85	31,36
mic4_07	0,46	3,89	13,43	15,74	38,89	56,20	61,39	27,14
mic3_18	0,19	2,22	8,89	17,59	34,54	48,33	55,19	23,85
mic6_09	0,28	4,72	12,22	12,87	22,31	26,20	31,48	15,73
mict_17	0,00	4,72	8,70	11,02	30,74	32,31	32,31	17,12
mic4_05	0,74	6,20	18,06	20,00	50,46	67,13	73,80	33,77
mic3_05	0,74	5,46	19,35	43,43	70,93	84,63	86,85	44,48
mic6_16	0,19	1,85	7,78	9,91	17,31	20,19	25,56	11,83
rank20	0,37	1,94	5,83	10,19	16,02	25,56	30,83	12,96
mic5_08	1,20	3,89	11,57	12,50	18,70	31,30	38,33	16,79
mic3_16	0,74	1,48	7,31	15,46	34,17	50,09	57,41	23,81
mic5_13	0,28	2,41	6,85	10,65	15,65	27,13	31,57	13,51
mic3_17	0,09	1,85	7,13	15,93	34,07	49,35	55,09	23,36
mic4_10	0,83	5,37	10,37	12,13	30,46	46,30	53,24	22,67
mic3_11	0,65	3,33	11,02	24,17	47,22	62,22	68,61	31,03
mic3_04	0,74	7,22	23,98	51,85	79,35	90,28	92,59	49,43
mic5_11	0,83	2,78	11,11	11,76	17,04	28,61	32,78	14,99
mic4_15	0,28	3,24	9,81	9,72	25,65	37,41	42,78	18,41
mic4_08	1,20	3,33	12,59	14,07	36,20	52,59	58,61	25,52
mic6_18	0,09	1,30	6,76	10,74	16,11	19,91	25,74	11,52
mic3_19	1,11	3,43	8,70	17,41	32,78	48,15	52,78	23,48
mic6_10	0,19	3,52	11,94	13,80	19,72	24,54	31,02	14,96
mic6_17	0,19	2,87	8,70	10,74	18,43	22,04	26,11	12,72
mic3_10	0,65	2,96	11,85	25,19	48,33	64,63	69,72	31,90
mic4_03	1,48	7,69	26,85	31,76	69,72	83,43	87,41	44,05
mic5_12	1,20	2,59	9,91	10,93	17,87	27,96	32,31	14,68
mic4_13	0,37	3,15	9,44	10,19	25,65	39,72	46,67	19,31
mic6_15	0,46	2,31	8,89	11,76	15,09	20,65	23,98	11,88
mic6_13	0,09	1,85	12,13	13,61	18,70	21,57	26,20	13,45

Taula A.3: Mesures de pèrdua de confidencialitat per enllaç de registres obtingudes en aplicar els diversos mètodes (2/5).

Mètode	ERD-1	ERD-2	ERD-3	ERD-4	ERD-5	ERD-6	ERD-7	ERD
mic5_06	0,74	2,87	12,59	15,09	22,50	36,48	43,43	19,10
mic5_19	0,00	3,15	8,89	11,30	16,76	26,94	30,46	13,93
mic4_12	0,37	2,96	10,28	11,39	26,02	40,65	48,89	20,08
rank16	0,46	1,94	6,11	9,81	18,61	28,52	33,15	14,09
mic3_03	2,04	10,00	31,48	64,54	86,76	94,63	96,67	55,16
mic4_09	0,93	3,70	12,31	12,87	33,43	49,17	54,91	23,90
rank03	0,00	18,98	71,67	82,50	90,46	93,52	93,98	64,44
mic4_18	0,28	2,41	9,44	11,02	24,63	34,26	38,52	17,22
mic5_04	0,37	5,46	17,13	20,93	27,69	49,17	56,76	25,36
mict_03	1,11	5,00	24,17	30,83	48,33	55,83	57,22	31,79
mic6_06	0,93	3,33	12,13	16,48	22,59	27,41	33,70	16,65
mic4_14	0,37	3,89	9,26	9,72	23,15	37,04	43,80	18,17
mict_14	0,00	1,30	10,37	14,26	25,93	33,70	36,30	17,41
mic4_20	0,00	1,94	8,70	10,09	23,43	34,35	38,43	16,71
mic2_20	0,83	5,09	30,46	46,57	70,83	80,93	84,35	45,58
mic2_12	1,20	8,33	47,87	66,11	84,63	92,22	93,61	56,28
mic2_17	0,83	6,02	36,94	54,72	77,04	86,94	88,89	50,20
mic2_15	1,57	6,67	40,00	60,46	79,26	88,43	90,19	52,37
mic5_10	0,37	3,61	8,52	10,28	17,50	31,39	34,63	15,19
mic2_16	1,11	6,30	39,35	56,30	79,72	88,89	89,72	51,63
mict_07	1,30	4,54	12,96	20,09	35,19	46,20	43,61	23,41
mic2_14	1,11	7,04	43,89	63,52	82,96	89,72	91,57	54,26
mict_09	0,00	5,83	10,83	14,17	37,50	45,00	45,00	22,62
mic2_13	0,83	7,69	46,48	64,44	84,07	91,11	92,87	55,36
mic2_11	0,93	9,35	50,28	70,83	87,96	93,61	94,35	58,19
mic2_10	1,67	9,91	55,37	73,15	88,52	94,35	95,46	59,78
mic5_16	0,65	1,02	8,61	9,54	13,89	23,70	26,76	12,02
mic2_18	0,93	5,65	33,98	52,87	74,07	84,63	86,57	48,39
mic2_19	0,93	5,37	30,37	47,31	71,39	81,76	85,28	46,06
mic5_09	0,83	2,96	10,93	13,06	17,31	28,52	35,00	15,52
mic4_16	1,30	3,33	9,07	9,26	23,33	35,00	42,59	17,70
mic2_09	1,39	11,02	60,56	78,24	90,65	95,28	96,39	61,93
mic2_08	1,20	12,50	62,22	81,39	92,59	96,11	96,85	63,27
mict_20	0,00	3,70	5,56	5,56	22,22	25,93	29,63	13,23
mict_04	0,37	6,67	17,41	25,56	40,74	50,74	50,00	27,35
mic2_07	1,76	14,17	68,70	85,56	94,91	97,13	97,50	65,67
mic6_07	0,37	5,28	13,15	15,28	20,56	23,15	32,31	15,73
mic2_06	1,76	16,30	75,28	89,26	94,44	97,78	98,24	67,58
mic2_05	2,50	19,54	80,28	91,94	96,48	98,15	98,52	69,63
mic2_04	2,50	24,26	84,72	95,37	98,24	99,63	99,63	72,05
jpg055	1,02	5,28	13,24	18,89	25,74	36,85	40,37	20,20
mic2_03	2,96	32,13	91,48	98,06	99,17	99,91	99,91	74,80
mic5_07	0,00	5,00	11,94	15,09	20,83	32,87	38,80	17,79
rank02	0,00	34,44	85,93	93,61	95,46	96,20	96,76	71,77
mostreig1	5,37	50,56	94,44	98,98	99,44	99,81	99,91	78,36
rank01	0,00	72,22	96,20	98,06	98,15	98,33	98,52	80,21
jpg050	0,56	3,98	11,76	16,20	23,15	32,13	35,56	17,62
jpg060	0,93	3,89	16,57	22,96	31,11	41,94	45,46	23,27
micir_20	5,00	78,98	97,13	98,24	98,70	98,70	98,61	82,20
micir_19	5,09	80,93	97,78	98,33	98,70	98,70	98,70	82,61
micir_18	5,46	82,04	97,31	98,33	98,80	98,89	98,80	82,80
micir_17	5,74	85,37	97,04	98,24	98,89	98,89	98,98	83,31
micir_16	6,20	86,39	97,59	98,80	98,98	99,07	98,98	83,72
micir_15	6,67	87,04	98,06	98,61	98,98	98,98	99,07	83,92
mostreig3	6,85	70,56	97,78	99,72	100,00	100,00	100,00	82,13

Taula A.3: Mesures de pèrdua de confidencialitat per enllaç de registres obtingudes en aplicar els diversos mètodes (3/5).

Mètode	ERD-1	ERD-2	ERD-3	ERD-4	ERD-5	ERD-6	ERD-7	ERD
micir_14	7,13	88,89	98,33	98,89	99,07	99,07	99,07	84,35
micir_13	7,69	89,91	98,43	98,98	98,98	99,26	99,26	84,64
micir_12	8,33	91,48	98,52	99,35	99,44	99,63	99,63	85,20
jpg040	0,74	3,52	9,81	11,94	18,80	25,00	29,81	14,23
micir_11	8,98	92,50	99,07	99,35	99,54	99,63	99,63	85,53
micir_10	10,00	92,59	99,07	99,35	99,26	99,44	99,44	85,60
micir_09	11,02	94,91	99,44	99,54	99,54	99,54	99,63	86,23
micir_08	12,50	96,39	99,54	99,81	99,72	99,81	99,81	86,80
micir_07	14,17	97,04	99,44	99,72	99,63	99,72	99,72	87,06
jpg095	4,54	56,20	94,17	99,07	99,44	99,91	99,91	79,03
micir_06	16,67	97,78	99,72	99,72	99,81	99,81	99,81	87,62
soroll06	2,96	36,48	85,93	94,91	98,15	99,26	99,63	73,90
micir_05	19,91	98,70	99,81	99,81	99,81	99,91	99,91	88,27
soroll04	4,44	53,89	94,26	98,89	99,63	99,91	100,00	78,72
micir_04	25,00	98,70	99,91	99,91	100,00	100,00	100,00	89,07
jpg070	0,93	6,48	23,70	33,24	47,13	58,98	62,04	33,21
soroll02	8,43	83,52	99,72	99,91	99,91	100,00	100,00	84,50
soroll08	2,41	27,22	77,41	89,07	96,11	98,06	98,70	69,85
jpg100	8,15	90,37	99,91	100,00	100,00	100,00	100,00	85,49
micir_03	33,33	99,07	99,81	99,91	100,00	100,00	100,00	90,30
jpg035	0,65	2,69	7,41	10,56	14,91	24,63	27,41	12,61
jpg090	2,22	27,59	75,74	88,61	94,81	96,94	98,24	69,17
soroll01	17,13	94,07	99,91	100,00	100,00	100,00	100,00	87,30
soroll12	1,76	15,46	52,96	70,00	85,46	91,39	92,78	58,55
jpg080	1,85	14,44	41,94	59,07	72,59	80,09	82,13	50,30
micp_03	0,56	3,06	7,78	10,28	21,39	33,06	34,44	15,79
soroll10	1,85	19,72	63,70	79,72	90,65	94,72	96,20	63,80
jpg075	1,02	9,44	31,30	42,96	58,89	67,69	70,83	40,30
jpg045	0,74	3,61	11,11	14,54	21,02	28,80	32,78	16,08
micz_03	0,56	3,61	8,06	11,39	26,94	35,28	34,44	17,18
jpg085	1,85	18,70	57,50	75,37	86,57	91,48	93,06	60,65
micp_07	1,30	0,65	7,31	8,61	26,11	27,41	30,00	14,48
jpg030	0,83	2,41	6,39	8,33	11,67	18,33	22,50	10,07
soroll16	1,30	8,52	37,13	52,87	71,85	81,57	84,54	48,25
mict_06	0,00	3,89	15,00	17,78	29,44	40,56	40,56	21,03
micp_12	0,00	0,00	7,78	8,89	21,11	28,89	31,11	13,97
jpg065	0,46	6,02	18,33	25,56	37,78	48,70	53,06	27,13
soroll20	0,93	7,78	25,74	37,69	57,78	71,57	74,17	39,38
soroll14	1,48	11,48	45,19	62,31	80,28	88,70	90,28	54,25
micp_16	0,00	1,48	5,93	5,93	19,26	26,67	25,19	12,06
micp_06	0,56	2,22	7,22	12,78	20,00	28,89	26,67	14,05
soroll18	1,11	8,33	33,52	47,59	65,65	76,85	78,98	44,58
micz_19	0,00	0,00	7,04	12,31	21,11	26,39	28,15	13,57
micz_09	0,00	0,83	8,33	12,50	21,67	30,00	31,67	15,00
micp_08	0,00	4,44	8,15	11,85	22,22	32,59	30,37	15,66
micp_05	0,00	1,85	9,26	12,50	24,54	30,56	31,02	15,67
micz_18	0,00	1,67	5,00	13,33	25,00	26,67	30,00	14,52
micz_05	0,46	3,24	9,72	12,50	30,09	35,19	36,57	18,25
micz_13	0,00	2,41	2,41	6,02	19,26	25,37	30,19	12,24
micz_08	0,74	1,48	11,11	16,30	25,93	34,07	35,56	17,88
jpg025	0,46	2,22	6,11	9,17	13,33	20,19	22,69	10,60
distrib	2,59	35,56	83,33	93,52	96,39	96,11	96,11	71,94
micz_11	0,00	3,06	10,19	17,31	23,43	34,81	31,76	17,22
micp_04	0,74	1,85	7,41	8,15	27,04	35,19	35,56	16,56
micz_04	0,37	4,81	8,52	13,33	27,41	34,07	34,81	17,62

Taula A.3: Mesures de pèrdua de confidencialitat per enllaç de registres obtingudes en aplicar els diversos mètodes (4/5).

Mètode	ERD-1	ERD-2	ERD-3	ERD-4	ERD-5	ERD-6	ERD-7	ERD
micp_13	0,00	2,41	8,43	8,43	18,06	24,17	27,78	12,75
micz_15	1,39	1,39	2,78	8,33	15,28	26,39	26,39	11,71
micz_20	0,00	1,85	7,41	9,26	25,93	22,22	24,07	12,96
micz_06	0,00	5,00	7,78	11,11	26,67	30,56	32,78	16,27
micp_09	0,00	1,67	7,50	7,50	21,67	31,67	31,67	14,52
jpg020	0,37	1,94	4,81	6,85	8,61	16,30	16,85	7,96
micz_17	1,57	0,00	6,30	7,87	20,46	25,19	31,48	13,27
micp_15	0,00	1,39	6,94	4,17	13,89	20,83	20,83	9,72
micp_14	0,00	0,00	7,96	7,96	26,11	37,78	35,00	16,40
micz_16	0,00	1,48	4,44	10,37	25,19	29,63	28,15	14,18
micz_12	0,00	2,22	3,33	4,44	25,56	31,11	33,33	14,29
micp_11	0,00	2,04	6,30	11,39	14,44	27,69	28,52	12,91
jpg015	0,46	0,93	3,98	6,57	8,80	16,30	17,31	7,76
micp_17	0,00	3,98	4,72	7,87	11,02	22,87	21,30	10,25
micz_10	0,00	1,85	7,41	8,33	22,22	33,33	34,26	15,34
micp_10	0,00	3,70	1,85	8,33	20,37	30,56	27,78	13,23
micz_14	1,30	0,00	9,07	9,07	24,81	33,89	35,19	16,19
jpg010	0,56	1,30	3,61	4,81	7,50	14,07	14,91	6,68
micp_18	0,00	0,00	1,67	3,33	18,33	30,00	31,67	12,14
micz_07	0,65	2,59	9,72	14,26	27,41	36,30	35,65	18,08
micp_19	0,00	0,00	0,00	1,76	12,31	20,83	21,11	8,00
micp_20	0,00	5,56	5,56	9,26	20,37	35,19	31,48	15,34

Taula A.3: Mesures de pèrdua de confidencialitat per enllaç de registres obtingudes en aplicar els diversos mètodes (5/5).

Mètode	ICN-1	ICN-2	ICN-3	ICN-4	ICN-5	ICN-6	ICN-7	ICN-8	ICN-9	ICN-10	ICN
rank07	14,27	26,71	39,69	53,23	67,86	80,95	94,34	95,61	96,23	96,64	66,55
rank14	8,88	16,30	23,46	30,63	38,11	44,44	50,73	57,05	63,55	69,20	40,23
rank12	9,99	18,14	26,45	34,46	42,34	49,53	56,89	64,47	71,23	77,69	45,12
mict_16	10,88	29,00	37,32	43,19	52,02	56,52	60,68	65,93	70,09	71,11	49,68
rank06	16,44	31,34	46,87	63,19	80,04	94,64	95,78	96,48	96,85	97,18	71,88
rank08	12,86	23,90	35,13	47,25	58,91	69,77	82,23	92,86	94,30	95,04	61,22
rank10	11,05	20,24	29,10	38,20	47,21	55,80	65,19	74,74	83,67	91,25	51,64
mict_10	14,46	32,91	43,59	51,78	57,91	63,53	67,45	70,66	73,86	75,93	55,21
rank17	8,55	15,63	22,22	28,58	34,76	40,44	46,11	51,76	57,22	61,83	36,71
rank09	11,41	21,32	31,47	41,10	51,52	61,46	71,89	82,26	92,11	93,32	55,79
rank11	10,21	18,97	27,61	36,04	44,37	52,14	60,82	68,88	77,08	84,25	48,04
mic6_19	10,20	28,99	28,99	44,11	44,44	54,67	60,29	63,17	69,06	69,06	47,30
mic6_03	26,20	43,30	55,90	64,51	71,06	75,76	79,65	83,18	85,66	87,38	67,26
rank15	8,44	15,88	23,28	30,19	36,68	42,96	49,33	55,29	60,90	66,09	38,90
rank13	9,54	17,88	25,79	33,29	40,80	47,76	54,53	61,26	67,27	73,19	43,13
mic6_08	18,18	28,37	40,34	49,59	54,38	61,03	66,22	69,26	74,58	76,39	53,83
mict_15	8,44	25,00	39,74	39,74	50,64	60,58	64,32	66,45	71,05	74,25	50,02
mict_18	11,54	32,82	32,82	46,28	52,05	57,05	64,36	64,36	70,51	73,08	50,49
mic5_18	9,67	26,94	26,94	40,51	46,63	51,42	60,33	60,33	67,27	70,25	46,03
rank18	8,38	14,98	21,69	27,82	33,77	39,80	45,41	50,76	55,83	60,33	35,88
mic5_15	8,50	24,47	36,69	36,69	48,11	56,11	60,43	63,18	68,34	72,31	47,48
mict_11	15,67	28,92	40,95	49,38	52,39	60,22	64,37	69,09	72,22	74,10	52,73
mic6_04	20,50	38,51	49,42	59,34	65,58	70,42	75,18	78,63	81,72	84,04	62,33
mic4_17	13,13	34,68	34,99	48,39	57,94	58,33	65,30	70,83	71,46	75,33	53,04

Taula A.4: Mesures de pèrdua de confidencialitat per intervals de confiança, basats en rangs, obtingudes en aplicar diversos mètodes (1/5).

Mètode	ICN-1	ICN-2	ICN-3	ICN-4	ICN-5	ICN-6	ICN-7	ICN-8	ICN-9	ICN-10	ICN
mict_08	16,07	27,12	41,37	52,31	57,44	64,73	69,80	72,19	76,52	78,52	55,61
rank05	18,63	35,88	54,60	74,66	94,57	95,92	96,56	96,86	97,12	97,39	76,22
mic4_11	23,82	37,21	47,60	55,38	58,82	64,20	69,33	73,16	76,20	78,80	58,45
mict_19	12,18	30,43	30,43	45,29	45,29	55,28	61,47	63,75	69,95	69,95	48,40
mic6_05	18,68	35,06	45,48	54,03	61,20	67,51	71,57	75,80	78,59	81,20	58,91
mic4_04	34,41	52,08	61,77	70,18	75,89	80,31	83,93	86,25	88,53	89,98	72,33
mic6_14	8,06	24,28	37,17	47,05	47,12	55,10	62,04	67,39	69,78	71,82	48,98
mic3_15	18,99	42,96	54,59	54,59	63,22	70,07	72,84	75,01	78,86	81,72	61,28
mic3_14	18,61	42,18	53,95	62,91	63,00	69,24	74,27	78,11	79,79	81,47	62,35
rank04	23,28	45,95	70,25	94,91	96,55	97,19	97,45	97,69	97,97	98,13	81,94
mic3_08	39,57	49,46	60,09	68,58	72,54	77,57	81,63	83,75	87,03	88,24	70,85
mic5_20	9,71	28,27	28,27	42,54	42,54	53,96	53,96	62,31	62,31	69,25	45,31
mic3_13	17,63	41,65	53,81	62,28	68,65	71,12	73,65	77,58	80,63	83,15	63,01
mict_12	7,61	22,48	34,36	45,04	54,10	61,97	67,27	71,79	74,70	76,24	51,56
mic3_20	18,23	44,00	44,00	57,32	57,32	66,25	66,25	72,93	72,93	77,64	57,69
mic5_17	9,37	27,07	27,31	40,19	50,26	50,98	58,43	65,24	65,90	70,79	46,55
mic3_06	37,40	54,11	65,38	73,57	78,10	81,47	85,24	87,83	89,63	90,87	74,36
mic4_06	26,26	44,37	55,78	64,11	69,19	73,30	78,16	81,52	84,11	85,79	66,26
mict_05	15,31	31,05	42,84	52,60	59,47	66,63	71,15	75,68	78,13	80,66	57,35
mic6_11	14,88	27,49	38,02	46,70	50,52	56,53	61,59	66,08	69,52	73,08	50,44
mic3_07	39,17	52,41	61,75	71,25	75,28	80,78	82,83	85,90	87,81	89,34	72,66
mict_13	8,89	25,10	38,26	50,31	57,81	61,14	64,30	68,76	72,76	75,45	52,28
mic3_09	39,88	50,38	57,53	64,51	71,85	77,47	80,40	82,70	84,82	87,17	69,67
mic6_12	7,84	22,31	33,99	43,82	51,47	58,26	63,63	67,76	71,60	73,42	49,41
mic5_14	9,05	23,97	36,50	46,75	46,89	54,92	61,77	66,64	69,16	71,37	48,70
mic6_20	11,24	29,94	29,94	44,25	44,25	55,48	55,48	63,89	63,89	69,78	46,81
mic4_19	12,89	35,99	35,99	49,99	50,37	59,59	63,91	67,04	72,47	72,47	52,07
rank19	8,43	15,23	21,76	27,98	33,37	38,98	44,10	49,35	54,34	58,79	35,23
mic5_05	17,22	34,42	45,48	54,15	60,74	67,53	71,89	76,61	79,22	81,32	58,86
mic3_12	18,98	41,20	52,52	60,69	67,23	72,34	76,55	79,56	82,28	83,43	63,48
mic5_03	24,37	42,00	55,68	64,57	71,25	76,31	80,16	83,52	85,72	87,74	67,13
mic4_07	27,38	42,04	51,81	62,19	66,18	72,74	75,59	79,59	82,24	84,29	64,40
mic3_18	18,57	44,08	44,08	57,35	62,22	65,90	72,26	72,26	76,97	78,86	59,26
mic6_09	17,64	28,60	37,95	45,76	54,27	61,40	65,42	69,04	71,84	75,39	52,73
mict_17	13,26	30,64	31,13	44,17	53,80	54,22	62,22	68,45	68,94	72,93	49,98
mic4_05	31,36	48,26	58,06	65,53	71,55	77,32	80,52	83,84	86,18	87,67	69,03
mic3_05	42,48	58,50	67,43	74,49	79,84	84,77	87,52	89,86	91,22	92,43	76,85
mic6_16	9,25	26,75	34,22	40,21	50,24	55,06	58,77	64,72	70,03	70,58	47,98
rank20	8,40	15,04	21,35	27,75	33,42	38,57	43,17	48,64	53,45	57,86	34,76
mic5_08	17,17	26,13	39,15	49,10	54,35	61,13	66,11	69,17	74,35	76,24	53,29
mic3_16	19,04	42,47	50,22	55,27	64,07	67,99	71,24	75,97	80,14	80,51	60,69
mic5_13	7,53	21,62	33,22	42,59	51,50	55,61	59,10	64,44	69,30	73,08	47,80
mic3_17	17,58	42,72	43,15	56,00	64,36	64,95	71,72	76,52	77,04	80,02	59,41
mic4_10	21,65	41,23	50,16	57,14	63,10	68,03	71,81	74,97	77,68	79,99	60,58
mic3_11	34,53	46,87	56,25	63,63	66,89	71,97	76,05	79,42	82,11	84,06	66,18
mic3_04	45,46	61,84	71,59	79,22	83,70	87,14	89,94	91,45	92,81	93,81	79,70
mic5_11	14,09	25,86	36,25	45,19	49,10	55,24	60,80	65,66	69,65	72,72	49,45
mic4_15	12,76	33,40	46,08	46,08	55,61	62,98	66,32	68,65	73,32	76,70	54,19
mic4_08	27,34	38,41	50,56	58,97	63,81	69,86	74,72	77,25	81,40	82,96	62,53
mic6_18	9,72	28,07	28,07	41,86	48,61	53,73	61,78	61,78	67,98	71,19	47,28
mic3_19	16,81	43,84	43,84	56,77	57,24	66,06	70,06	72,75	77,45	77,45	58,23
mic6_10	14,17	31,58	40,32	47,81	54,30	59,79	64,05	67,71	71,03	73,98	52,48
mic6_17	10,11	27,32	27,81	41,45	52,46	52,91	60,82	67,28	67,83	71,95	47,99
mic3_10	33,96	51,12	58,90	65,47	70,55	74,71	77,93	80,87	83,12	85,23	68,19
mic4_03	41,04	56,23	68,49	76,00	81,51	85,41	88,21	90,32	91,76	92,78	77,18

Taula A.4: Mesures de pèrdua de confidencialitat per intervals de confiança, basats en rangs, obtingudes en aplicar diversos mètodes (2/5).

Mètode	ICN-1	ICN-2	ICN-3	ICN-4	ICN-5	ICN-6	ICN-7	ICN-8	ICN-9	ICN-10	ICN
mic5_12	7,46	21,62	33,05	43,53	51,65	58,13	63,48	68,05	71,73	73,32	49,20
mic4_13	11,27	30,59	43,72	53,19	60,66	63,79	66,45	71,50	75,11	77,91	55,42
mic6_15	8,58	26,13	38,73	38,73	48,92	56,94	60,76	63,51	68,48	72,47	48,32
mic6_13	7,89	22,71	36,10	46,05	54,34	57,49	60,60	65,85	70,21	73,54	49,48
mic5_06	14,02	29,16	41,92	52,36	59,03	64,11	69,26	73,81	77,27	79,22	56,02
mic5_19	9,27	25,77	25,77	39,99	40,22	51,67	57,07	60,04	67,26	67,26	44,43
mic4_12	11,00	29,35	43,12	52,92	59,74	65,46	70,09	73,83	77,04	78,32	56,09
rank16	8,87	16,00	22,76	29,59	35,77	41,54	47,78	53,65	59,00	63,82	37,88
mic3_03	50,98	66,26	77,61	84,00	88,32	90,88	92,73	94,21	95,16	95,92	83,61
mic4_09	28,67	40,55	49,05	55,37	63,33	69,59	73,62	76,26	78,81	81,93	61,72
rank03	32,86	64,29	95,20	96,99	97,69	97,97	98,16	98,38	98,54	98,74	87,88
mic4_18	13,92	36,09	36,09	49,34	54,62	58,92	66,24	66,24	72,02	74,49	52,80
mic5_04	19,55	37,76	48,79	59,30	66,10	71,15	75,84	79,29	82,18	84,20	62,42
mict_03	18,95	35,79	50,77	59,83	66,67	72,12	76,71	80,41	83,08	85,02	62,93
mic6_06	14,25	30,56	43,30	52,83	59,06	63,83	69,27	73,56	77,27	79,45	56,34
mic4_14	12,08	31,94	45,33	54,77	54,86	62,31	67,94	72,61	74,44	76,34	55,26
mict_14	9,27	26,85	38,55	49,53	49,63	57,93	64,13	68,63	71,03	72,32	50,79
mic4_20	13,54	36,92	36,92	51,10	51,10	60,85	60,85	68,01	68,01	73,68	52,10
mic2_20	22,62	53,03	53,03	67,50	67,50	76,98	76,98	82,77	82,77	86,05	66,92
mic2_12	21,90	48,45	63,08	72,53	78,98	83,28	86,39	88,67	90,02	90,61	72,39
mic2_17	21,39	51,10	51,66	66,70	75,30	75,83	81,44	85,03	85,45	87,44	68,13
mic2_15	21,43	50,40	64,89	64,89	74,32	80,56	82,74	84,27	86,97	88,87	69,93
mic5_10	13,64	30,39	39,80	47,47	53,81	59,10	64,22	67,91	71,44	73,91	52,17
mic2_16	22,44	51,18	60,61	66,71	75,08	78,89	81,43	84,61	87,30	87,62	69,59
mict_07	17,31	32,23	42,73	53,75	59,05	66,34	69,59	73,98	76,23	78,77	57,00
mic2_14	21,30	50,09	64,34	73,64	73,76	79,82	84,24	86,87	87,96	88,78	71,08
mict_09	18,72	29,68	39,87	47,50	58,08	64,29	69,36	72,37	74,68	77,37	55,19
mic2_13	19,80	48,50	63,58	73,16	79,76	82,04	84,01	86,64	88,63	89,87	71,60
mic2_11	38,96	57,46	67,97	75,54	78,31	83,16	86,34	88,46	89,98	91,03	75,72
mic2_10	39,12	61,25	70,81	77,39	82,47	85,80	87,93	89,69	90,81	91,67	77,69
mic5_16	8,19	24,96	32,82	38,06	48,30	53,54	57,38	64,08	68,92	69,30	46,56
mic2_18	21,62	52,34	52,34	66,88	72,44	75,95	81,53	81,53	85,22	86,65	67,65
mic2_19	22,74	52,35	52,35	66,90	67,29	76,05	79,57	81,52	85,12	85,12	66,90
mic5_09	17,68	28,04	37,41	44,25	53,82	60,97	64,90	68,33	71,16	74,93	52,15
mic4_16	13,14	34,52	42,04	47,80	56,08	60,51	64,33	69,69	73,79	74,32	53,62
mic2_09	46,77	60,77	70,01	77,04	83,90	87,69	89,52	90,84	91,66	92,84	79,10
mic2_08	46,00	60,04	73,35	80,95	84,51	88,48	90,36	91,37	92,88	93,31	80,13
mict_20	11,68	34,05	34,05	46,58	46,58	58,12	58,12	65,67	65,67	72,51	49,30
mict_04	16,27	34,90	46,24	56,35	63,28	69,06	74,67	77,66	80,63	83,36	60,24
mic2_07	44,72	63,96	74,68	83,61	86,65	89,92	91,03	92,57	93,31	94,03	81,45
mic6_07	15,84	29,57	40,45	51,97	56,97	64,41	67,78	72,46	75,38	77,88	55,27
mic2_06	44,54	67,42	79,17	85,90	88,89	90,81	92,44	93,75	94,62	95,14	83,27
mic2_05	50,85	72,81	82,05	87,64	90,73	93,01	94,30	95,16	95,68	96,08	85,83
mic2_04	54,86	76,61	85,63	90,63	92,96	94,24	95,48	96,15	96,76	97,17	88,05
jpg055	13,87	24,15	34,19	42,91	50,33	57,09	63,30	68,31	72,76	76,60	50,35
mic2_03	63,50	81,62	90,09	93,77	95,28	96,26	96,90	97,42	97,77	98,03	91,06
mic5_07	16,27	28,92	39,83	51,08	55,86	63,45	66,45	71,27	74,64	77,30	54,51
rank02	48,42	95,52	97,57	98,35	98,61	98,70	98,80	98,86	99,00	99,05	93,29
mostreig1	76,15	94,03	98,94	99,64	99,72	99,81	99,84	99,84	99,84	99,84	96,76
rank01	95,68	98,13	98,78	99,03	99,15	99,21	99,24	99,27	99,31	99,40	98,72
jpg050	13,38	23,25	32,54	40,29	48,53	54,75	61,21	66,13	70,78	74,58	48,54
jpg060	14,47	24,89	36,07	44,38	52,05	58,88	64,87	70,05	74,70	78,58	51,90
micir_20	49,49	97,66	97,66	97,71	97,71	97,77	97,77	97,86	97,86	97,93	92,94
micir_19	50,62	96,16	96,16	97,83	97,83	97,90	97,95	97,98	98,05	98,05	92,85
micir_18	50,04	97,98	97,98	98,01	98,05	98,06	98,16	98,16	98,21	98,23	93,29

Taula A.4: Mesures de pèrdua de confidencialitat per intervals de confiança, basats en rangs, obtingudes en aplicar diversos mètodes (3/5).

Mètode	ICN-1	ICN-2	ICN-3	ICN-4	ICN-5	ICN-6	ICN-7	ICN-8	ICN-9	ICN-10	ICN
micir_17	51,20	96,79	98,07	98,13	98,17	98,18	98,27	98,33	98,34	98,36	93,38
micir_16	50,36	97,04	98,25	98,26	98,30	98,33	98,36	98,42	98,48	98,48	93,43
micir_15	51,01	98,21	98,24	98,24	98,30	98,33	98,37	98,43	98,49	98,54	93,62
mostreig3	85,79	99,62	99,77	99,77	99,77	99,77	99,77	99,77	99,77	99,77	98,36
micir_14	51,52	98,24	98,27	98,32	98,33	98,35	98,41	98,46	98,49	98,54	93,69
micir_13	52,45	98,48	98,50	98,53	98,55	98,56	98,56	98,63	98,68	98,73	93,97
micir_12	51,79	98,72	98,73	98,73	98,77	98,77	98,81	98,82	98,87	98,87	94,09
jpg040	12,29	21,79	30,54	38,92	45,63	52,21	57,89	63,37	67,95	71,70	46,23
micir_11	98,16	98,79	98,82	98,84	98,85	98,86	98,87	98,89	98,92	98,95	98,79
micir_10	98,75	98,78	98,79	98,83	98,85	98,85	98,87	98,90	98,95	98,96	98,85
micir_09	99,03	99,04	99,04	99,07	99,10	99,11	99,11	99,11	99,15	99,15	99,09
micir_08	99,22	99,22	99,23	99,25	99,25	99,25	99,25	99,25	99,25	99,25	99,24
micir_07	99,28	99,29	99,29	99,32	99,32	99,34	99,34	99,35	99,36	99,36	99,33
jpg095	58,91	82,13	90,39	92,55	93,77	94,10	94,38	94,99	95,46	95,51	89,22
micir_06	99,43	99,44	99,44	99,47	99,47	99,47	99,47	99,47	99,47	99,47	99,46
soroll06	42,30	67,74	81,25	87,90	91,26	92,81	93,74	94,29	94,91	95,40	84,16
micir_05	99,51	99,52	99,52	99,54	99,54	99,54	99,54	99,54	99,54	99,54	99,53
soroll04	54,89	80,43	89,49	92,49	93,79	94,54	95,24	95,74	96,22	96,71	88,95
micir_04	99,67	99,67	99,67	99,67	99,67	99,67	99,67	99,67	99,67	99,67	99,67
jpg070	17,02	29,20	40,38	50,71	58,13	65,40	71,47	76,01	80,35	83,69	57,24
soroll02	77,84	91,77	94,15	95,31	96,12	96,83	97,38	97,86	98,28	98,62	94,42
soroll08	34,31	57,65	73,06	82,31	87,58	90,35	92,06	93,11	93,77	94,40	79,86
jpg100	97,56	98,78	99,14	99,22	99,37	99,42	99,48	99,48	99,48	99,48	99,14
micir_03	99,79	99,79	99,79	99,79	99,79	99,79	99,79	99,79	99,79	99,79	99,79
jpg035	11,38	21,06	30,06	38,21	44,80	50,94	56,63	62,16	66,78	70,90	45,29
jpg090	37,49	60,87	75,14	83,01	87,54	90,04	92,15	92,90	94,56	95,01	80,87
soroll01	88,42	94,70	96,25	97,26	97,93	98,53	98,95	99,27	99,43	99,57	97,03
soroll12	25,14	43,96	58,63	69,42	77,27	82,71	86,65	89,23	91,06	92,25	71,63
jpg080	22,07	37,82	50,83	61,25	70,22	76,03	80,94	84,52	87,74	90,08	66,15
micp_03	13,91	26,73	38,93	47,67	54,81	60,64	66,22	70,26	73,33	75,90	52,84
soroll10	28,72	49,72	64,42	75,13	81,92	86,25	89,16	91,10	92,41	93,33	75,22
jpg075	19,45	33,58	45,73	56,28	63,65	69,89	75,68	79,84	83,48	86,17	61,38
jpg045	13,11	22,54	31,79	40,56	47,02	53,85	59,73	65,09	69,83	74,00	47,75
micz_03	14,91	26,71	38,38	47,09	53,57	59,55	64,64	69,34	72,78	75,60	52,26
jpg085	27,04	46,36	60,58	70,82	78,52	83,68	87,24	89,42	91,54	92,57	72,78
micp_07	13,36	25,59	34,19	45,41	50,59	58,33	61,52	67,64	71,59	74,05	50,23
jpg030	11,61	21,03	29,46	37,24	43,83	49,85	55,30	60,85	65,35	69,34	44,39
soroll16	19,59	35,88	48,89	58,92	67,36	74,04	79,69	83,87	87,12	89,18	64,45
mict_06	12,78	27,52	39,44	50,26	56,88	62,65	68,50	72,74	76,03	77,99	54,48
micp_12	7,18	19,32	29,66	40,68	49,91	56,41	61,71	66,50	70,26	71,79	47,34
jpg065	15,26	26,74	37,61	46,98	55,19	61,34	67,78	73,06	76,99	80,76	54,17
soroll20	16,11	29,54	41,57	51,64	59,82	66,87	72,59	77,20	81,08	84,00	58,04
soroll14	21,43	38,53	53,28	64,25	72,27	78,14	82,74	86,52	89,10	90,86	67,71
micp_16	9,17	24,90	32,36	38,58	48,26	53,85	57,49	64,67	69,12	69,86	46,83
micp_06	10,90	24,74	35,51	45,43	51,20	55,04	60,90	66,24	70,30	72,91	49,32
soroll18	17,66	32,55	45,01	55,36	63,46	70,11	76,07	80,66	84,15	86,68	61,17
micz_19	10,92	28,38	28,38	43,35	43,74	54,27	59,53	62,10	68,33	68,33	46,73
micz_09	17,44	27,18	36,15	42,44	52,44	59,04	63,14	66,47	69,29	72,24	50,58
micp_08	15,90	24,62	36,35	45,07	49,46	56,52	62,45	65,53	70,94	73,50	50,03
micp_05	13,57	27,39	36,25	44,20	51,85	59,47	63,50	68,52	71,69	74,11	51,05
micz_18	10,51	28,33	28,33	43,46	49,36	54,10	61,54	61,54	66,54	69,23	47,29
micz_05	13,89	26,42	36,40	44,94	52,10	60,26	64,85	69,05	71,47	73,82	51,32
micz_13	8,73	23,73	35,31	46,45	53,58	56,55	59,79	64,23	68,50	71,46	48,83
micz_08	16,07	25,36	37,09	47,52	52,02	59,37	64,67	67,52	72,42	74,02	51,61
jpg025	11,19	20,30	28,20	35,40	42,02	47,96	53,92	59,00	63,80	67,80	42,96

Taula A.4: Mesures de pèrdua de confidencialitat per intervals de confiança, basats en rangs, obtingudes en aplicar diversos mètodes (4/5).

Mètode	ICN-1	ICN-2	ICN-3	ICN-4	ICN-5	ICN-6	ICN-7	ICN-8	ICN-9	ICN-10	ICN
distrib	40,89	70,20	85,54	92,93	95,61	96,77	97,46	97,86	98,04	98,11	87,34
micz_11	14,37	26,99	38,15	46,37	48,90	55,41	60,77	64,47	67,83	70,73	49,40
micp_04	13,25	27,15	37,32	47,15	53,99	59,29	65,10	68,86	73,70	75,81	52,16
micz_04	13,02	29,29	39,03	48,26	54,44	59,80	65,07	69,06	72,54	74,87	52,54
micp_13	9,26	21,42	33,18	43,01	50,79	54,86	58,48	64,31	67,46	70,53	47,33
micz_15	7,91	22,65	36,22	36,22	48,61	56,30	61,11	64,53	69,98	73,29	47,68
micz_20	10,54	29,77	29,77	45,44	45,44	58,55	58,55	65,10	65,10	71,08	47,93
micz_06	11,88	25,73	37,09	47,27	53,72	58,97	64,06	68,21	71,32	73,85	51,21
micp_09	16,15	26,47	34,55	42,24	50,96	58,91	63,21	66,47	70,32	73,72	50,30
jpg020	10,82	18,93	26,30	33,44	39,67	45,34	50,85	56,04	60,85	64,85	40,71
micz_17	10,41	27,37	27,37	42,75	52,44	52,69	59,59	66,25	66,50	71,22	47,66
micp_15	7,91	22,54	33,76	33,76	46,69	55,98	60,26	62,18	66,45	69,55	45,91
micp_14	9,70	24,50	35,60	45,47	45,77	53,76	60,94	66,74	68,33	70,04	48,09
micz_16	9,80	29,23	35,50	41,37	50,54	55,33	59,49	65,76	71,05	71,57	48,96
micz_12	8,38	22,74	32,91	42,05	51,03	57,44	63,68	66,84	69,49	70,77	48,53
micp_11	12,16	23,40	32,57	41,83	45,14	52,98	58,00	63,20	66,18	68,76	46,42
jpg015	9,99	18,16	25,07	31,34	37,05	42,40	47,69	52,72	57,12	61,25	38,28
micp_17	7,09	22,89	23,08	38,46	47,43	48,46	58,03	63,96	64,69	69,65	44,37
micz_10	13,39	29,49	38,18	45,94	52,64	57,98	63,75	67,31	69,87	71,79	51,03
micp_10	12,25	28,35	37,46	46,15	51,78	56,62	60,83	65,10	68,66	71,51	49,87
micz_14	9,49	23,77	36,47	47,46	47,66	57,86	64,96	69,25	71,24	72,74	50,09
jpg010	9,72	16,51	23,03	28,85	34,07	39,07	44,03	48,99	53,38	57,41	35,51
micp_18	9,74	26,15	26,15	38,97	44,62	50,38	58,72	58,72	65,77	69,23	44,85
micz_07	13,54	24,92	35,87	47,61	52,55	59,99	63,33	68,82	71,56	74,42	51,26
micp_19	8,66	26,91	26,91	41,84	41,98	51,57	57,52	60,77	66,97	66,97	45,01
micp_20	11,25	31,05	31,05	45,87	45,87	55,98	55,98	62,82	62,82	68,38	47,11

Taula A.4: Mesures de pèrdua de confidencialitat per intervals de confiança, basats en rangs, obtingudes en aplicar diversos mètodes (5/5).

Mètode	ICD-1	ICD-2	ICD-3	ICD-4	ICD-5	ICD-6	ICD-7	ICD-8	ICD-9	ICD-10	ICD
rank07	8,97	13,41	17,54	20,31	24,54	27,93	30,98	34,23	37,37	42,34	25,76
rank14	4,81	7,63	10,33	12,55	15,75	18,05	19,93	22,02	23,56	26,45	16,11
rank12	5,44	8,64	11,54	14,00	17,69	20,17	22,38	24,41	26,27	29,30	17,98
mict_16	4,96	9,52	14,30	18,46	22,91	29,46	33,11	37,27	40,34	42,79	25,31
rank06	9,89	14,92	18,97	22,31	27,83	31,94	35,05	38,78	41,89	48,31	28,99
rank08	7,39	11,14	14,90	17,57	21,99	25,17	28,02	30,68	33,30	37,91	22,81
rank10	6,20	9,86	13,21	15,95	19,60	22,19	24,42	26,65	29,10	32,46	19,97
mict_10	5,41	9,62	14,32	18,80	23,15	27,28	31,98	35,76	39,74	43,73	24,98
rank17	4,20	6,85	9,44	11,59	14,44	16,70	18,70	20,54	22,31	25,06	14,98
rank09	6,55	10,63	13,77	16,34	20,63	23,77	25,80	28,19	30,71	34,54	21,09
rank11	5,59	9,38	12,28	14,81	18,22	20,93	22,88	25,25	27,22	30,61	18,72
mic6_19	3,97	8,39	12,49	16,79	20,51	24,05	28,23	32,49	34,99	37,93	21,98
mic6_03	9,29	18,46	24,98	32,74	39,29	44,08	48,96	52,66	56,56	59,93	38,70
rank15	4,31	6,99	9,88	11,95	15,17	17,65	19,37	21,36	23,18	26,04	15,59
rank13	4,74	7,73	10,73	13,13	16,69	19,42	21,11	23,60	25,61	28,66	17,14
mic6_08	5,21	10,73	15,88	21,23	25,61	29,99	34,35	38,13	41,88	45,43	26,85
mict_15	3,21	7,91	12,71	17,09	21,26	26,07	30,66	35,04	38,35	43,16	23,55
mict_18	7,31	12,69	18,08	23,59	26,41	31,92	35,38	38,46	41,92	45,00	28,08
mic5_18	3,69	6,97	10,84	14,99	18,53	22,19	25,72	29,44	32,65	36,32	20,13
rank18	4,37	6,94	9,50	11,55	14,37	16,59	18,46	20,26	22,04	24,47	14,85

Taula A.5: Mesures de pèrdua de confidencialitat per intervals de confiança, basats en desviacions típiques, obtingudes en aplicar els diferents mètodes (1/5).

Mètode	ICD-1	ICD-2	ICD-3	ICD-4	ICD-5	ICD-6	ICD-7	ICD-8	ICD-9	ICD-10	ICD
mic5_15	4,67	8,74	12,78	16,50	20,21	23,89	27,41	30,78	34,00	37,44	21,64
mict_11	4,86	9,64	14,67	19,60	25,24	29,17	33,67	38,23	40,90	44,68	26,07
mic6_04	7,46	14,54	21,28	27,67	33,10	37,83	42,49	46,47	50,42	54,29	33,55
mic4_17	6,58	12,61	17,91	23,73	28,53	33,34	37,46	40,73	43,67	47,05	29,16
mict_08	4,16	9,06	14,64	19,09	24,27	29,00	33,39	37,15	41,31	44,27	25,64
rank05	10,94	16,24	20,89	24,55	30,89	35,68	39,55	43,90	48,06	55,25	32,59
mic4_11	9,09	16,37	23,71	29,28	34,06	38,57	42,19	45,61	48,93	52,13	33,99
mict_19	4,06	7,85	12,18	16,24	20,01	23,39	27,86	34,04	38,62	41,87	22,61
mic6_05	6,87	13,25	18,86	24,14	29,15	34,26	38,93	43,01	46,52	49,94	30,49
mic4_04	16,95	27,85	35,89	42,30	47,74	52,23	56,40	60,06	63,38	66,62	46,94
mic6_14	4,36	8,77	13,08	17,65	21,97	25,95	30,00	33,21	36,90	40,03	23,19
mic3_15	12,68	22,89	30,55	36,47	41,27	44,79	48,90	52,35	55,11	57,83	40,28
mic3_14	13,08	22,54	30,95	36,83	41,87	46,22	50,04	53,15	56,20	58,85	40,97
rank04	13,48	19,32	24,49	29,16	37,32	43,30	47,87	53,48	58,82	66,76	39,40
mic3_08	20,66	31,23	39,15	45,09	49,91	53,60	57,31	60,42	63,45	66,19	48,70
mic5_20	2,89	6,65	10,55	14,47	17,38	21,17	24,52	28,11	31,94	34,84	19,25
mic3_13	13,88	23,70	31,46	37,74	42,28	46,48	50,37	53,85	57,00	59,88	41,66
mict_12	5,73	9,23	13,25	17,86	22,05	25,90	29,40	33,68	37,52	41,45	23,61
mic3_20	8,97	18,01	24,16	30,86	35,78	40,10	43,82	47,05	50,17	52,77	35,17
mic5_17	4,21	8,33	12,13	16,20	19,59	23,30	26,45	31,47	34,76	38,11	21,45
mic3_06	23,89	34,29	41,89	48,05	53,45	57,64	61,33	64,66	67,78	70,63	52,36
mic4_06	14,86	23,73	30,66	36,45	42,09	46,49	50,19	53,58	57,04	60,02	41,51
mict_05	5,09	10,51	14,92	20,16	25,89	30,45	34,12	38,92	42,91	46,23	26,92
mic6_11	4,59	9,16	14,09	18,07	22,74	27,19	31,05	35,26	38,62	41,66	24,24
mic3_07	22,39	33,64	41,04	47,51	52,24	56,14	59,67	63,18	66,15	68,76	51,07
mict_13	3,80	9,17	13,35	20,20	24,09	28,16	32,71	37,43	40,68	43,28	25,29
mic3_09	18,67	29,25	37,34	43,48	47,89	51,84	55,59	58,69	62,03	65,10	46,99
mic6_12	4,58	8,42	13,19	17,41	21,43	25,45	29,40	33,47	37,61	40,96	23,19
mic5_14	3,61	8,13	12,41	16,71	20,21	23,84	28,26	32,28	35,63	38,67	21,98
mic6_20	3,65	7,49	11,32	15,74	19,93	23,84	27,44	31,06	34,42	37,26	21,22
mic4_19	5,82	11,10	17,12	22,01	26,97	31,32	34,90	38,55	42,04	45,36	27,52
rank19	4,19	6,52	8,87	11,06	13,85	16,08	18,05	19,95	21,63	24,36	14,46
mic5_05	6,22	11,91	17,51	23,07	28,54	33,43	37,68	41,50	45,41	49,23	29,45
mic3_12	16,53	25,36	33,46	39,24	44,23	48,29	51,90	55,08	57,99	60,95	43,30
mic5_03	8,25	16,56	22,72	29,97	36,49	41,43	46,70	50,58	54,62	58,30	36,56
mic4_07	12,03	20,75	28,11	34,06	39,10	43,50	47,31	51,05	54,52	57,42	38,79
mic3_18	10,97	19,50	26,20	32,55	38,21	42,86	46,73	50,09	52,83	55,59	37,55
mic6_09	5,09	9,39	13,74	18,64	23,95	28,58	32,34	35,82	39,83	43,28	25,07
mict_17	7,21	12,84	17,81	22,90	26,90	30,90	33,45	37,08	40,05	42,59	27,17
mic4_05	15,80	25,49	33,52	39,47	44,66	49,41	53,21	56,97	60,10	63,03	44,16
mic3_05	26,45	37,21	44,59	50,94	55,76	59,97	63,66	66,88	69,90	73,07	54,84
mic6_16	4,18	8,85	12,90	17,07	21,01	25,24	29,05	32,89	36,47	39,63	22,73
rank20	3,81	6,18	8,94	10,90	13,75	16,18	18,13	20,01	21,75	24,44	14,41
mic5_08	5,14	9,86	14,54	19,40	24,11	28,67	32,59	36,70	40,43	44,10	25,55
mic3_16	11,50	19,94	28,60	34,52	39,37	43,43	47,30	50,48	53,45	56,03	38,46
mic5_13	4,13	7,99	12,04	16,03	20,41	24,52	28,16	32,30	35,60	38,78	22,00
mic3_17	9,20	18,85	26,94	33,35	38,26	42,65	46,55	50,04	53,01	55,76	37,46
mic4_10	8,77	17,40	24,37	30,76	35,32	39,21	43,09	46,47	49,36	52,76	34,75
mic3_11	16,23	25,68	34,82	40,84	45,61	49,27	52,81	56,13	59,22	62,38	44,30
mic3_04	28,80	40,06	48,44	54,57	59,24	63,40	67,64	70,80	73,65	76,52	58,31
mic5_11	4,71	9,05	13,42	17,93	21,94	25,73	29,62	33,92	37,91	41,27	23,55
mic4_15	6,26	13,59	20,33	25,20	29,96	34,26	37,86	41,72	44,96	48,03	30,22
mic4_08	11,25	20,38	26,79	32,81	37,96	41,91	45,80	49,81	52,83	55,76	37,53
mic6_18	3,91	8,11	12,72	16,96	20,76	24,00	27,77	31,31	35,31	37,95	21,88
mic3_19	9,39	18,12	26,02	32,23	37,09	41,66	45,32	48,54	51,41	54,29	36,41

Taula A.5: Mesures de pèrdua de confidencialitat per intervals de confiança, basats en desviacions típiques, obtingudes en aplicar els diferents mètodes (2/5).

Mètode	ICD-1	ICD-2	ICD-3	ICD-4	ICD-5	ICD-6	ICD-7	ICD-8	ICD-9	ICD-10	ICD
mic6_10	5,03	9,47	14,33	19,57	24,02	28,06	31,53	35,28	38,90	42,28	24,85
mic6_17	4,11	7,82	12,41	16,51	20,16	24,27	27,71	31,97	35,45	39,06	21,95
mic3_10	17,91	28,18	36,05	42,17	46,42	50,51	53,84	57,20	60,17	62,97	45,54
mic4_03	22,62	33,92	40,99	47,54	53,11	57,53	61,69	65,46	69,04	72,17	52,41
mic5_12	4,58	9,27	13,56	17,96	22,42	26,45	29,73	33,30	36,73	40,28	23,43
mic4_13	7,11	14,23	21,27	26,94	31,66	36,02	39,91	43,15	46,50	49,67	31,65
mic6_15	3,71	8,43	12,98	17,43	21,52	25,44	29,52	33,43	37,09	39,75	22,93
mic6_13	3,77	8,21	12,67	17,66	21,74	25,81	29,67	33,75	37,54	41,05	23,19
mic5_06	5,38	10,41	15,21	20,16	24,85	29,39	34,24	38,13	42,15	45,98	26,59
mic5_19	3,82	7,13	11,28	14,54	18,18	22,05	25,63	28,97	32,32	35,66	19,96
mic4_12	8,06	15,97	21,77	28,43	33,16	37,60	41,01	44,77	48,26	51,31	33,03
rank16	4,70	7,60	10,26	12,22	15,28	17,46	19,25	21,40	23,38	25,80	15,74
mic3_03	33,41	45,11	53,12	59,41	64,46	68,72	72,63	75,67	78,59	81,50	63,26
mic4_09	10,09	18,84	26,00	31,92	36,85	40,95	44,84	48,47	51,77	54,74	36,45
rank03	16,65	23,80	30,72	37,36	49,20	57,49	63,40	69,71	74,22	80,46	50,30
mic4_18	6,40	12,34	19,12	24,34	29,42	33,29	36,74	40,59	43,21	46,12	29,16
mic5_04	7,20	14,11	20,07	26,32	31,75	35,98	40,57	44,75	48,75	52,61	32,21
mict_03	6,88	13,40	17,71	23,97	29,34	33,78	39,04	43,53	47,93	51,58	30,72
mic6_06	6,22	12,09	17,55	22,09	27,35	31,52	35,69	39,42	43,48	46,96	28,24
mic4_14	7,98	14,71	21,12	26,57	31,94	36,18	39,96	43,38	46,37	49,64	31,78
mict_14	2,69	8,30	12,19	16,98	21,07	26,25	30,44	34,53	37,92	41,51	23,19
mic4_20	5,88	10,98	17,31	21,46	25,82	30,06	33,94	37,82	41,44	44,74	26,95
mic2_20	12,23	22,29	30,61	38,05	44,49	49,84	54,51	58,28	61,75	64,67	43,67
mic2_12	17,24	30,03	38,77	47,19	53,52	58,59	63,02	66,92	70,55	74,00	51,98
mic2_17	12,17	22,20	32,14	39,49	46,03	51,80	56,89	60,92	64,39	67,68	45,37
mic2_15	14,25	26,78	35,60	43,40	48,73	54,10	58,71	62,69	66,60	69,74	48,06
mic5_10	4,54	8,54	13,68	18,27	22,76	26,72	30,81	34,69	38,35	41,69	24,01
mic2_16	14,24	24,83	34,53	41,98	47,61	53,01	57,90	61,82	65,41	68,97	47,03
mict_07	6,40	11,98	16,43	22,02	27,01	32,26	36,05	39,44	43,89	47,48	28,29
mic2_14	16,00	27,45	36,36	44,09	50,54	55,83	60,36	64,18	67,81	71,32	49,39
mict_09	3,91	8,97	13,78	19,04	23,27	27,63	32,63	37,12	41,03	44,94	25,23
mic2_13	15,93	27,38	37,23	45,36	51,98	57,33	61,87	66,09	69,57	72,55	50,53
mic2_11	18,60	30,87	40,66	48,38	55,05	60,00	64,81	69,05	72,69	75,96	53,61
mic2_10	19,96	32,01	42,03	50,24	56,73	62,33	66,46	70,36	73,78	77,09	55,10
mic5_16	3,70	7,85	11,87	16,21	20,13	24,39	28,21	31,37	34,47	37,69	21,59
mic2_18	11,89	22,50	32,31	39,74	45,83	50,99	55,25	59,58	63,30	66,61	44,80
mic2_19	12,02	22,51	31,15	39,41	45,63	50,04	55,18	59,16	62,56	65,85	44,35
mic5_09	4,56	9,32	14,87	19,40	24,36	28,57	32,51	36,05	39,07	42,25	25,10
mic4_16	7,71	13,23	19,89	25,29	30,26	34,49	38,25	42,14	45,19	47,69	30,41
mic2_09	21,27	34,55	44,62	52,28	58,40	64,09	68,38	72,29	75,70	78,81	57,04
mic2_08	22,44	36,09	46,20	54,78	60,87	66,17	70,94	74,58	77,77	80,57	59,04
mict_20	3,28	8,12	13,11	17,66	21,79	25,93	29,34	33,90	37,32	41,74	23,22
mict_04	6,24	11,65	17,55	22,34	27,58	31,62	36,13	40,57	44,76	48,63	28,71
mic2_07	24,25	38,03	48,62	56,32	62,93	68,18	72,90	76,83	80,02	83,01	61,11
mic6_07	5,26	10,46	16,38	21,65	26,31	30,62	34,91	39,03	42,77	46,11	27,35
mic2_06	26,33	40,27	51,47	59,79	66,20	71,40	75,83	79,71	82,53	85,42	63,90
mic2_05	29,15	44,32	55,68	63,56	70,13	75,25	79,60	82,86	85,75	87,87	67,42
mic2_04	32,54	48,16	59,57	68,09	74,72	79,50	83,53	86,52	89,11	90,99	71,27
jpg055	3,68	6,92	10,11	13,25	15,92	18,97	21,83	24,75	27,29	29,96	17,27
mic2_03	38,38	55,66	66,09	74,53	80,34	84,86	88,41	90,90	92,79	94,05	76,60
mic5_07	5,35	10,32	15,35	20,21	25,04	29,30	33,43	37,16	40,56	44,15	26,08
rank02	24,16	33,59	43,70	53,94	67,59	75,38	80,58	84,59	87,14	90,09	64,08
mostreig1	42,68	57,68	67,86	74,18	82,28	88,19	91,09	93,26	94,76	96,13	78,81
rank01	37,81	57,02	73,80	82,22	90,07	92,56	93,75	94,60	95,11	95,89	81,28
jpg050	3,63	6,60	9,48	12,57	15,36	18,01	20,83	23,23	25,95	28,33	16,40

Taula A.5: Mesures de pèrdua de confidencialitat per intervals de confiança, basats en desviacions típiques, obtingudes en aplicar els diferents mètodes (3/5).

A. Taules (comparació de mètodes pertorbatius)

Mètode	ICD-1	ICD-2	ICD-3	ICD-4	ICD-5	ICD-6	ICD-7	ICD-8	ICD-9	ICD-10	ICD
jpg060	3,90	7,32	10,61	13,63	16,81	19,70	23,00	25,95	28,71	31,40	18,10
micir_20	39,55	64,11	80,59	88,72	92,04	93,95	94,99	95,77	96,27	96,65	84,26
micir_19	41,63	65,71	81,47	89,00	92,17	93,97	95,12	95,80	96,40	96,88	84,81
micir_18	43,08	67,84	83,82	90,33	92,94	94,62	95,64	96,34	96,71	97,05	85,84
micir_17	45,21	71,27	84,81	90,88	93,55	94,86	95,84	96,44	97,09	97,40	86,74
micir_16	47,04	72,95	86,50	91,55	94,23	95,46	96,30	96,87	97,21	97,56	87,57
micir_15	49,87	75,80	87,77	92,67	94,80	95,78	96,47	96,99	97,30	97,54	88,50
mostreig3	42,61	65,73	77,67	86,77	92,24	95,02	96,82	97,68	98,23	98,53	85,13
micir_14	52,69	77,99	88,64	93,24	95,07	96,15	96,75	97,10	97,40	97,69	89,27
micir_13	54,92	80,05	90,15	94,01	95,43	96,32	96,94	97,39	97,71	97,91	90,08
micir_12	57,72	82,01	91,42	94,70	96,09	96,74	97,26	97,60	97,96	98,10	90,96
jpg040	3,40	6,33	8,98	11,73	14,38	16,72	19,19	21,59	24,09	26,46	15,29
micir_11	62,32	84,87	92,43	95,32	96,46	96,98	97,51	97,83	98,02	98,25	92,00
micir_10	65,84	86,94	93,45	95,97	96,87	97,45	97,86	98,05	98,28	98,40	92,91
micir_09	69,57	88,73	94,16	96,29	97,17	97,78	98,13	98,35	98,45	98,62	93,73
micir_08	73,70	89,99	95,15	96,92	97,66	98,10	98,40	98,67	98,77	98,90	94,63
micir_07	78,00	92,27	95,95	97,41	97,99	98,23	98,58	98,67	98,80	98,92	95,48
jpg095	18,63	36,10	51,13	64,46	74,92	81,97	87,17	91,13	93,95	95,99	69,55
micir_06	82,43	93,81	96,79	97,88	98,39	98,65	98,85	98,95	99,06	99,13	96,40
soroll06	13,68	26,70	38,55	49,97	60,16	68,85	76,21	81,95	86,79	90,52	59,34
micir_05	85,87	95,06	97,44	98,23	98,60	98,82	98,97	99,10	99,17	99,25	97,05
soroll04	19,03	37,70	53,65	67,65	78,76	86,57	91,93	95,56	97,69	98,75	72,73
micir_04	89,59	96,57	98,07	98,62	98,91	99,12	99,28	99,40	99,44	99,49	97,85
jpg070	4,42	8,53	12,15	15,97	19,69	23,14	26,65	30,06	33,75	36,70	21,11
soroll02	39,05	68,83	87,02	95,33	98,88	99,71	99,92	100,00	100,00	100,00	88,87
soroll08	10,14	19,94	30,09	38,70	46,84	54,64	61,91	68,44	74,15	79,30	48,42
jpg100	46,06	84,08	93,13	96,84	99,15	99,52	99,67	99,79	99,94	100,00	91,82
micir_03	92,86	97,56	98,43	99,02	99,27	99,44	99,49	99,55	99,64	99,67	98,49
jpg035	3,02	5,74	8,60	11,21	13,68	15,71	18,06	20,41	22,78	25,21	14,44
jpg090	10,81	20,44	30,20	38,76	47,24	54,81	61,40	67,29	73,05	77,44	48,15
soroll01	67,92	94,80	99,73	99,99	100,00	100,00	100,00	100,00	100,00	100,00	96,24
soroll12	7,03	13,84	20,01	26,92	33,05	38,65	44,24	49,74	54,88	59,61	34,80
jpg080	5,96	10,99	16,25	21,20	26,34	30,93	35,58	40,25	44,44	48,36	28,03
micp_03	3,74	7,80	11,50	15,28	19,08	22,97	26,41	29,70	33,89	36,92	20,73
soroll10	7,86	15,80	23,69	31,15	38,40	45,24	51,70	57,86	63,33	68,40	40,34
jpg075	5,21	10,06	14,49	18,63	22,84	26,89	30,74	34,74	38,52	41,86	24,40
jpg045	3,60	6,63	9,52	12,27	14,91	17,41	19,90	22,46	25,00	27,11	15,88
micz_03	3,59	8,27	12,31	16,35	19,87	24,29	27,50	31,43	35,11	37,91	21,66
jpg085	7,34	13,78	20,12	26,44	32,85	38,65	44,20	49,49	54,75	59,27	34,69
micp_07	3,14	7,63	10,92	14,81	17,50	21,55	24,49	27,41	30,57	32,93	19,10
jpg030	3,29	6,01	8,70	11,25	13,84	16,02	18,16	20,29	22,46	24,72	14,47
soroll16	5,12	9,96	15,06	20,13	25,26	30,36	34,94	39,43	43,60	47,82	27,17
mict_06	4,66	9,06	14,62	19,02	23,38	27,61	32,05	35,81	39,32	43,46	24,90
micp_12	3,33	5,64	8,55	11,45	14,19	18,12	21,71	24,70	27,61	31,11	16,64
jpg065	4,28	7,80	11,23	14,72	18,02	21,55	24,70	27,81	30,91	33,94	19,50
soroll20	3,85	7,96	12,16	16,00	20,17	23,90	27,86	31,55	35,09	39,00	21,75
soroll14	5,84	11,49	17,13	22,41	27,94	32,94	38,16	43,23	48,42	52,88	30,05
micp_16	3,25	6,50	9,12	12,19	15,73	19,03	22,17	24,44	26,72	28,43	16,76
micp_06	4,15	7,39	10,38	13,63	16,97	20,73	24,15	27,56	30,47	33,42	18,88
soroll18	4,30	8,80	13,33	18,10	22,38	26,61	31,19	35,36	39,30	43,22	24,26
micz_19	3,36	6,45	10,09	13,87	18,21	21,99	23,73	27,39	30,09	32,24	18,74
micz_09	5,58	9,42	12,50	16,15	19,68	23,21	26,41	29,36	32,63	35,90	21,08
micp_08	3,19	7,81	10,77	14,19	17,44	20,34	23,02	25,81	29,57	33,50	18,56
micp_05	3,85	7,05	10,04	13,53	17,27	21,08	24,39	28,10	31,30	34,26	19,09
micz_18	2,69	6,28	9,74	12,31	15,38	18,97	21,79	25,64	30,26	32,69	17,58

Taula A.5: Mesures de pèrdua de confidencialitat per intervals de confiança, basats en desviacions típiques, obtingudes en aplicar els diferents mètodes (4/5).

Mètode	ICD-1	ICD-2	ICD-3	ICD-4	ICD-5	ICD-6	ICD-7	ICD-8	ICD-9	ICD-10	ICD
micz_05	4,27	7,19	10,79	14,53	18,87	22,01	25,96	28,88	31,98	35,15	19,96
micz_13	3,70	7,41	11,11	14,91	18,43	21,94	25,01	28,63	31,87	33,63	19,66
micz_08	4,10	8,15	12,02	15,21	19,60	22,74	25,76	28,60	32,71	35,50	20,44
jpg025	3,11	5,85	8,55	10,95	13,47	15,44	17,50	19,61	21,62	23,69	13,98
distrib	20,30	32,99	42,22	52,87	61,55	68,00	72,78	76,97	80,50	83,66	59,18
micz_11	3,93	8,57	13,11	16,79	20,09	23,23	27,07	30,93	34,61	37,43	21,58
micp_04	3,56	7,69	11,00	14,81	18,69	21,88	26,27	29,60	32,85	35,38	20,17
micz_04	3,25	6,89	11,65	15,56	19,23	23,08	26,84	30,85	34,19	37,52	20,91
micp_13	4,17	6,94	9,45	13,54	18,35	21,60	24,56	26,97	29,84	33,54	18,90
micz_15	3,31	6,52	10,04	13,03	15,49	19,34	22,54	25,11	27,35	30,56	17,33
micz_20	2,56	6,13	9,40	14,67	16,67	21,37	24,79	28,63	32,19	34,76	19,12
micz_06	3,85	7,14	10,26	14,19	18,33	22,27	25,94	28,93	32,74	35,81	19,94
micp_09	3,33	7,18	10,51	13,27	16,60	20,71	23,78	27,12	29,36	32,63	18,45
jpg020	2,72	5,43	7,50	9,79	12,07	14,26	16,08	18,00	19,81	21,59	12,72
micz_17	2,91	5,93	10,53	14,29	18,40	21,13	24,28	27,07	29,37	32,27	18,62
micp_15	3,85	7,69	9,40	11,43	14,42	17,52	20,51	23,08	26,07	28,53	16,25
micp_14	3,89	8,18	11,57	15,46	17,35	20,30	23,22	25,41	28,50	32,09	18,60
micz_16	2,74	5,70	9,34	13,79	17,38	21,37	24,16	27,29	29,91	33,22	18,49
micz_12	3,76	7,26	10,34	14,62	17,09	20,77	23,16	26,58	30,34	32,99	18,69
micp_11	2,49	6,17	9,15	12,83	16,27	19,10	21,60	25,68	27,87	30,86	17,20
jpg015	2,57	5,06	7,36	9,42	11,15	13,00	14,74	16,67	18,33	20,22	11,85
micp_17	4,06	5,94	8,12	10,30	13,33	15,75	18,53	21,45	24,78	27,27	14,95
micz_10	3,13	7,05	11,40	15,03	19,02	23,65	28,13	31,05	34,47	37,32	21,03
micp_10	4,27	6,91	9,97	13,60	17,17	20,87	24,00	26,14	29,20	31,62	18,38
micz_14	3,49	6,38	10,47	13,26	16,07	19,26	22,95	26,84	31,62	35,63	18,60
jpg010	2,43	5,04	7,15	9,00	10,77	12,24	13,91	15,66	17,29	18,72	11,22
micp_18	2,56	5,77	8,33	12,18	15,77	18,33	20,64	23,59	25,51	28,85	16,15
micz_07	4,34	7,38	10,83	14,75	18,54	21,43	24,57	28,23	31,97	35,62	19,77
micp_19	1,89	4,60	8,66	12,41	14,82	17,93	21,18	24,02	26,19	29,84	16,16
micp_20	4,27	7,41	9,83	13,53	17,52	20,37	23,36	25,21	28,21	31,20	18,09

Taula A.5: Mesures de pèrdua de confidencialitat per intervals de confiança, basats en desviacions típiques, obtingudes en aplicar els diferents mètodes (5/5).

Apèndix B

Correlacions entre estadístics i mesures amb diverses particions de variables

Estad.	Suma rel.	Mic1-12mul	Mic2-11mul	Mic3-10mul	Mic4-9mul	Mic5-8mul	Mic6-7mul
5_11	-4,5337	-0,3335	-0,3393	-0,3030	-0,2521	-0,1413	-0,1254
5_14	-4,2901	-0,2849	-0,1142	-0,2347	-0,2710	-0,2028	-0,2331
5_9	-3,9800	-0,2575	-0,1412	-0,2435	-0,2449	-0,1771	-0,1898
5_33	-3,7445	-0,2849	-0,1142	-0,1886	-0,1590	-0,1720	-0,2845
5_34	-3,7445	-0,2849	-0,1142	-0,1886	-0,1590	-0,1720	-0,2845
5_18	-3,1376	-0,2575	0,1525	-0,1594	-0,1688	-0,1990	-0,2816
5_19	-3,1376	-0,2575	0,1525	-0,1594	-0,1688	-0,1990	-0,2816
5_24	-3,1333	-0,3335	0,1669	-0,1301	-0,1642	-0,2005	-0,2469
5_25	-3,1333	-0,3335	0,1669	-0,1301	-0,1642	-0,2005	-0,2469
5_21	-3,0616	-0,1284	0,0318	-0,1867	-0,1582	-0,1861	-0,3022
5_22	-3,0616	-0,1284	0,0318	-0,1867	-0,1582	-0,1861	-0,3022
5_10	-2,9686	-0,1284	0,0318	-0,1565	-0,2033	-0,1909	-0,2364
3_3	-2,7183			-0,1562	-0,1644	-0,2235	-0,2682
6_3	-2,7183			-0,1562	-0,1644	-0,2235	-0,2682
22_1	-2,5871	-0,1155	-0,0637	0,0085	-0,1294	-0,1906	-0,3375
3_33	-2,5135		0,0650	-0,1339	-0,1637	-0,1980	-0,3077
3_34	-2,5135		0,0650	-0,1339	-0,1637	-0,1980	-0,3077
6_33	-2,5135		0,0650	-0,1339	-0,1637	-0,1980	-0,3077
6_34	-2,5135		0,0650	-0,1339	-0,1637	-0,1980	-0,3077
6_7	-2,2909			-0,1307	-0,1381	-0,1717	-0,2516
3_7	-2,2909			-0,1307	-0,1381	-0,1717	-0,2516
6_21	-2,2410		0,1939	-0,1317	-0,1408	-0,2152	-0,3163
6_22	-2,2410		0,1939	-0,1317	-0,1408	-0,2152	-0,3163
3_21	-2,2410		0,1939	-0,1317	-0,1408	-0,2152	-0,3163
3_22	-2,2410		0,1939	-0,1317	-0,1408	-0,2152	-0,3163
3_18	-2,1149		0,2418	-0,1158	-0,1464	-0,2181	-0,3145
3_19	-2,1149		0,2418	-0,1158	-0,1464	-0,2181	-0,3145
10_33	-2,0738	-0,0341	0,2240	-0,1113	-0,1406	-0,2000	-0,2872
10_34	-2,0738	-0,0341	0,2240	-0,1113	-0,1406	-0,2000	-0,2872
10_3	-2,0618		0,2159	-0,1317	-0,1597	-0,2131	-0,2446
3_2	-2,0618		0,2159	-0,1317	-0,1597	-0,2131	-0,2446

Taula B.1: Suma estandarditzada i correlacions entre els diversos estadístics i les mesures de pèrdua d'informació obtingudes en aplicar les diverses versions de microagregació (1/5).

B. Correlacions entre estadístics i mesures amb diverses particions de variables

Estad.	Suma rel.	Mic1-12mul	Mic2-11mul	Mic3-10mul	Mic4-9mul	Mic5-8mul	Mic6-7mul
5_3	-2,0473			-0,1562	-0,0961	-0,1291	-0,2488
6_11	-2,0286	-0,2943	-0,1593	-0,1452	-0,0961	-0,0418	0,0682
22_6	-1,9793	-0,0291	-0,2959	-0,0699	-0,0422	-0,0402	-0,2678
13_1	-1,9627	0,1133	0,2310	-0,1325	-0,1596	-0,1963	-0,3683
6_8	-1,8933	-0,2962	-0,1770	-0,1461	-0,0909	-0,0216	0,0990
10_8	-1,8933	-0,2962	-0,1770	-0,1461	-0,0909	-0,0216	0,0990
10_11	-1,8933	-0,2962	-0,1770	-0,1461	-0,0909	-0,0216	0,0990
5_7	-1,8069			-0,1307	-0,0803	-0,1117	-0,2376
3_24	-1,7806		0,2282	-0,0779	-0,1257	-0,1935	-0,2902
3_25	-1,7806		0,2282	-0,0779	-0,1257	-0,1935	-0,2902
6_30	-1,7234		-0,1621	-0,0459	-0,2039	-0,0744	-0,0698
6_31	-1,7234		-0,1621	-0,0459	-0,2039	-0,0744	-0,0698
3_4	-1,6189		0,2159	-0,0894	-0,1347	-0,1838	-0,2087
4_18	-1,5643	-0,1133	0,1985	-0,0995	-0,0934	-0,1432	-0,1521
4_19	-1,5643	-0,1133	0,1985	-0,0995	-0,0934	-0,1432	-0,1521
4_30	-1,5643	-0,1133	0,1985	-0,0995	-0,0934	-0,1432	-0,1521
4_31	-1,5643	-0,1133	0,1985	-0,0995	-0,0934	-0,1432	-0,1521
4_27	-1,5643	-0,1133	0,1985	-0,0995	-0,0934	-0,1432	-0,1521
4_28	-1,5643	-0,1133	0,1985	-0,0995	-0,0934	-0,1432	-0,1521
5_27	-1,5643	-0,1133	0,1985	-0,0995	-0,0934	-0,1432	-0,1521
5_28	-1,5643	-0,1133	0,1985	-0,0995	-0,0934	-0,1432	-0,1521
4_15	-1,5643	-0,1133	0,1985	-0,0995	-0,0934	-0,1432	-0,1521
4_16	-1,5643	-0,1133	0,1985	-0,0995	-0,0934	-0,1432	-0,1521
4_24	-1,5643	-0,1133	0,1985	-0,0995	-0,0934	-0,1432	-0,1521
4_25	-1,5643	-0,1133	0,1985	-0,0995	-0,0934	-0,1432	-0,1521
5_2	-1,4453		0,2159	-0,1136	-0,1008	-0,1414	-0,2222
10_7	-1,3656		0,2159	-0,0992	-0,1025	-0,1575	-0,1850
10_21	-1,2268	0,2962	0,2418	-0,1158	-0,1464	-0,2181	-0,3145
10_22	-1,2268	0,2962	0,2418	-0,1158	-0,1464	-0,2181	-0,3145
3_27	-1,1635		0,2085	-0,0951	-0,0877	-0,1418	-0,1524
3_28	-1,1635		0,2085	-0,0951	-0,0877	-0,1418	-0,1524
5_4	-1,1466		0,2159	-0,0786	-0,0916	-0,1356	-0,1756
6_14	-1,0886	0,0377	0,0650	-0,0265	-0,1005	-0,0995	-0,1817
3_14	-1,0886	0,0377	0,0650	-0,0265	-0,1005	-0,0995	-0,1817
6_13	-0,9050	0,1191	0,3125	-0,1773	-0,1570	-0,1478	-0,0666
6_18	-0,8053		-0,2145	0,0047	0,0216	-0,0247	-0,1302
6_19	-0,8053		-0,2145	0,0047	0,0216	-0,0247	-0,1302
18_1	-0,7776	-0,1133	-0,2162	0,0363	0,0063	0,0319	-0,0920
18_2	-0,7776	-0,1133	-0,2162	0,0363	0,0063	0,0319	-0,0920
18_5	-0,7776	-0,1133	-0,2162	0,0363	0,0063	0,0319	-0,0920
18_6	-0,7776	-0,1133	-0,2162	0,0363	0,0063	0,0319	-0,0920
18_4	-0,7776	-0,1133	-0,2162	0,0363	0,0063	0,0319	-0,0920
19_1	-0,7772	-0,1104	-0,2099	-0,0058	-0,0203	0,0194	0,0052
19_2	-0,7772	-0,1104	-0,2099	-0,0058	-0,0203	0,0194	0,0052
19_5	-0,7772	-0,1104	-0,2099	-0,0058	-0,0203	0,0194	0,0052
19_6	-0,7772	-0,1104	-0,2099	-0,0058	-0,0203	0,0194	0,0052
19_4	-0,7772	-0,1104	-0,2099	-0,0058	-0,0203	0,0194	0,0052
4_2	-0,7657		0,2159	-0,0851	-0,0670	-0,1181	-0,0858
4_6	-0,7657		0,2159	-0,0851	-0,0670	-0,1181	-0,0858
4_5	-0,7657		0,2159	-0,0851	-0,0670	-0,1181	-0,0858
5_5	-0,7657		0,2159	-0,0851	-0,0670	-0,1181	-0,0858
3_5	-0,7657		0,2159	-0,0851	-0,0670	-0,1181	-0,0858
4_4	-0,7657		0,2159	-0,0851	-0,0670	-0,1181	-0,0858
4_1	-0,7657		0,2159	-0,0851	-0,0670	-0,1181	-0,0858
5_8	-0,7604	-0,1254	-0,3089	-0,1104	-0,0263	0,0776	0,1566

Taula B.1: Suma estandarditzada i correlacions entre els diversos estadístics i les mesures de pèrdua d'informació obtingudes en aplicar les diverses versions de microagregació (2/5).

Estad.	Suma rel.	Mic1-12mul	Mic2-11mul	Mic3-10mul	Mic4-9mul	Mic5-8mul	Mic6-7mul
10_14	-0,6734	-0,0341	-0,0753	0,0060	-0,0469	-0,0122	-0,0781
5_35	-0,5559	-0,2849	-0,1142	0,0489	-0,0167	0,0103	0,1469
6_9	-0,5340	0,0578	0,2065	-0,0153	-0,0577	-0,1342	-0,1316
6_2	-0,4877		-0,2159	0,0051	0,0256	0,0032	-0,0580
5_6	-0,3409		0,2159	-0,0906	-0,0313	-0,0928	-0,0073
10_13	-0,2658	0,1579	0,3412	-0,0446	-0,0724	-0,1628	-0,1511
5_20	-0,2509	-0,2575	-0,2896	0,0487	0,0162	0,0708	0,2340
10_30	-0,2347	0,1579	0,3412	-0,0330	-0,0868	-0,1010	-0,2226
10_31	-0,2347	0,1579	0,3412	-0,0330	-0,0868	-0,1010	-0,2226
3_6	-0,1610		0,2159	-0,0712	-0,0182	-0,0625	-0,0258
5_26	-0,1169	-0,3335	-0,2647	0,0217	0,0334	0,1210	0,2855
4_11	-0,0940	-0,1133	-0,2200	0,0487	0,0300	0,0843	0,0404
4_8	-0,0940	-0,1133	-0,2200	0,0487	0,0300	0,0843	0,0404
3_12	-0,0940	-0,1133	-0,2200	0,0487	0,0300	0,0843	0,0404
4_13	-0,0940	-0,1133	-0,2200	0,0487	0,0300	0,0843	0,0404
4_12	-0,0940	-0,1133	-0,2200	0,0487	0,0300	0,0843	0,0404
5_12	-0,0940	-0,1133	-0,2200	0,0487	0,0300	0,0843	0,0404
4_9	-0,0940	-0,1133	-0,2200	0,0487	0,0300	0,0843	0,0404
10_6	0,0011			0,0051	-0,0145	0,0335	-0,0339
10_12	0,0940	0,1133	0,2200	-0,0487	-0,0300	-0,0843	-0,0404
6_12	0,0940	0,1133	0,2200	-0,0487	-0,0300	-0,0843	-0,0404
29_1	0,1494	-0,1133	-0,2260	0,0661	0,0480	0,1020	0,0637
6_6	0,1610		-0,2159	0,0712	0,0182	0,0625	0,0258
5_23	0,3014	-0,1284	0,0318	0,0863	0,0014	0,0149	0,1002
5_1	0,3480		0,2159	-0,0024	-0,0318	-0,0464	0,0735
5_15	0,3724	-0,1254	0,1827	-0,0318	-0,0506	-0,0090	0,2544
5_16	0,3724	-0,1254	0,1827	-0,0318	-0,0506	-0,0090	0,2544
3_1	0,4877		0,2159	-0,0051	-0,0256	-0,0032	0,0580
22_5	0,5302	-0,1133	-0,2247	0,0779	0,0645	0,1242	0,1344
3_8	0,5340	-0,0578	-0,2065	0,0153	0,0577	0,1342	0,1316
10_5	0,7657		-0,2159	0,0851	0,0670	0,1181	0,0858
6_5	0,7657		-0,2159	0,0851	0,0670	0,1181	0,0858
3_15	0,8053		0,2145	-0,0047	-0,0216	0,0247	0,1302
3_16	0,8053		0,2145	-0,0047	-0,0216	0,0247	0,1302
6_20	0,8404		0,2171	-0,0067	-0,0210	0,0261	0,1407
3_17	0,8404		0,2171	-0,0067	-0,0210	0,0261	0,1407
3_13	0,9050	-0,1191	-0,3125	0,1773	0,1570	0,1478	0,0666
10_32	1,0228	-0,1579	-0,3412	0,0852	0,1442	0,1637	0,2819
5_32	1,1290	0,0195	-0,1992	0,0159	0,1249	0,0847	0,2384
6_27	1,1635		-0,2085	0,0951	0,0877	0,1418	0,1524
6_28	1,1635		-0,2085	0,0951	0,0877	0,1418	0,1524
25_3	1,1722	-0,1062	-0,1811	0,1146	0,0895	0,1386	0,2302
25_7	1,1722	-0,1062	-0,1811	0,1146	0,0895	0,1386	0,2302
10_15	1,2268	-0,2962	-0,2418	0,1158	0,1464	0,2181	0,3145
10_16	1,2268	-0,2962	-0,2418	0,1158	0,1464	0,2181	0,3145
10_24	1,2268	-0,2962	-0,2418	0,1158	0,1464	0,2181	0,3145
10_25	1,2268	-0,2962	-0,2418	0,1158	0,1464	0,2181	0,3145
22_4	1,4536	-0,1155	-0,1915	0,1210	0,1047	0,1696	0,2794
5_30	1,4601	0,0195	0,2488	-0,0390	0,0588	0,0494	0,2201
5_31	1,4601	0,0195	0,2488	-0,0390	0,0588	0,0494	0,2201
13_4	1,5643	0,1133	-0,1985	0,0995	0,0934	0,1432	0,1521
13_5	1,5643	0,1133	-0,1985	0,0995	0,0934	0,1432	0,1521
13_6	1,5643	0,1133	-0,1985	0,0995	0,0934	0,1432	0,1521

Taula B.1: Suma estandarditzada i correlacions entre els diversos estadístics i les mesures de pèrdua d'informació obtingudes en aplicar les diverses versions de microagregació (3/5).

B. Correlacions entre estadístics i mesures amb diverses particions de variables

Estad.	Suma rel.	Mic1-12mul	Mic2-11mul	Mic3-10mul	Mic4-9mul	Mic5-8mul	Mic6-7mul
6_4	1,6189		-0,2159	0,0894	0,1347	0,1838	0,2087
10_27	1,6202	0,1133	-0,1857	0,1033	0,0983	0,1433	0,1514
10_28	1,6202	0,1133	-0,1857	0,1033	0,0983	0,1433	0,1514
5_17	1,6855	-0,1254	0,1473	0,0799	0,0513	0,1103	0,3205
3_30	1,7234		0,1621	0,0459	0,2039	0,0744	0,0698
3_31	1,7234		0,1621	0,0459	0,2039	0,0744	0,0698
6_10	1,7430	0,2979	0,1939	0,1455	0,0841	-0,0001	-0,1282
3_10	1,7430	0,2979	0,1939	0,1455	0,0841	-0,0001	-0,1282
28_1	1,7469	0,1133	0,1530	0,0422	0,0780	0,0762	0,1335
26_1	1,7469	0,1133	0,1530	0,0422	0,0780	0,0762	0,1335
26_5	1,7469	0,1133	0,1530	0,0422	0,0780	0,0762	0,1335
26_6	1,7469	0,1133	0,1530	0,0422	0,0780	0,0762	0,1335
26_2	1,7469	0,1133	0,1530	0,0422	0,0780	0,0762	0,1335
26_3	1,7469	0,1133	0,1530	0,0422	0,0780	0,0762	0,1335
15_4	1,7469	0,1133	0,1530	0,0422	0,0780	0,0762	0,1335
15_1	1,7469	0,1133	0,1530	0,0422	0,0780	0,0762	0,1335
15_2	1,7469	0,1133	0,1530	0,0422	0,0780	0,0762	0,1335
15_5	1,7469	0,1133	0,1530	0,0422	0,0780	0,0762	0,1335
15_6	1,7469	0,1133	0,1530	0,0422	0,0780	0,0762	0,1335
26_4	1,7469	0,1133	0,1530	0,0422	0,0780	0,0762	0,1335
26_7	1,7469	0,1133	0,1530	0,0422	0,0780	0,0762	0,1335
6_24	1,7806		-0,2282	0,0779	0,1257	0,1935	0,2902
6_25	1,7806		-0,2282	0,0779	0,1257	0,1935	0,2902
16_7	1,8406	0,0805	0,0400	0,1016	0,0641	0,0879	0,2216
16_3	1,8406	0,0805	0,0400	0,1016	0,0641	0,0879	0,2216
3_9	1,8933	0,2962	0,1770	0,1461	0,0909	0,0216	-0,0990
10_10	1,8933	0,2962	0,1770	0,1461	0,0909	0,0216	-0,0990
3_11	2,0286	0,2943	0,1593	0,1452	0,0961	0,0418	-0,0682
10_29	2,0311	-0,1133	-0,2241	0,1268	0,1498	0,2312	0,3583
6_1	2,0618		-0,2159	0,1317	0,1597	0,2131	0,2446
10_1	2,0618		-0,2159	0,1317	0,1597	0,2131	0,2446
10_4	2,0618		-0,2159	0,1317	0,1597	0,2131	0,2446
4_20	2,0990	-0,1133	-0,2523	0,1497	0,1735	0,2228	0,3571
13_7	2,0990	-0,1133	-0,2523	0,1497	0,1735	0,2228	0,3571
4_29	2,0990	-0,1133	-0,2523	0,1497	0,1735	0,2228	0,3571
5_29	2,0990	-0,1133	-0,2523	0,1497	0,1735	0,2228	0,3571
4_32	2,0990	-0,1133	-0,2523	0,1497	0,1735	0,2228	0,3571
14_4	2,0990	-0,1133	-0,2523	0,1497	0,1735	0,2228	0,3571
4_17	2,0990	-0,1133	-0,2523	0,1497	0,1735	0,2228	0,3571
13_3	2,0990	-0,1133	-0,2523	0,1497	0,1735	0,2228	0,3571
14_1	2,0990	-0,1133	-0,2523	0,1497	0,1735	0,2228	0,3571
14_2	2,0990	-0,1133	-0,2523	0,1497	0,1735	0,2228	0,3571
14_5	2,0990	-0,1133	-0,2523	0,1497	0,1735	0,2228	0,3571
14_6	2,0990	-0,1133	-0,2523	0,1497	0,1735	0,2228	0,3571
4_26	2,0990	-0,1133	-0,2523	0,1497	0,1735	0,2228	0,3571
6_15	2,1149		-0,2418	0,1158	0,1464	0,2181	0,3145
6_16	2,1149		-0,2418	0,1158	0,1464	0,2181	0,3145
5_13	2,2563	0,0195	-0,0524	0,1716	0,1594	0,1823	0,1638
13_2	2,2766		-0,2345	0,1352	0,1563	0,2179	0,3318
22_2	2,2766		-0,2345	0,1352	0,1563	0,2179	0,3318
25_2	2,2883		-0,2411	0,1391	0,1597	0,2179	0,3318
3_26	2,3394		-0,1788	0,1395	0,1441	0,2178	0,3232
6_26	2,3394		-0,1788	0,1395	0,1441	0,2178	0,3232

Taula B.1: Suma estandarditzada i correlacions entre els diversos estadístics i les mesures de pèrdua d'informació obtingudes en aplicar les diverses versions de microagregació (4/5).

Estad.	Suma rel.	Mic1-12mul	Mic2-11mul	Mic3-10mul	Mic4-9mul	Mic5-8mul	Mic6-7mul
25_1	2,3611	0,1062	-0,4753	0,1321	0,2054	0,2278	0,3551
3_35	2,3652		0,0650	0,1513	0,1308	0,1444	0,2525
6_35	2,3652		0,0650	0,1513	0,1308	0,1444	0,2525
3_20	2,3932		-0,2113	0,1302	0,1638	0,2302	0,3349
6_17	2,3932		-0,2113	0,1302	0,1638	0,2302	0,3349
27_1	2,4393	0,0216	-0,2363	0,1419	0,1706	0,2094	0,3538
16_1	2,5519	0,0093	-0,2831	0,1178	0,2022	0,2574	0,3628
6_32	2,6762		-0,1033	0,1249	0,2336	0,1683	0,3557
3_32	2,6762		-0,1033	0,1249	0,2336	0,1683	0,3557
22_3	2,6850	0,1155	-0,2429	0,1249	0,1739	0,2244	0,3402
10_35	2,7122	-0,0341	-0,2000	0,2258	0,2117	0,2132	0,3245
3_29	2,7359		-0,1335	0,1515	0,2052	0,2061	0,3531
6_29	2,7359		-0,1335	0,1515	0,2052	0,2061	0,3531
16_2	2,7507	0,1133	-0,1927	0,1513	0,1699	0,2056	0,3285
25_5	2,7673	0,1062	-0,2954	0,1526	0,1907	0,2358	0,3489
25_6	2,7673	0,1062	-0,2954	0,1526	0,1907	0,2358	0,3489
25_4	2,7673	0,1062	-0,2954	0,1526	0,1907	0,2358	0,3489
22_7	2,8209	0,1259	-0,2113	0,1426	0,1783	0,2185	0,3353
16_5	2,9455	0,1425	-0,2439	0,1460	0,1906	0,2320	0,3481
16_6	2,9455	0,1425	-0,2439	0,1460	0,1906	0,2320	0,3481
16_4	2,9455	0,1425	-0,2439	0,1460	0,1906	0,2320	0,3481
10_26	3,2813	0,2962	-0,2113	0,1302	0,1638	0,2302	0,3349
10_17	3,2813	0,2962	-0,2113	0,1302	0,1638	0,2302	0,3349
10_23	3,2813	0,2962	-0,2113	0,1302	0,1638	0,2302	0,3349
6_23	3,2967		0,1939	0,1719	0,1706	0,2122	0,3197
3_23	3,2967		0,1939	0,1719	0,1706	0,2122	0,3197

Taula B.1: Suma estandarditzada i correlacions entre els diversos estadístics i les mesures de pèrdua d'informació obtingudes en aplicar les diverses versions de microagregació (5/5).

Estad.	Suma rel.	Mic1-12mul	Mic2-11mul	Mic3-10mul	Mic4-9mul	Mic5-8mul	Mic6-7mul
5_8	-6,0000	-0,6418	-0,5751	-0,5785	-0,5238	-0,4130	-0,2098
18_4	-4,6583	-0,5234	-0,4127	-0,4465	-0,3999	-0,3085	-0,1768
18_1	-4,6583	-0,5234	-0,4127	-0,4465	-0,3999	-0,3085	-0,1768
18_2	-4,6583	-0,5234	-0,4127	-0,4465	-0,3999	-0,3085	-0,1768
18_5	-4,6583	-0,5234	-0,4127	-0,4465	-0,3999	-0,3085	-0,1768
18_6	-4,6583	-0,5234	-0,4127	-0,4465	-0,3999	-0,3085	-0,1768
4_11	-4,3804	-0,5234	-0,4154	-0,4424	-0,3976	-0,3026	-0,1230
4_13	-4,3804	-0,5234	-0,4154	-0,4424	-0,3976	-0,3026	-0,1230
4_8	-4,3804	-0,5234	-0,4154	-0,4424	-0,3976	-0,3026	-0,1230
4_12	-4,3804	-0,5234	-0,4154	-0,4424	-0,3976	-0,3026	-0,1230
5_12	-4,3804	-0,5234	-0,4154	-0,4424	-0,3976	-0,3026	-0,1230
4_9	-4,3804	-0,5234	-0,4154	-0,4424	-0,3976	-0,3026	-0,1230
3_12	-4,3804	-0,5234	-0,4154	-0,4424	-0,3976	-0,3026	-0,1230
22_6	-4,2270	-0,5395	-0,4124	-0,4821	-0,3403	-0,2094	-0,1425
19_4	-4,2149	-0,5286	-0,4116	-0,4270	-0,3436	-0,2357	-0,1491
19_1	-4,2149	-0,5286	-0,4116	-0,4270	-0,3436	-0,2357	-0,1491
19_2	-4,2149	-0,5286	-0,4116	-0,4270	-0,3436	-0,2357	-0,1491
19_5	-4,2149	-0,5286	-0,4116	-0,4270	-0,3436	-0,2357	-0,1491
19_6	-4,2149	-0,5286	-0,4116	-0,4270	-0,3436	-0,2357	-0,1491
29_1	-4,1687	-0,5234	-0,4126	-0,4240	-0,3829	-0,2860	-0,1005

Taula B.2: Suma estandarditzada i correlacions entre els diversos estadístics i les mesures de pèrdua de confidencialitat per enllaç de registres obtingudes en aplicar les diverses versions de microagregació multivariant amb k=10 (1/5).

Estad.	Suma rel.	Mic1-12mul	Mic2-11mul	Mic3-10mul	Mic4-9mul	Mic5-8mul	Mic6-7mul
3_13	-3,7103	-0,5800	-0,5312	-0,2748	-0,2865	-0,2749	-0,0410
5_11	-3,5675	-0,2836	-0,3590	-0,3133	-0,2930	-0,2482	-0,1677
22_5	-3,4391	-0,5234	-0,4008	-0,4006	-0,3575	-0,2450	0,0087
22_1	-3,1774	-0,4428	-0,3158	-0,3443	-0,2520	-0,1701	-0,0944
6_6	-3,1315		-0,3721	-0,3420	-0,3867	-0,2847	-0,0977
6_5	-2,9620		-0,3721	-0,3801	-0,3524	-0,2597	-0,0748
10_5	-2,9620		-0,3721	-0,3801	-0,3524	-0,2597	-0,0748
6_30	-2,8555		-0,2822	-0,3411	-0,3947	-0,1486	-0,1388
6_31	-2,8555		-0,2822	-0,3411	-0,3947	-0,1486	-0,1388
5_13	-2,7760	-0,2609	-0,2620	-0,2467	-0,2343	-0,2421	-0,0952
10_6	-2,7587			-0,4437	-0,4204	-0,2582	-0,1183
6_2	-2,4779		-0,3721	-0,4437	-0,3256	-0,1805	-0,0011
3_8	-2,4043	-0,2577	-0,0770	-0,1735	-0,2089	-0,2078	-0,1399
10_14	-2,3876	-0,2723	-0,1794	-0,2324	-0,2106	-0,1777	-0,0875
22_4	-2,3275	-0,4428	-0,3651	-0,3297	-0,2909	-0,1571	0,1055
5_26	-2,3234	-0,2836	-0,4271	-0,4143	-0,3045	-0,1135	0,0909
10_32	-2,3008	-0,6187	-0,4563	-0,4019	-0,1694	-0,0054	0,1023
6_4	-2,1342		-0,3721	-0,3850	-0,2799	-0,1504	0,0161
6_18	-1,9816		-0,3719	-0,4413	-0,3139	-0,1416	0,0776
6_19	-1,9816		-0,3719	-0,4413	-0,3139	-0,1416	0,0776
10_29	-1,9696	-0,5234	-0,3929	-0,2818	-0,1896	-0,0349	0,0970
10_15	-1,8815	-0,2093	-0,4078	-0,3011	-0,2160	-0,0844	0,0610
10_16	-1,8815	-0,2093	-0,4078	-0,3011	-0,2160	-0,0844	0,0610
10_24	-1,8815	-0,2093	-0,4078	-0,3011	-0,2160	-0,0844	0,0610
10_25	-1,8815	-0,2093	-0,4078	-0,3011	-0,2160	-0,0844	0,0610
6_24	-1,8508		-0,3899	-0,4076	-0,2973	-0,1274	0,0855
6_25	-1,8508		-0,3899	-0,4076	-0,2973	-0,1274	0,0855
16_1	-1,8388	-0,5138	-0,2677	-0,2784	-0,1417	-0,0189	0,0471
25_7	-1,7491	-0,2908	-0,2721	-0,2765	-0,2816	-0,1668	0,1251
25_3	-1,7491	-0,2908	-0,2721	-0,2765	-0,2816	-0,1668	0,1251
6_27	-1,7149		-0,3527	-0,3397	-0,2934	-0,1572	0,0894
6_28	-1,7149		-0,3527	-0,3397	-0,2934	-0,1572	0,0894
6_1	-1,6958		-0,3721	-0,2747	-0,2053	-0,1106	0,0180
10_1	-1,6958		-0,3721	-0,2747	-0,2053	-0,1106	0,0180
10_4	-1,6958		-0,3721	-0,2747	-0,2053	-0,1106	0,0180
6_15	-1,5554		-0,4078	-0,3011	-0,2160	-0,0844	0,0610
6_16	-1,5554		-0,4078	-0,3011	-0,2160	-0,0844	0,0610
3_20	-1,3464		-0,3639	-0,2831	-0,2104	-0,0741	0,0748
6_17	-1,3464		-0,3639	-0,2831	-0,2104	-0,0741	0,0748
4_17	-1,2583	-0,5234	-0,3261	-0,1343	-0,0770	0,0089	0,1011
4_20	-1,2583	-0,5234	-0,3261	-0,1343	-0,0770	0,0089	0,1011
14_4	-1,2583	-0,5234	-0,3261	-0,1343	-0,0770	0,0089	0,1011
4_29	-1,2583	-0,5234	-0,3261	-0,1343	-0,0770	0,0089	0,1011
5_29	-1,2583	-0,5234	-0,3261	-0,1343	-0,0770	0,0089	0,1011
4_32	-1,2583	-0,5234	-0,3261	-0,1343	-0,0770	0,0089	0,1011
13_7	-1,2583	-0,5234	-0,3261	-0,1343	-0,0770	0,0089	0,1011
4_26	-1,2583	-0,5234	-0,3261	-0,1343	-0,0770	0,0089	0,1011
13_3	-1,2583	-0,5234	-0,3261	-0,1343	-0,0770	0,0089	0,1011
14_1	-1,2583	-0,5234	-0,3261	-0,1343	-0,0770	0,0089	0,1011
14_2	-1,2583	-0,5234	-0,3261	-0,1343	-0,0770	0,0089	0,1011
14_5	-1,2583	-0,5234	-0,3261	-0,1343	-0,0770	0,0089	0,1011
14_6	-1,2583	-0,5234	-0,3261	-0,1343	-0,0770	0,0089	0,1011
5_20	-1,1901	0,1368	-0,4266	-0,2785	-0,2205	-0,0884	0,0954
10_17	-1,0203	0,2093	-0,3639	-0,2831	-0,2104	-0,0741	0,0748
10_26	-1,0203	0,2093	-0,3639	-0,2831	-0,2104	-0,0741	0,0748

Taula B.2: Suma estandarditzada i correlacions entre els diversos estadístics i les mesures de pèrdua de confidencialitat per enllaç de registres obtingudes en aplicar les diverses versions de microagregació multivariant amb k=10 (2/5).

Estad.	Suma rel.	Mic1-12mul	Mic2-11mul	Mic3-10mul	Mic4-9mul	Mic5-8mul	Mic6-7mul
10_23	-1,0203	0,2093	-0,3639	-0,2831	-0,2104	-0,0741	0,0748
5_14	-0,9649	-0,1111	-0,0750	-0,0942	-0,0635	-0,0583	-0,0495
25_2	-0,8465		-0,3434	-0,2163	-0,1590	-0,0367	0,1084
13_2	-0,8416		-0,3346	-0,2201	-0,1610	-0,0367	0,1084
22_2	-0,8416		-0,3346	-0,2201	-0,1610	-0,0367	0,1084
5_9	-0,7830	0,1368	-0,0018	-0,0323	-0,0774	-0,1107	-0,1094
13_4	-0,6540	0,5234	-0,3283	-0,3154	-0,2696	-0,1306	0,1001
13_5	-0,6540	0,5234	-0,3283	-0,3154	-0,2696	-0,1306	0,1001
13_6	-0,6540	0,5234	-0,3283	-0,3154	-0,2696	-0,1306	0,1001
10_35	-0,6058	-0,2723	-0,2529	-0,1369	-0,0919	0,0296	0,1255
5_33	-0,4862	-0,1111	-0,0750	-0,0474	0,0380	0,0006	-0,0366
5_34	-0,4862	-0,1111	-0,0750	-0,0474	0,0380	0,0006	-0,0366
10_27	-0,4105	0,5234	-0,2989	-0,2886	-0,2445	-0,1055	0,1080
10_28	-0,4105	0,5234	-0,2989	-0,2886	-0,2445	-0,1055	0,1080
6_8	-0,3208	-0,2093	-0,2140	-0,0607	0,0189	0,0707	0,0576
10_8	-0,3208	-0,2093	-0,2140	-0,0607	0,0189	0,0707	0,0576
10_11	-0,3208	-0,2093	-0,2140	-0,0607	0,0189	0,0707	0,0576
27_1	-0,2662	-0,5000	-0,1141	0,0425	0,0203	0,0439	0,1034
5_15	-0,2286	-0,6418	0,3077	0,4170	0,1864	0,0292	-0,1911
5_16	-0,2286	-0,6418	0,3077	0,4170	0,1864	0,0292	-0,1911
3_14	-0,2050	-0,0148	0,0490	-0,0093	-0,0063	-0,0289	-0,0355
6_14	-0,2050	-0,0148	0,0490	-0,0093	-0,0063	-0,0289	-0,0355
3_26	-0,1897		-0,1439	-0,0429	-0,1094	-0,0194	0,0819
6_26	-0,1897		-0,1439	-0,0429	-0,1094	-0,0194	0,0819
3_33	-0,1807		0,0490	-0,0220	0,0195	-0,0083	-0,0514
3_34	-0,1807		0,0490	-0,0220	0,0195	-0,0083	-0,0514
6_33	-0,1807		0,0490	-0,0220	0,0195	-0,0083	-0,0514
6_34	-0,1807		0,0490	-0,0220	0,0195	-0,0083	-0,0514
5_35	-0,1749	-0,1111	-0,0750	-0,0129	-0,1135	-0,0773	0,1164
3_7	-0,1605			-0,0204	0,0224	0,0053	-0,0379
6_7	-0,1605			-0,0204	0,0224	0,0053	-0,0379
25_4	-0,1147	0,2908	-0,4069	-0,1573	-0,0648	0,0312	0,0964
25_5	-0,1147	0,2908	-0,4069	-0,1573	-0,0648	0,0312	0,0964
25_6	-0,1147	0,2908	-0,4069	-0,1573	-0,0648	0,0312	0,0964
5_3	-0,1066			-0,0400	0,0665	0,0080	-0,0385
6_11	-0,0312	-0,2042	-0,2058	-0,0423	0,0451	0,1006	0,0814
3_11	0,0312	0,2042	0,2058	0,0423	-0,0451	-0,1006	-0,0814
13_1	0,0576	0,5234	0,1773	-0,1537	-0,1093	-0,0807	-0,0832
3_3	0,0727			-0,0400	0,0599	0,0485	-0,0188
6_3	0,0727			-0,0400	0,0599	0,0485	-0,0188
5_7	0,1061			-0,0204	0,0694	0,0327	-0,0147
16_5	0,3006	0,3205	-0,2242	-0,1273	-0,0567	0,0267	0,0954
16_6	0,3006	0,3205	-0,2242	-0,1273	-0,0567	0,0267	0,0954
16_4	0,3006	0,3205	-0,2242	-0,1273	-0,0567	0,0267	0,0954
3_9	0,3208	0,2093	0,2140	0,0607	-0,0189	-0,0707	-0,0576
10_10	0,3208	0,2093	0,2140	0,0607	-0,0189	-0,0707	-0,0576
3_35	0,3519		0,0490	-0,0306	-0,1434	-0,0238	0,1365
6_35	0,3519		0,0490	-0,0306	-0,1434	-0,0238	0,1365
6_21	0,3595		0,2209	-0,0259	0,0627	0,0297	-0,0360
6_22	0,3595		0,2209	-0,0259	0,0627	0,0297	-0,0360
3_21	0,3595		0,2209	-0,0259	0,0627	0,0297	-0,0360
3_22	0,3595		0,2209	-0,0259	0,0627	0,0297	-0,0360
6_23	0,3913		0,2209	0,0461	-0,0884	-0,0701	0,0558
3_23	0,3913		0,2209	0,0461	-0,0884	-0,0701	0,0558
5_32	0,4095	-0,2609	-0,1185	0,0886	0,1583	0,0715	0,0825

Taula B.2: Suma estandarditzada i correlacions entre els diversos estadístics i les mesures de pèrdua de confidencialitat per enllaç de registres obtingudes en aplicar les diverses versions de microagregació multivariant amb k=10 (3/5).

B. Correlacions entre estadístics i mesures amb diverses particions de variables

Estad.	Suma rel.	Mic1-12mul	Mic2-11mul	Mic3-10mul	Mic4-9mul	Mic5-8mul	Mic6-7mul
5_24	0,4196	-0,2836	0,3150	0,2977	0,1929	0,0388	-0,1391
5_25	0,4196	-0,2836	0,3150	0,2977	0,1929	0,0388	-0,1391
10_30	0,5589	0,6187	0,4563	-0,1594	-0,2496	-0,0593	-0,0635
10_31	0,5589	0,6187	0,4563	-0,1594	-0,2496	-0,0593	-0,0635
6_10	0,6384	0,2144	0,2209	0,0788	0,0084	-0,0375	-0,0296
3_10	0,6384	0,2144	0,2209	0,0788	0,0084	-0,0375	-0,0296
4_15	0,6540	-0,5234	0,3283	0,3154	0,2696	0,1306	-0,1001
4_16	0,6540	-0,5234	0,3283	0,3154	0,2696	0,1306	-0,1001
4_27	0,6540	-0,5234	0,3283	0,3154	0,2696	0,1306	-0,1001
4_28	0,6540	-0,5234	0,3283	0,3154	0,2696	0,1306	-0,1001
5_27	0,6540	-0,5234	0,3283	0,3154	0,2696	0,1306	-0,1001
5_28	0,6540	-0,5234	0,3283	0,3154	0,2696	0,1306	-0,1001
4_24	0,6540	-0,5234	0,3283	0,3154	0,2696	0,1306	-0,1001
4_25	0,6540	-0,5234	0,3283	0,3154	0,2696	0,1306	-0,1001
4_30	0,6540	-0,5234	0,3283	0,3154	0,2696	0,1306	-0,1001
4_31	0,6540	-0,5234	0,3283	0,3154	0,2696	0,1306	-0,1001
4_18	0,6540	-0,5234	0,3283	0,3154	0,2696	0,1306	-0,1001
4_19	0,6540	-0,5234	0,3283	0,3154	0,2696	0,1306	-0,1001
22_3	0,7708	0,4428	-0,1869	-0,0585	-0,0394	0,0322	0,1057
25_1	0,7799	0,2908	-0,5680	0,1501	0,1832	0,1549	0,0693
22_7	0,9796	0,4794	-0,1665	-0,0465	-0,0192	0,0454	0,1110
5_18	1,0837	0,1368	0,3169	0,2147	0,1583	0,0317	-0,0903
5_19	1,0837	0,1368	0,3169	0,2147	0,1583	0,0317	-0,0903
3_29	1,1069		-0,0283	0,1329	0,1128	0,0853	0,1058
6_29	1,1069		-0,0283	0,1329	0,1128	0,0853	0,1058
10_33	1,1387	-0,2723	0,3789	0,3024	0,2231	0,0903	-0,0552
10_34	1,1387	-0,2723	0,3789	0,3024	0,2231	0,0903	-0,0552
5_21	1,1942	0,4155	0,2544	0,0423	0,1030	0,0230	-0,0464
5_22	1,1942	0,4155	0,2544	0,0423	0,1030	0,0230	-0,0464
16_2	1,2195	0,5234	-0,1714	-0,0049	0,0028	0,0533	0,1208
5_17	1,3091	-0,6418	0,2689	0,4022	0,2271	0,1376	0,0796
3_18	1,5554		0,4078	0,3011	0,2160	0,0844	-0,0610
3_19	1,5554		0,4078	0,3011	0,2160	0,0844	-0,0610
5_2	1,5619		0,3721	0,2520	0,2079	0,0968	-0,0319
5_10	1,5858	0,4155	0,2544	0,1972	0,1239	0,0370	-0,0359
5_23	1,6076	0,4155	0,2544	0,1371	-0,0117	-0,0007	0,0640
10_3	1,6958		0,3721	0,2747	0,2053	0,1106	-0,0180
3_2	1,6958		0,3721	0,2747	0,2053	0,1106	-0,0180
3_27	1,7149		0,3527	0,3397	0,2934	0,1572	-0,0894
3_28	1,7149		0,3527	0,3397	0,2934	0,1572	-0,0894
5_30	1,8010	-0,2609	0,3781	0,4093	0,3891	0,0705	-0,0149
5_31	1,8010	-0,2609	0,3781	0,4093	0,3891	0,0705	-0,0149
3_24	1,8508		0,3899	0,4076	0,2973	0,1274	-0,0855
3_25	1,8508		0,3899	0,4076	0,2973	0,1274	-0,0855
10_21	1,8815	0,2093	0,4078	0,3011	0,2160	0,0844	-0,0610
10_22	1,8815	0,2093	0,4078	0,3011	0,2160	0,0844	-0,0610
3_32	1,9730		0,0663	0,2548	0,2771	0,1074	0,1317
6_32	1,9730		0,0663	0,2548	0,2771	0,1074	0,1317
3_15	1,9816		0,3719	0,4413	0,3139	0,1416	-0,0776
3_16	1,9816		0,3719	0,4413	0,3139	0,1416	-0,0776
5_4	2,0934		0,3721	0,3641	0,2838	0,1584	-0,0227
3_4	2,1342		0,3721	0,3850	0,2799	0,1504	-0,0161
10_7	2,1962		0,3721	0,3152	0,2580	0,1665	0,0228
6_20	2,2450		0,3723	0,4454	0,3260	0,1658	-0,0411
3_17	2,2450		0,3723	0,4454	0,3260	0,1658	-0,0411

Taula B.2: Suma estandarditzada i correlacions entre els diversos estadístics i les mesures de pèrdua de confidencialitat per enllaç de registres obtingudes en aplicar les diverses versions de microagregació multivariant amb k=10 (4/5).

Estad.	Suma rel.	Mic1-12mul	Mic2-11mul	Mic3-10mul	Mic4-9mul	Mic5-8mul	Mic6-7mul
6_9	2,4043	0,2577	0,0770	0,1735	0,2089	0,2078	0,1399
3_1	2,4779		0,3721	0,4437	0,3256	0,1805	0,0011
5_1	2,7858		0,3721	0,4896	0,3360	0,2325	0,0185
5_6	2,8527		0,3721	0,3759	0,3679	0,2061	0,0744
3_30	2,8555		0,2822	0,3411	0,3947	0,1486	0,1388
3_31	2,8555		0,2822	0,3411	0,3947	0,1486	0,1388
4_2	2,9620		0,3721	0,3801	0,3524	0,2597	0,0748
4_6	2,9620		0,3721	0,3801	0,3524	0,2597	0,0748
3_5	2,9620		0,3721	0,3801	0,3524	0,2597	0,0748
4_5	2,9620		0,3721	0,3801	0,3524	0,2597	0,0748
5_5	2,9620		0,3721	0,3801	0,3524	0,2597	0,0748
4_1	2,9620		0,3721	0,3801	0,3524	0,2597	0,0748
4_4	2,9620		0,3721	0,3801	0,3524	0,2597	0,0748
16_7	2,9781	0,5793	0,0560	0,3013	0,1632	0,1186	0,1801
16_3	2,9781	0,5793	0,0560	0,3013	0,1632	0,1186	0,1801
3_6	3,1315		0,3721	0,3420	0,3867	0,2847	0,0977
28_1	3,6334	0,5234	0,3279	0,3412	0,2513	0,1842	0,1536
26_1	3,6334	0,5234	0,3279	0,3412	0,2513	0,1842	0,1536
26_7	3,6334	0,5234	0,3279	0,3412	0,2513	0,1842	0,1536
26_3	3,6334	0,5234	0,3279	0,3412	0,2513	0,1842	0,1536
26_5	3,6334	0,5234	0,3279	0,3412	0,2513	0,1842	0,1536
26_6	3,6334	0,5234	0,3279	0,3412	0,2513	0,1842	0,1536
26_2	3,6334	0,5234	0,3279	0,3412	0,2513	0,1842	0,1536
15_4	3,6334	0,5234	0,3279	0,3412	0,2513	0,1842	0,1536
26_4	3,6334	0,5234	0,3279	0,3412	0,2513	0,1842	0,1536
15_1	3,6334	0,5234	0,3279	0,3412	0,2513	0,1842	0,1536
15_2	3,6334	0,5234	0,3279	0,3412	0,2513	0,1842	0,1536
15_5	3,6334	0,5234	0,3279	0,3412	0,2513	0,1842	0,1536
15_6	3,6334	0,5234	0,3279	0,3412	0,2513	0,1842	0,1536
6_13	3,7103	0,5800	0,5312	0,2748	0,2865	0,2749	0,0410
10_12	4,3804	0,5234	0,4154	0,4424	0,3976	0,3026	0,1230
6_12	4,3804	0,5234	0,4154	0,4424	0,3976	0,3026	0,1230
10_13	4,5991	0,6187	0,4563	0,4788	0,4508	0,3216	0,0786

Taula B.2: Suma estandarditzada i correlacions entre els diversos estadístics i les mesures de pèrdua de confidencialitat per enllaç de registres obtingudes en aplicar les diverses versions de microagregació multivariant amb k=10 (5/5).

Estad.	Suma rel.	Mic1-12mul	Mic2-11mul	Mic3-10mul	Mic4-9mul	Mic5-8mul	Mic6-7mul
19_4	-5,6423	-0,6272	-0,6725	-0,6556	-0,6340	-0,5905	-0,5729
19_1	-5,6423	-0,6272	-0,6725	-0,6556	-0,6340	-0,5905	-0,5729
19_2	-5,6423	-0,6272	-0,6725	-0,6556	-0,6340	-0,5905	-0,5729
19_5	-5,6423	-0,6272	-0,6725	-0,6556	-0,6340	-0,5905	-0,5729
19_6	-5,6423	-0,6272	-0,6725	-0,6556	-0,6340	-0,5905	-0,5729
5_8	-5,2933	-0,7308	-0,7087	-0,6914	-0,6454	-0,5167	-0,3010
18_4	-5,2646	-0,6172	-0,6385	-0,6222	-0,5961	-0,5287	-0,5116
18_1	-5,2646	-0,6172	-0,6385	-0,6222	-0,5961	-0,5287	-0,5116
18_2	-5,2646	-0,6172	-0,6385	-0,6222	-0,5961	-0,5287	-0,5116
18_5	-5,2646	-0,6172	-0,6385	-0,6222	-0,5961	-0,5287	-0,5116
18_6	-5,2646	-0,6172	-0,6385	-0,6222	-0,5961	-0,5287	-0,5116
4_13	-4,3890	-0,6172	-0,6276	-0,5975	-0,5456	-0,4059	-0,2028
4_12	-4,3890	-0,6172	-0,6276	-0,5975	-0,5456	-0,4059	-0,2028
5_12	-4,3890	-0,6172	-0,6276	-0,5975	-0,5456	-0,4059	-0,2028
4_9	-4,3890	-0,6172	-0,6276	-0,5975	-0,5456	-0,4059	-0,2028

Taula B.3: Suma estandarditzada i correlacions entre els diversos estadístics i les mesures de pèrdua de confidencialitat per intervals de confiança, basats en rangs, obtingudes en aplicar les diverses versions de microagregació multivariant amb k=10 (1/5).

Estad.	Suma rel.	Mic1-12mul	Mic2-11mul	Mic3-10mul	Mic4-9mul	Mic5-8mul	Mic6-7mul
4_8	-4,3890	-0,6172	-0,6276	-0,5975	-0,5456	-0,4059	-0,2028
3_12	-4,3890	-0,6172	-0,6276	-0,5975	-0,5456	-0,4059	-0,2028
4_11	-4,3890	-0,6172	-0,6276	-0,5975	-0,5456	-0,4059	-0,2028
3_13	-4,0385	-0,7030	-0,7371	-0,4523	-0,4417	-0,3416	-0,1293
29_1	-3,8904	-0,6172	-0,5762	-0,5438	-0,4875	-0,3361	-0,1209
22_1	-3,4948	-0,5213	-0,5360	-0,4729	-0,3708	-0,2823	-0,2100
22_6	-3,3923	-0,6401	-0,3694	-0,5125	-0,4070	-0,2670	-0,1441
5_13	-3,3813	-0,4926	-0,5529	-0,4476	-0,4104	-0,3084	-0,1133
22_5	-2,8345	-0,6172	-0,5268	-0,4907	-0,4161	-0,2095	0,2155
6_6	-2,5673		-0,4429	-0,4180	-0,4435	-0,2901	-0,1051
10_5	-2,3590		-0,4429	-0,4489	-0,4034	-0,2509	-0,0337
6_5	-2,3590		-0,4429	-0,4489	-0,4034	-0,2509	-0,0337
3_8	-2,2129	-0,3730	-0,1820	-0,2635	-0,2663	-0,2533	-0,1535
22_4	-2,2048	-0,5213	-0,5149	-0,4311	-0,3332	-0,0804	0,2486
16_1	-2,1052	-0,7538	-0,2647	-0,3803	-0,2188	-0,0220	0,0806
6_2	-2,1030		-0,4429	-0,4824	-0,3467	-0,1635	0,0056
10_14	-2,0758	-0,2434	-0,2759	-0,2709	-0,2686	-0,2171	-0,1234
10_29	-2,0672	-0,6172	-0,5072	-0,3795	-0,2242	0,0131	0,1613
10_32	-1,9378	-0,7193	-0,6160	-0,5137	-0,1713	0,1960	0,2813
6_4	-1,8637		-0,4429	-0,4463	-0,3287	-0,1262	0,0605
6_18	-1,8381		-0,4420	-0,4781	-0,3310	-0,1147	0,0918
6_19	-1,8381		-0,4420	-0,4781	-0,3310	-0,1147	0,0918
10_6	-1,8209			-0,4824	-0,4430	-0,1944	-0,0616
6_1	-1,5291		-0,4429	-0,3406	-0,2576	-0,0885	0,0649
10_1	-1,5291		-0,4429	-0,3406	-0,2576	-0,0885	0,0649
10_4	-1,5291		-0,4429	-0,3406	-0,2576	-0,0885	0,0649
6_24	-1,4655		-0,4545	-0,4580	-0,3190	-0,0481	0,2230
6_25	-1,4655		-0,4545	-0,4580	-0,3190	-0,0481	0,2230
5_26	-1,4646	-0,0146	-0,4657	-0,4452	-0,2848	-0,0327	0,1869
3_20	-1,4505		-0,4536	-0,3695	-0,2740	-0,0602	0,1294
6_17	-1,4505		-0,4536	-0,3695	-0,2740	-0,0602	0,1294
14_4	-1,4398	-0,6172	-0,4633	-0,2178	-0,0836	0,0745	0,1682
4_26	-1,4398	-0,6172	-0,4633	-0,2178	-0,0836	0,0745	0,1682
13_3	-1,4398	-0,6172	-0,4633	-0,2178	-0,0836	0,0745	0,1682
14_1	-1,4398	-0,6172	-0,4633	-0,2178	-0,0836	0,0745	0,1682
14_2	-1,4398	-0,6172	-0,4633	-0,2178	-0,0836	0,0745	0,1682
14_5	-1,4398	-0,6172	-0,4633	-0,2178	-0,0836	0,0745	0,1682
14_6	-1,4398	-0,6172	-0,4633	-0,2178	-0,0836	0,0745	0,1682
4_32	-1,4398	-0,6172	-0,4633	-0,2178	-0,0836	0,0745	0,1682
4_29	-1,4398	-0,6172	-0,4633	-0,2178	-0,0836	0,0745	0,1682
5_29	-1,4398	-0,6172	-0,4633	-0,2178	-0,0836	0,0745	0,1682
13_7	-1,4398	-0,6172	-0,4633	-0,2178	-0,0836	0,0745	0,1682
4_20	-1,4398	-0,6172	-0,4633	-0,2178	-0,0836	0,0745	0,1682
4_17	-1,4398	-0,6172	-0,4633	-0,2178	-0,0836	0,0745	0,1682
10_15	-1,4223	-0,1827	-0,4666	-0,3513	-0,2347	-0,0089	0,1893
10_16	-1,4223	-0,1827	-0,4666	-0,3513	-0,2347	-0,0089	0,1893
10_24	-1,4223	-0,1827	-0,4666	-0,3513	-0,2347	-0,0089	0,1893
10_25	-1,4223	-0,1827	-0,4666	-0,3513	-0,2347	-0,0089	0,1893
6_30	-1,3779		-0,2974	-0,3034	-0,2774	-0,0760	0,0132
6_31	-1,3779		-0,2974	-0,3034	-0,2774	-0,0760	0,0132
5_11	-1,2909	-0,0146	-0,0951	-0,1102	-0,1747	-0,2153	-0,1996
25_3	-1,2638	-0,3407	-0,3874	-0,3298	-0,2586	-0,0186	0,3466
25_7	-1,2638	-0,3407	-0,3874	-0,3298	-0,2586	-0,0186	0,3466
10_17	-1,2176	0,1827	-0,4536	-0,3695	-0,2740	-0,0602	0,1294
10_26	-1,2176	0,1827	-0,4536	-0,3695	-0,2740	-0,0602	0,1294

Taula B.3: Suma estandarditzada i correlacions entre els diversos estadístics i les mesures de pèrdua de confidencialitat per intervals de confiança, basats en rangs, obtingudes en aplicar les diverses versions de microagregació multivariant amb k=10 (2/5).

Estad.	Suma rel.	Mic1-12mul	Mic2-11mul	Mic3-10mul	Mic4-9mul	Mic5-8mul	Mic6-7mul
10_23	-1,2176	0,1827	-0,4536	-0,3695	-0,2740	-0,0602	0,1294
6_15	-1,1894		-0,4666	-0,3513	-0,2347	-0,0089	0,1893
6_16	-1,1894		-0,4666	-0,3513	-0,2347	-0,0089	0,1893
10_35	-1,1824	-0,2434	-0,4303	-0,2978	-0,1517	0,0513	0,1664
3_11	-0,7971	0,1757	0,0502	-0,0850	-0,1819	-0,2109	-0,1874
6_27	-0,7793		-0,3949	-0,3728	-0,2727	-0,0024	0,4137
6_28	-0,7793		-0,3949	-0,3728	-0,2727	-0,0024	0,4137
27_1	-0,7661	-0,7848	-0,2280	-0,0223	0,0432	0,1220	0,1729
25_2	-0,6522		-0,4310	-0,2760	-0,1477	0,0742	0,2491
13_2	-0,6386		-0,4196	-0,2774	-0,1477	0,0742	0,2491
22_2	-0,6386		-0,4196	-0,2774	-0,1477	0,0742	0,2491
3_9	-0,5930	0,1827	0,0648	-0,0595	-0,1517	-0,1778	-0,1669
10_10	-0,5930	0,1827	0,0648	-0,0595	-0,1517	-0,1778	-0,1669
6_26	-0,4584		-0,2670	-0,1112	-0,1315	0,0080	0,1460
3_26	-0,4584		-0,2670	-0,1112	-0,1315	0,0080	0,1460
6_10	-0,3566	0,1898	0,0793	-0,0329	-0,1179	-0,1385	-0,1382
3_10	-0,3566	0,1898	0,0793	-0,0329	-0,1179	-0,1385	-0,1382
4_24	-0,3500	-0,6172	0,3395	0,3294	0,2251	-0,0520	-0,4361
4_25	-0,3500	-0,6172	0,3395	0,3294	0,2251	-0,0520	-0,4361
4_30	-0,3500	-0,6172	0,3395	0,3294	0,2251	-0,0520	-0,4361
4_31	-0,3500	-0,6172	0,3395	0,3294	0,2251	-0,0520	-0,4361
4_27	-0,3500	-0,6172	0,3395	0,3294	0,2251	-0,0520	-0,4361
4_28	-0,3500	-0,6172	0,3395	0,3294	0,2251	-0,0520	-0,4361
5_27	-0,3500	-0,6172	0,3395	0,3294	0,2251	-0,0520	-0,4361
5_28	-0,3500	-0,6172	0,3395	0,3294	0,2251	-0,0520	-0,4361
4_18	-0,3500	-0,6172	0,3395	0,3294	0,2251	-0,0520	-0,4361
4_19	-0,3500	-0,6172	0,3395	0,3294	0,2251	-0,0520	-0,4361
4_15	-0,3500	-0,6172	0,3395	0,3294	0,2251	-0,0520	-0,4361
4_16	-0,3500	-0,6172	0,3395	0,3294	0,2251	-0,0520	-0,4361
5_20	-0,1847	0,4710	-0,4078	-0,2779	-0,1647	0,0336	0,2112
5_15	-0,1564	-0,7308	0,3621	0,4617	0,2165	-0,0559	-0,3581
5_16	-0,1564	-0,7308	0,3621	0,4617	0,2165	-0,0559	-0,3581
6_23	-0,1139		0,0793	-0,0182	-0,1625	-0,0687	0,0991
3_23	-0,1139		0,0793	-0,0182	-0,1625	-0,0687	0,0991
6_21	-0,0903		0,0793	0,0015	0,0733	-0,0323	-0,1484
6_22	-0,0903		0,0793	0,0015	0,0733	-0,0323	-0,1484
3_21	-0,0903		0,0793	0,0015	0,0733	-0,0323	-0,1484
3_22	-0,0903		0,0793	0,0015	0,0733	-0,0323	-0,1484
25_4	-0,0707	0,3407	-0,4663	-0,2193	-0,0621	0,1129	0,2006
25_5	-0,0707	0,3407	-0,4663	-0,2193	-0,0621	0,1129	0,2006
25_6	-0,0707	0,3407	-0,4663	-0,2193	-0,0621	0,1129	0,2006
5_3	0,0128			0,0076	0,0612	-0,0124	-0,0413
3_35	0,0338		0,0655	-0,0616	-0,1732	-0,0101	0,1831
6_35	0,0338		0,0655	-0,0616	-0,1732	-0,0101	0,1831
3_3	0,1163			0,0076	0,1071	0,0315	-0,0653
6_3	0,1163			0,0076	0,1071	0,0315	-0,0653
5_7	0,1216			0,0259	0,0635	0,0134	-0,0212
3_7	0,2029			0,0259	0,0749	0,0325	-0,0032
6_7	0,2029			0,0259	0,0749	0,0325	-0,0032
3_14	0,2489	0,0991	0,0655	0,0546	0,0207	-0,0112	-0,0334
6_14	0,2489	0,0991	0,0655	0,0546	0,0207	-0,0112	-0,0334
3_33	0,2632		0,0655	0,0367	0,0782	0,0239	-0,0232
3_34	0,2632		0,0655	0,0367	0,0782	0,0239	-0,0232
6_33	0,2632		0,0655	0,0367	0,0782	0,0239	-0,0232
6_34	0,2632		0,0655	0,0367	0,0782	0,0239	-0,0232

Taula B.3: Suma estandarditzada i correlacions entre els diversos estadístics i les mesures de pèrdua de confidencialitat per intervals de confiança, basats en rangs, obtingudes en aplicar les diverses versions de microagregació multivariant amb k=10 (3/5).

Estad.	Suma rel.	Mic1-12mul	Mic2-11mul	Mic3-10mul	Mic4-9mul	Mic5-8mul	Mic6-7mul
16_5	0,3166	0,2883	-0,2085	-0,1691	-0,0428	0,1136	0,2008
16_6	0,3166	0,2883	-0,2085	-0,1691	-0,0428	0,1136	0,2008
16_4	0,3166	0,2883	-0,2085	-0,1691	-0,0428	0,1136	0,2008
13_5	0,3500	0,6172	-0,3395	-0,3294	-0,2251	0,0520	0,4361
13_6	0,3500	0,6172	-0,3395	-0,3294	-0,2251	0,0520	0,4361
13_4	0,3500	0,6172	-0,3395	-0,3294	-0,2251	0,0520	0,4361
5_32	0,4275	-0,4926	-0,1944	0,0764	0,2595	0,2296	0,2393
6_8	0,5930	-0,1827	-0,0648	0,0595	0,1517	0,1778	0,1669
10_8	0,5930	-0,1827	-0,0648	0,0595	0,1517	0,1778	0,1669
10_11	0,5930	-0,1827	-0,0648	0,0595	0,1517	0,1778	0,1669
10_27	0,6854	0,6172	-0,2772	-0,2833	-0,1772	0,1008	0,4519
10_28	0,6854	0,6172	-0,2772	-0,2833	-0,1772	0,1008	0,4519
3_29	0,7294		-0,0905	0,0434	0,1303	0,1648	0,1768
6_29	0,7294		-0,0905	0,0434	0,1303	0,1648	0,1768
3_27	0,7793		0,3949	0,3728	0,2727	0,0024	-0,4137
3_28	0,7793		0,3949	0,3728	0,2727	0,0024	-0,4137
5_33	0,7903	0,3360	0,1317	0,0603	0,0807	0,0169	-0,0329
5_34	0,7903	0,3360	0,1317	0,0603	0,0807	0,0169	-0,0329
6_11	0,7971	-0,1757	-0,0502	0,0850	0,1819	0,2109	0,1874
25_1	0,8183	0,3407	-0,3610	0,1086	0,1641	0,1663	0,1037
5_35	0,8568	0,3360	0,1317	0,0165	-0,0242	0,0044	0,1467
5_14	0,8691	0,3360	0,1317	0,1200	0,0793	0,0155	-0,0346
5_30	0,9665	-0,4926	0,3788	0,3762	0,3481	0,0460	-0,0465
5_31	0,9665	-0,4926	0,3788	0,3762	0,3481	0,0460	-0,0465
22_3	1,0989	0,5213	-0,1349	-0,0566	0,0146	0,1518	0,2405
13_1	1,0997	0,6172	0,4666	-0,0298	-0,0653	-0,0634	-0,0390
3_18	1,1894		0,4666	0,3513	0,2347	0,0089	-0,1893
3_19	1,1894		0,4666	0,3513	0,2347	0,0089	-0,1893
5_9	1,2495	0,4710	0,3305	0,2192	0,0936	-0,0384	-0,1124
5_24	1,2573	-0,0146	0,4151	0,4052	0,2704	0,0296	-0,1962
5_25	1,2573	-0,0146	0,4151	0,4052	0,2704	0,0296	-0,1962
10_33	1,2648	-0,2434	0,4420	0,3743	0,2779	0,0874	-0,0829
10_34	1,2648	-0,2434	0,4420	0,3743	0,2779	0,0874	-0,0829
22_7	1,2663	0,5485	-0,1103	-0,0474	0,0376	0,1710	0,2509
3_30	1,3779		0,2974	0,3034	0,2774	0,0760	-0,0132
3_31	1,3779		0,2974	0,3034	0,2774	0,0760	-0,0132
10_21	1,4223	0,1827	0,4666	0,3513	0,2347	0,0089	-0,1893
10_22	1,4223	0,1827	0,4666	0,3513	0,2347	0,0089	-0,1893
5_17	1,4249	-0,7308	0,3411	0,4330	0,3621	0,2527	0,1594
5_2	1,4556		0,4429	0,3230	0,2327	0,0723	-0,0547
3_24	1,4655		0,4545	0,4580	0,3190	0,0481	-0,2230
3_25	1,4655		0,4545	0,4580	0,3190	0,0481	-0,2230
3_32	1,4847		0,0893	0,2293	0,2737	0,2228	0,1320
6_32	1,4847		0,0893	0,2293	0,2737	0,2228	0,1320
16_2	1,4922	0,6172	-0,1473	0,0010	0,0753	0,1914	0,2654
10_3	1,5291		0,4429	0,3406	0,2576	0,0885	-0,0649
3_2	1,5291		0,4429	0,3406	0,2576	0,0885	-0,0649
10_30	1,5700	0,7193	0,6160	-0,0920	-0,1725	0,0900	0,0376
10_31	1,5700	0,7193	0,6160	-0,0920	-0,1725	0,0900	0,0376
3_15	1,8381		0,4420	0,4781	0,3310	0,1147	-0,0918
3_16	1,8381		0,4420	0,4781	0,3310	0,1147	-0,0918
5_4	1,8601		0,4429	0,4377	0,3272	0,1333	-0,0611
3_4	1,8637		0,4429	0,4463	0,3287	0,1262	-0,0605
5_18	1,9511	0,4710	0,4517	0,3383	0,2359	0,0437	-0,1092
5_19	1,9511	0,4710	0,4517	0,3383	0,2359	0,0437	-0,1092

Taula B.3: Suma estandarditzada i correlacions entre els diversos estadístics i les mesures de pèrdua de confidencialitat per intervals de confiança, basats en rangs, obtingudes en aplicar les diverses versions de microagregació multivariant amb k=10 (4/5).

Estad.	Suma rel.	Mic1-12mul	Mic2-11mul	Mic3-10mul	Mic4-9mul	Mic5-8mul	Mic6-7mul
6_20	2,0205		0,4438	0,4851	0,3474	0,1462	-0,0395
3_17	2,0205		0,4438	0,4851	0,3474	0,1462	-0,0395
10_7	2,0224		0,4429	0,3866	0,3235	0,1846	0,0278
3_1	2,1030		0,4429	0,4824	0,3467	0,1635	-0,0056
6_9	2,2129	0,3730	0,1820	0,2635	0,2663	0,2533	0,1535
5_21	2,2204	0,7056	0,5621	0,1759	0,1735	0,0545	-0,0326
5_22	2,2204	0,7056	0,5621	0,1759	0,1735	0,0545	-0,0326
5_1	2,3301		0,4429	0,5536	0,4033	0,2194	-0,0390
4_4	2,3590		0,4429	0,4489	0,4034	0,2509	0,0337
3_5	2,3590		0,4429	0,4489	0,4034	0,2509	0,0337
4_2	2,3590		0,4429	0,4489	0,4034	0,2509	0,0337
4_1	2,3590		0,4429	0,4489	0,4034	0,2509	0,0337
4_6	2,3590		0,4429	0,4489	0,4034	0,2509	0,0337
4_5	2,3590		0,4429	0,4489	0,4034	0,2509	0,0337
5_5	2,3590		0,4429	0,4489	0,4034	0,2509	0,0337
5_6	2,3904		0,4429	0,4271	0,4188	0,2359	0,0707
3_6	2,5673		0,4429	0,4180	0,4435	0,2901	0,1051
5_23	2,6127	0,7056	0,5621	0,2084	0,0915	0,1163	0,1780
5_10	3,0205	0,7056	0,5621	0,4466	0,3171	0,1394	-0,0083
16_7	3,6275	0,7382	0,0926	0,4234	0,3732	0,3746	0,4218
16_3	3,6275	0,7382	0,0926	0,4234	0,3732	0,3746	0,4218
6_13	4,0385	0,7030	0,7371	0,4523	0,4417	0,3416	0,1293
10_13	4,3716	0,7193	0,6160	0,6119	0,5771	0,3920	0,1010
10_12	4,3890	0,6172	0,6276	0,5975	0,5456	0,4059	0,2028
6_12	4,3890	0,6172	0,6276	0,5975	0,5456	0,4059	0,2028
28_1	5,0974	0,6172	0,6268	0,5795	0,5370	0,5182	0,5229
26_1	5,0974	0,6172	0,6268	0,5795	0,5370	0,5182	0,5229
26_5	5,0974	0,6172	0,6268	0,5795	0,5370	0,5182	0,5229
26_6	5,0974	0,6172	0,6268	0,5795	0,5370	0,5182	0,5229
26_7	5,0974	0,6172	0,6268	0,5795	0,5370	0,5182	0,5229
26_2	5,0974	0,6172	0,6268	0,5795	0,5370	0,5182	0,5229
26_4	5,0974	0,6172	0,6268	0,5795	0,5370	0,5182	0,5229
15_4	5,0974	0,6172	0,6268	0,5795	0,5370	0,5182	0,5229
15_1	5,0974	0,6172	0,6268	0,5795	0,5370	0,5182	0,5229
15_2	5,0974	0,6172	0,6268	0,5795	0,5370	0,5182	0,5229
15_5	5,0974	0,6172	0,6268	0,5795	0,5370	0,5182	0,5229
15_6	5,0974	0,6172	0,6268	0,5795	0,5370	0,5182	0,5229
26_3	5,0974	0,6172	0,6268	0,5795	0,5370	0,5182	0,5229

Taula B.3: Suma estandarditzada i correlacions entre els diversos estadístics i les mesures de pèrdua de confidencialitat per intervals de confiança, basats en rangs, obtingudes en aplicar les diverses versions de microagregació multivariant amb k=10 (5/5).

Estad.	Suma rel.	Mic1-12mul	Mic2-11mul	Mic3-10mul	Mic4-9mul	Mic5-8mul	Mic6-7mul
19_1	-5,9154	-0,7955	-0,7069	-0,7215	-0,6535	-0,5834	-0,5439
19_2	-5,9154	-0,7955	-0,7069	-0,7215	-0,6535	-0,5834	-0,5439
19_5	-5,9154	-0,7955	-0,7069	-0,7215	-0,6535	-0,5834	-0,5439
19_6	-5,9154	-0,7955	-0,7069	-0,7215	-0,6535	-0,5834	-0,5439
19_4	-5,9154	-0,7955	-0,7069	-0,7215	-0,6535	-0,5834	-0,5439
18_4	-5,5353	-0,7878	-0,6729	-0,6849	-0,6154	-0,5259	-0,4810
18_1	-5,5353	-0,7878	-0,6729	-0,6849	-0,6154	-0,5259	-0,4810
18_2	-5,5353	-0,7878	-0,6729	-0,6849	-0,6154	-0,5259	-0,4810
18_5	-5,5353	-0,7878	-0,6729	-0,6849	-0,6154	-0,5259	-0,4810
18_6	-5,5353	-0,7878	-0,6729	-0,6849	-0,6154	-0,5259	-0,4810

Taula B.4: Suma estandarditzada i correlacions entre els diversos estadístics i les mesures de pèrdua de confidencialitat per intervals de confiança, basats en desviacions típiques, obtingudes en aplicar les diverses versions de microagregació multivariant amb k=10 (1/5).

Estad.	Suma rel.	Mic1-12mul	Mic2-11mul	Mic3-10mul	Mic4-9mul	Mic5-8mul	Mic6-7mul
5_8	-4,9113	-0,7248	-0,6384	-0,6667	-0,6041	-0,4769	-0,2765
4_13	-4,6592	-0,7878	-0,6598	-0,6542	-0,5634	-0,4071	-0,1919
4_12	-4,6592	-0,7878	-0,6598	-0,6542	-0,5634	-0,4071	-0,1919
5_12	-4,6592	-0,7878	-0,6598	-0,6542	-0,5634	-0,4071	-0,1919
4_9	-4,6592	-0,7878	-0,6598	-0,6542	-0,5634	-0,4071	-0,1919
4_11	-4,6592	-0,7878	-0,6598	-0,6542	-0,5634	-0,4071	-0,1919
3_12	-4,6592	-0,7878	-0,6598	-0,6542	-0,5634	-0,4071	-0,1919
4_8	-4,6592	-0,7878	-0,6598	-0,6542	-0,5634	-0,4071	-0,1919
3_13	-4,2899	-0,8455	-0,7020	-0,5423	-0,4671	-0,3217	-0,1664
29_1	-4,1358	-0,7878	-0,6002	-0,5898	-0,5036	-0,3388	-0,1150
22_1	-3,8006	-0,7109	-0,6027	-0,5526	-0,3563	-0,2690	-0,1945
5_13	-3,4132	-0,5649	-0,5621	-0,4908	-0,4268	-0,2801	-0,0839
22_6	-3,2396	-0,6148	-0,3778	-0,5160	-0,4236	-0,2716	-0,0918
22_5	-3,0640	-0,7878	-0,5445	-0,5277	-0,4301	-0,2144	0,2011
10_14	-2,9157	-0,4814	-0,3842	-0,4204	-0,3554	-0,2719	-0,1226
6_6	-2,5401		-0,4516	-0,4678	-0,4374	-0,2616	-0,0735
22_4	-2,5270	-0,7109	-0,5331	-0,4757	-0,3428	-0,0833	0,2023
10_5	-2,4393		-0,4516	-0,4792	-0,4170	-0,2551	-0,0331
6_5	-2,4393		-0,4516	-0,4792	-0,4170	-0,2551	-0,0331
10_29	-2,3500	-0,7878	-0,5234	-0,4095	-0,2410	0,0006	0,1263
3_8	-2,1107	-0,3824	-0,2235	-0,2491	-0,2769	-0,2230	-0,1105
6_4	-2,0966		-0,4516	-0,4634	-0,3654	-0,1756	0,0243
16_1	-2,0837	-0,7656	-0,2152	-0,3955	-0,2656	-0,0431	0,0708
6_2	-2,0808		-0,4516	-0,4358	-0,3213	-0,1764	-0,0239
3_11	-1,9012	-0,0690	-0,1579	-0,2747	-0,2866	-0,2676	-0,1742
6_18	-1,8597		-0,4504	-0,4317	-0,3044	-0,1342	0,0390
6_19	-1,8597		-0,4504	-0,4317	-0,3044	-0,1342	0,0390
10_4	-1,8569		-0,4516	-0,3987	-0,3162	-0,1414	0,0331
6_1	-1,8569		-0,4516	-0,3987	-0,3162	-0,1414	0,0331
10_1	-1,8569		-0,4516	-0,3987	-0,3162	-0,1414	0,0331
10_17	-1,7914	-0,0624	-0,4645	-0,4239	-0,3193	-0,0920	0,0932
10_26	-1,7914	-0,0624	-0,4645	-0,4239	-0,3193	-0,0920	0,0932
10_23	-1,7914	-0,0624	-0,4645	-0,4239	-0,3193	-0,0920	0,0932
3_9	-1,7307	-0,0624	-0,1422	-0,2539	-0,2577	-0,2416	-0,1618
10_10	-1,7307	-0,0624	-0,1422	-0,2539	-0,2577	-0,2416	-0,1618
6_17	-1,7196		-0,4645	-0,4239	-0,3193	-0,0920	0,0932
3_20	-1,7196		-0,4645	-0,4239	-0,3193	-0,0920	0,0932
14_4	-1,6471	-0,7878	-0,4420	-0,2324	-0,1017	0,0668	0,1348
4_32	-1,6471	-0,7878	-0,4420	-0,2324	-0,1017	0,0668	0,1348
4_29	-1,6471	-0,7878	-0,4420	-0,2324	-0,1017	0,0668	0,1348
5_29	-1,6471	-0,7878	-0,4420	-0,2324	-0,1017	0,0668	0,1348
13_3	-1,6471	-0,7878	-0,4420	-0,2324	-0,1017	0,0668	0,1348
14_1	-1,6471	-0,7878	-0,4420	-0,2324	-0,1017	0,0668	0,1348
14_2	-1,6471	-0,7878	-0,4420	-0,2324	-0,1017	0,0668	0,1348
14_5	-1,6471	-0,7878	-0,4420	-0,2324	-0,1017	0,0668	0,1348
14_6	-1,6471	-0,7878	-0,4420	-0,2324	-0,1017	0,0668	0,1348
4_26	-1,6471	-0,7878	-0,4420	-0,2324	-0,1017	0,0668	0,1348
13_7	-1,6471	-0,7878	-0,4420	-0,2324	-0,1017	0,0668	0,1348
4_20	-1,6471	-0,7878	-0,4420	-0,2324	-0,1017	0,0668	0,1348
4_17	-1,6471	-0,7878	-0,4420	-0,2324	-0,1017	0,0668	0,1348
10_32	-1,6314	-0,8128	-0,5744	-0,4490	-0,1101	0,2241	0,2845
10_35	-1,6057	-0,4814	-0,3671	-0,3077	-0,1818	0,0079	0,0863
6_24	-1,6048		-0,4532	-0,4554	-0,3350	-0,0790	0,1716
6_25	-1,6048		-0,4532	-0,4554	-0,3350	-0,0790	0,1716
10_6	-1,5579			-0,4358	-0,3941	-0,1679	-0,0342

Taula B.4: Suma estandarditzada i correlacions entre els diversos estadístics i les mesures de pèrdua de confidencialitat per intervals de confiança, basats en desviacions típiques, obtingudes en aplicar les diverses versions de microagregació multivariant amb k=10 (2/5).

Estad.	Suma rel.	Mic1-12mul	Mic2-11mul	Mic3-10mul	Mic4-9mul	Mic5-8mul	Mic6-7mul
25_7	-1,5382	-0,5422	-0,3957	-0,3526	-0,2593	-0,0090	0,2971
25_3	-1,5382	-0,5422	-0,3957	-0,3526	-0,2593	-0,0090	0,2971
6_10	-1,5195	-0,0558	-0,1253	-0,2300	-0,2234	-0,2080	-0,1418
3_10	-1,5195	-0,0558	-0,1253	-0,2300	-0,2234	-0,2080	-0,1418
5_26	-1,4224	0,1659	-0,4503	-0,4715	-0,3176	-0,0577	0,1426
6_15	-1,3584		-0,4551	-0,3813	-0,2678	-0,0349	0,1542
6_16	-1,3584		-0,4551	-0,3813	-0,2678	-0,0349	0,1542
10_15	-1,2865	0,0624	-0,4551	-0,3813	-0,2678	-0,0349	0,1542
10_16	-1,2865	0,0624	-0,4551	-0,3813	-0,2678	-0,0349	0,1542
10_24	-1,2865	0,0624	-0,4551	-0,3813	-0,2678	-0,0349	0,1542
10_25	-1,2865	0,0624	-0,4551	-0,3813	-0,2678	-0,0349	0,1542
6_30	-1,0260		-0,3215	-0,3044	-0,2247	-0,0520	0,1543
6_31	-1,0260		-0,3215	-0,3044	-0,2247	-0,0520	0,1543
3_14	-0,8680	-0,1369	-0,0805	-0,1354	-0,1018	-0,0995	-0,0449
6_14	-0,8680	-0,1369	-0,0805	-0,1354	-0,1018	-0,0995	-0,0449
6_23	-0,8470		-0,1253	-0,1558	-0,2500	-0,1181	0,0713
3_23	-0,8470		-0,1253	-0,1558	-0,2500	-0,1181	0,0713
6_27	-0,8431		-0,3990	-0,3918	-0,2822	-0,0099	0,3878
6_28	-0,8431		-0,3990	-0,3918	-0,2822	-0,0099	0,3878
27_1	-0,7465	-0,8690	-0,1631	-0,0124	0,0258	0,1187	0,1406
25_2	-0,7304		-0,4117	-0,2920	-0,1639	0,0657	0,2148
13_2	-0,7223		-0,4067	-0,2920	-0,1632	0,0657	0,2148
22_2	-0,7223		-0,4067	-0,2920	-0,1632	0,0657	0,2148
5_11	-0,6347	0,1659	0,0935	-0,0569	-0,1483	-0,2039	-0,1645
6_35	-0,6283		-0,0805	-0,1685	-0,2430	-0,0638	0,1089
3_35	-0,6283		-0,0805	-0,1685	-0,2430	-0,0638	0,1089
4_30	-0,4237	-0,7878	0,3385	0,3422	0,2331	-0,0439	-0,4089
4_31	-0,4237	-0,7878	0,3385	0,3422	0,2331	-0,0439	-0,4089
4_27	-0,4237	-0,7878	0,3385	0,3422	0,2331	-0,0439	-0,4089
4_28	-0,4237	-0,7878	0,3385	0,3422	0,2331	-0,0439	-0,4089
5_27	-0,4237	-0,7878	0,3385	0,3422	0,2331	-0,0439	-0,4089
5_28	-0,4237	-0,7878	0,3385	0,3422	0,2331	-0,0439	-0,4089
4_24	-0,4237	-0,7878	0,3385	0,3422	0,2331	-0,0439	-0,4089
4_25	-0,4237	-0,7878	0,3385	0,3422	0,2331	-0,0439	-0,4089
4_18	-0,4237	-0,7878	0,3385	0,3422	0,2331	-0,0439	-0,4089
4_19	-0,4237	-0,7878	0,3385	0,3422	0,2331	-0,0439	-0,4089
4_15	-0,4237	-0,7878	0,3385	0,3422	0,2331	-0,0439	-0,4089
4_16	-0,4237	-0,7878	0,3385	0,3422	0,2331	-0,0439	-0,4089
5_20	-0,2781	0,6514	-0,3761	-0,3389	-0,2198	-0,0092	0,1774
6_21	-0,1492		-0,1253	0,0764	0,1213	-0,0146	-0,1297
6_22	-0,1492		-0,1253	0,0764	0,1213	-0,0146	-0,1297
3_21	-0,1492		-0,1253	0,0764	0,1213	-0,0146	-0,1297
3_22	-0,1492		-0,1253	0,0764	0,1213	-0,0146	-0,1297
6_26	-0,1035		0,0140	-0,0733	-0,1438	-0,0047	0,1122
3_26	-0,1035		0,0140	-0,0733	-0,1438	-0,0047	0,1122
5_15	-0,0371	-0,7248	0,3836	0,4492	0,2171	-0,0505	-0,3339
5_16	-0,0371	-0,7248	0,3836	0,4492	0,2171	-0,0505	-0,3339
25_4	0,0623	0,5422	-0,4207	-0,2290	-0,0870	0,0962	0,1735
25_5	0,0623	0,5422	-0,4207	-0,2290	-0,0870	0,0962	0,1735
25_6	0,0623	0,5422	-0,4207	-0,2290	-0,0870	0,0962	0,1735
5_35	0,1540	0,2042	0,1443	-0,1165	-0,1067	-0,0602	0,0776
6_33	0,3056		-0,0805	0,0984	0,1353	0,0480	-0,0034
6_34	0,3056		-0,0805	0,0984	0,1353	0,0480	-0,0034
3_33	0,3056		-0,0805	0,0984	0,1353	0,0480	-0,0034
3_34	0,3056		-0,0805	0,0984	0,1353	0,0480	-0,0034

Taula B.4: Suma estandarditzada i correlacions entre els diversos estadístics i les mesures de pèrdua de confidencialitat per intervals de confiança, basats en desviacions típiques, obtingudes en aplicar les diverses versions de microagregació multivariant amb k=10 (3/5).

Estad.	Suma rel.	Mic1-12mul	Mic2-11mul	Mic3-10mul	Mic4-9mul	Mic5-8mul	Mic6-7mul
5_14	0,3795	0,2042	0,1443	0,0395	0,0170	-0,0375	-0,0414
13_5	0,4237	0,7878	-0,3385	-0,3422	-0,2331	0,0439	0,4089
13_6	0,4237	0,7878	-0,3385	-0,3422	-0,2331	0,0439	0,4089
13_4	0,4237	0,7878	-0,3385	-0,3422	-0,2331	0,0439	0,4089
5_3	0,4244			0,1177	0,1320	0,0385	-0,0037
5_7	0,4499			0,1286	0,1288	0,0488	-0,0049
16_5	0,4959	0,4872	-0,1547	-0,1676	-0,0690	0,0999	0,1744
16_6	0,4959	0,4872	-0,1547	-0,1676	-0,0690	0,0999	0,1744
16_4	0,4959	0,4872	-0,1547	-0,1676	-0,0690	0,0999	0,1744
3_3	0,5135			0,1177	0,1854	0,0826	-0,0407
6_3	0,5135			0,1177	0,1854	0,0826	-0,0407
3_7	0,6815			0,1286	0,1743	0,1044	0,0314
6_7	0,6815			0,1286	0,1743	0,1044	0,0314
10_27	0,7776	0,7878	-0,2710	-0,2898	-0,1837	0,0922	0,4238
10_28	0,7776	0,7878	-0,2710	-0,2898	-0,1837	0,0922	0,4238
5_33	0,7808	0,2042	0,1443	0,1260	0,1156	0,0218	-0,0257
5_34	0,7808	0,2042	0,1443	0,1260	0,1156	0,0218	-0,0257
3_27	0,8431		0,3990	0,3918	0,2822	0,0099	-0,3878
3_28	0,8431		0,3990	0,3918	0,2822	0,0099	-0,3878
3_30	1,0260		0,3215	0,3044	0,2247	0,0520	-0,1543
3_31	1,0260		0,3215	0,3044	0,2247	0,0520	-0,1543
5_32	1,0468	-0,5649	0,1123	0,1623	0,2775	0,2558	0,2447
5_30	1,0685	-0,5649	0,3867	0,3573	0,3124	0,1073	0,0078
5_31	1,0685	-0,5649	0,3867	0,3573	0,3124	0,1073	0,0078
3_29	1,1185		0,1885	0,1061	0,1070	0,1590	0,1460
6_29	1,1185		0,1885	0,1061	0,1070	0,1590	0,1460
10_33	1,1687	-0,4814	0,4359	0,4028	0,3090	0,1048	-0,0568
10_34	1,1687	-0,4814	0,4359	0,4028	0,3090	0,1048	-0,0568
25_1	1,1870	0,5422	-0,1487	0,1283	0,1263	0,1371	0,0909
13_1	1,2829	0,7878	0,4204	-0,0207	-0,0473	-0,0584	-0,0093
10_21	1,2865	-0,0624	0,4551	0,3813	0,2678	0,0349	-0,1542
10_22	1,2865	-0,0624	0,4551	0,3813	0,2678	0,0349	-0,1542
22_3	1,2937	0,7109	-0,0717	-0,0355	-0,0033	0,1409	0,2120
3_18	1,3584		0,4551	0,3813	0,2678	0,0349	-0,1542
3_19	1,3584		0,4551	0,3813	0,2678	0,0349	-0,1542
22_7	1,4337	0,7315	-0,0501	-0,0330	0,0202	0,1595	0,2198
3_24	1,6048		0,4532	0,4554	0,3350	0,0790	-0,1716
3_25	1,6048		0,4532	0,4554	0,3350	0,0790	-0,1716
16_2	1,6271	0,7878	-0,0950	0,0145	0,0645	0,1828	0,2300
5_9	1,7086	0,6514	0,4692	0,2596	0,1031	-0,0409	-0,0828
10_30	1,7240	0,8128	0,5744	-0,0699	-0,1532	0,0899	0,0834
10_31	1,7240	0,8128	0,5744	-0,0699	-0,1532	0,0899	0,0834
6_8	1,7307	0,0624	0,1422	0,2539	0,2577	0,2416	0,1618
10_8	1,7307	0,0624	0,1422	0,2539	0,2577	0,2416	0,1618
10_11	1,7307	0,0624	0,1422	0,2539	0,2577	0,2416	0,1618
5_24	1,7919	0,1659	0,4463	0,4548	0,3186	0,0655	-0,1416
5_25	1,7919	0,1659	0,4463	0,4548	0,3186	0,0655	-0,1416
5_2	1,7924		0,4516	0,3928	0,2905	0,1149	-0,0175
3_32	1,8447		0,3888	0,3132	0,2109	0,2447	0,0645
6_32	1,8447		0,3888	0,3132	0,2109	0,2447	0,0645
10_3	1,8569		0,4516	0,3987	0,3162	0,1414	-0,0331
3_2	1,8569		0,4516	0,3987	0,3162	0,1414	-0,0331
3_15	1,8597		0,4504	0,4317	0,3044	0,1342	-0,0390
3_16	1,8597		0,4504	0,4317	0,3044	0,1342	-0,0390
6_11	1,9012	0,0690	0,1579	0,2747	0,2866	0,2676	0,1742

Taula B.4: Suma estandarditzada i correlacions entre els diversos estadístics i les mesures de pèrdua de confidencialitat per intervals de confiança, basats en desviacions típiques, obtingudes en aplicar les diverses versions de microagregació multivariant amb k=10 (4/5).

Estad.	Suma rel.	Mic1-12mul	Mic2-11mul	Mic3-10mul	Mic4-9mul	Mic5-8mul	Mic6-7mul
5_17	1,9628	-0,7248	0,5296	0,4654	0,3831	0,2858	0,1775
3_17	2,0213		0,4528	0,4388	0,3206	0,1609	0,0033
6_20	2,0213		0,4528	0,4388	0,3206	0,1609	0,0033
3_1	2,0808		0,4516	0,4358	0,3213	0,1764	0,0239
5_4	2,0962		0,4516	0,4790	0,3683	0,1643	-0,0281
3_4	2,0966		0,4516	0,4634	0,3654	0,1756	-0,0243
6_9	2,1107	0,3824	0,2235	0,2491	0,2769	0,2230	0,1105
5_1	2,2546		0,4516	0,5372	0,3892	0,2033	-0,0397
10_7	2,3137		0,4516	0,4434	0,3786	0,2274	0,0496
5_6	2,4299		0,4516	0,4300	0,4019	0,2542	0,0786
4_4	2,4393		0,4516	0,4792	0,4170	0,2551	0,0331
3_5	2,4393		0,4516	0,4792	0,4170	0,2551	0,0331
4_2	2,4393		0,4516	0,4792	0,4170	0,2551	0,0331
4_5	2,4393		0,4516	0,4792	0,4170	0,2551	0,0331
5_5	2,4393		0,4516	0,4792	0,4170	0,2551	0,0331
4_6	2,4393		0,4516	0,4792	0,4170	0,2551	0,0331
4_1	2,4393		0,4516	0,4792	0,4170	0,2551	0,0331
5_18	2,4722	0,6514	0,4901	0,4097	0,2900	0,0784	-0,0634
5_19	2,4722	0,6514	0,4901	0,4097	0,2900	0,0784	-0,0634
5_23	2,5075	0,8357	0,6345	0,1055	0,0181	0,0712	0,1917
3_6	2,5401		0,4516	0,4678	0,4374	0,2616	0,0735
5_21	2,7462	0,8357	0,6345	0,2746	0,2293	0,0861	0,0044
5_22	2,7462	0,8357	0,6345	0,2746	0,2293	0,0861	0,0044
5_10	3,2196	0,8357	0,6345	0,4701	0,3103	0,1236	0,0121
16_7	3,7529	0,8691	0,1168	0,4610	0,4070	0,3828	0,3643
16_3	3,7529	0,8691	0,1168	0,4610	0,4070	0,3828	0,3643
10_13	4,1252	0,8128	0,5744	0,5676	0,5137	0,3522	0,1092
6_13	4,2899	0,8455	0,7020	0,5423	0,4671	0,3217	0,1664
6_12	4,6592	0,7878	0,6598	0,6542	0,5634	0,4071	0,1919
10_12	4,6592	0,7878	0,6598	0,6542	0,5634	0,4071	0,1919
26_1	5,3952	0,7878	0,6743	0,6572	0,5537	0,5098	0,4910
26_7	5,3952	0,7878	0,6743	0,6572	0,5537	0,5098	0,4910
26_5	5,3952	0,7878	0,6743	0,6572	0,5537	0,5098	0,4910
26_6	5,3952	0,7878	0,6743	0,6572	0,5537	0,5098	0,4910
26_2	5,3952	0,7878	0,6743	0,6572	0,5537	0,5098	0,4910
26_4	5,3952	0,7878	0,6743	0,6572	0,5537	0,5098	0,4910
15_4	5,3952	0,7878	0,6743	0,6572	0,5537	0,5098	0,4910
15_1	5,3952	0,7878	0,6743	0,6572	0,5537	0,5098	0,4910
15_2	5,3952	0,7878	0,6743	0,6572	0,5537	0,5098	0,4910
15_5	5,3952	0,7878	0,6743	0,6572	0,5537	0,5098	0,4910
15_6	5,3952	0,7878	0,6743	0,6572	0,5537	0,5098	0,4910
26_3	5,3952	0,7878	0,6743	0,6572	0,5537	0,5098	0,4910
28_1	5,3952	0,7878	0,6743	0,6572	0,5537	0,5098	0,4910

Taula B.4: Suma estandarditzada i correlacions entre els diversos estadístics i les mesures de pèrdua de confidencialitat per intervals de confiança, basats en desviacions típiques, obtingudes en aplicar les diverses versions de microagregació multivariant amb k=10 (5/5).

Estad.	Suma rel.	Mic1-12mul	Mic2-11mul	Mic3-10mul	Mic4-9mul	Mic5-8mul	Mic6-7mul
5_11	-4,8378	-0,3354	-0,3514	-0,3348	-0,2907	-0,1871	-0,1583
5_14	-3,9401	-0,2582	-0,1013	-0,2200	-0,2546	-0,2044	-0,2380
22_1	-3,6445	-0,1928	-0,1443	-0,0957	-0,1875	-0,2289	-0,3588
5_9	-3,4442	-0,1876	-0,0956	-0,2046	-0,2292	-0,1892	-0,2078
5_33	-3,3398	-0,2582	-0,1013	-0,1701	-0,1321	-0,1642	-0,2863
5_34	-3,3398	-0,2582	-0,1013	-0,1701	-0,1321	-0,1642	-0,2863
22_6	-3,1768	-0,1280	-0,3468	-0,1912	-0,1256	-0,0889	-0,2901
5_8	-2,5918	-0,2336	-0,4050	-0,2584	-0,1570	-0,0213	0,1052
18_1	-2,4600	-0,2055	-0,2969	-0,1007	-0,1057	-0,0532	-0,1505
18_2	-2,4600	-0,2055	-0,2969	-0,1007	-0,1057	-0,0532	-0,1505
18_5	-2,4600	-0,2055	-0,2969	-0,1007	-0,1057	-0,0532	-0,1505
18_6	-2,4600	-0,2055	-0,2969	-0,1007	-0,1057	-0,0532	-0,1505
18_4	-2,4600	-0,2055	-0,2969	-0,1007	-0,1057	-0,0532	-0,1505
5_24	-2,4421	-0,3354	0,2236	-0,0280	-0,0968	-0,1843	-0,2731
5_25	-2,4421	-0,3354	0,2236	-0,0280	-0,0968	-0,1843	-0,2731
3_3	-2,3998			-0,1424	-0,1299	-0,2045	-0,2699
6_3	-2,3998			-0,1424	-0,1299	-0,2045	-0,2699
19_1	-2,3971	-0,2041	-0,2935	-0,1376	-0,1240	-0,0598	-0,0562
19_2	-2,3971	-0,2041	-0,2935	-0,1376	-0,1240	-0,0598	-0,0562
19_5	-2,3971	-0,2041	-0,2935	-0,1376	-0,1240	-0,0598	-0,0562
19_6	-2,3971	-0,2041	-0,2935	-0,1376	-0,1240	-0,0598	-0,0562
19_4	-2,3971	-0,2041	-0,2935	-0,1376	-0,1240	-0,0598	-0,0562
6_30	-2,3043		-0,2062	-0,1274	-0,2699	-0,0974	-0,0801
6_31	-2,3043		-0,2062	-0,1274	-0,2699	-0,0974	-0,0801
3_33	-2,2822		0,0639	-0,1186	-0,1388	-0,1893	-0,3097
3_34	-2,2822		0,0639	-0,1186	-0,1388	-0,1893	-0,3097
6_33	-2,2822		0,0639	-0,1186	-0,1388	-0,1893	-0,3097
6_34	-2,2822		0,0639	-0,1186	-0,1388	-0,1893	-0,3097
5_18	-2,2501	-0,1876	0,2142	-0,0738	-0,1090	-0,1828	-0,2947
5_19	-2,2501	-0,1876	0,2142	-0,0738	-0,1090	-0,1828	-0,2947
5_21	-2,1588	-0,0238	0,1086	-0,1416	-0,1130	-0,1708	-0,3039
5_22	-2,1588	-0,0238	0,1086	-0,1416	-0,1130	-0,1708	-0,3039
6_7	-2,0339			-0,1149	-0,1136	-0,1597	-0,2509
3_7	-2,0339			-0,1149	-0,1136	-0,1597	-0,2509
6_21	-2,0316		0,2023	-0,1197	-0,1116	-0,2053	-0,3257
6_22	-2,0316		0,2023	-0,1197	-0,1116	-0,2053	-0,3257
3_21	-2,0316		0,2023	-0,1197	-0,1116	-0,2053	-0,3257
3_22	-2,0316		0,2023	-0,1197	-0,1116	-0,2053	-0,3257
5_10	-1,8488	-0,0238	0,1086	-0,0674	-0,1420	-0,1686	-0,2366
5_3	-1,8203			-0,1424	-0,0702	-0,1225	-0,2511
13_1	-1,7235	0,2055	0,2566	-0,1484	-0,1682	-0,2052	-0,3742
4_11	-1,6683	-0,2055	-0,2998	-0,0867	-0,0795	0,0075	0,0091
4_8	-1,6683	-0,2055	-0,2998	-0,0867	-0,0795	0,0075	0,0091
3_12	-1,6683	-0,2055	-0,2998	-0,0867	-0,0795	0,0075	0,0091
4_13	-1,6683	-0,2055	-0,2998	-0,0867	-0,0795	0,0075	0,0091
4_12	-1,6683	-0,2055	-0,2998	-0,0867	-0,0795	0,0075	0,0091
5_12	-1,6683	-0,2055	-0,2998	-0,0867	-0,0795	0,0075	0,0091
4_9	-1,6683	-0,2055	-0,2998	-0,0867	-0,0795	0,0075	0,0091
6_11	-1,5833	-0,2967	-0,1681	-0,1241	-0,0629	-0,0078	0,0911
6_18	-1,5674		-0,2751	-0,1127	-0,0553	-0,0531	-0,1129
6_19	-1,5674		-0,2751	-0,1127	-0,0553	-0,0531	-0,1129
5_7	-1,5617			-0,1149	-0,0553	-0,1009	-0,2363
6_8	-1,5286	-0,2993	-0,1856	-0,1301	-0,0647	0,0051	0,1170
10_8	-1,5286	-0,2993	-0,1856	-0,1301	-0,0647	0,0051	0,1170
10_11	-1,5286	-0,2993	-0,1856	-0,1301	-0,0647	0,0051	0,1170

Taula B.5: Suma estandarditzada i correlacions entre els diversos estadístics i les mesures generals obtingudes en aplicar les diverses versions de microagregació multivariant amb k=10 (1/4).

Estad.	Suma rel.	Mic1-12mul	Mic2-11mul	Mic3-10mul	Mic4-9mul	Mic5-8mul	Mic6-7mul
10_14	-1,4872	-0,0837	-0,1157	-0,0637	-0,1008	-0,0554	-0,0972
10_33	-1,4534	-0,0837	0,2839	-0,0136	-0,0704	-0,1729	-0,2944
10_34	-1,4534	-0,0837	0,2839	-0,0136	-0,0704	-0,1729	-0,2944
3_18	-1,4033		0,3054	-0,0196	-0,0802	-0,1969	-0,3297
3_19	-1,4033		0,3054	-0,0196	-0,0802	-0,1969	-0,3297
6_2	-1,3512		-0,2763	-0,1132	-0,0548	-0,0351	-0,0577
29_1	-1,3366	-0,2055	-0,3009	-0,0632	-0,0560	0,0324	0,0408
10_3	-1,2702		0,2763	-0,0385	-0,0914	-0,1812	-0,2463
3_2	-1,2702		0,2763	-0,0385	-0,0914	-0,1812	-0,2463
4_18	-1,2425	-0,2055	0,2470	-0,0048	-0,0246	-0,1237	-0,1942
4_19	-1,2425	-0,2055	0,2470	-0,0048	-0,0246	-0,1237	-0,1942
4_30	-1,2425	-0,2055	0,2470	-0,0048	-0,0246	-0,1237	-0,1942
4_31	-1,2425	-0,2055	0,2470	-0,0048	-0,0246	-0,1237	-0,1942
4_27	-1,2425	-0,2055	0,2470	-0,0048	-0,0246	-0,1237	-0,1942
4_28	-1,2425	-0,2055	0,2470	-0,0048	-0,0246	-0,1237	-0,1942
5_27	-1,2425	-0,2055	0,2470	-0,0048	-0,0246	-0,1237	-0,1942
5_28	-1,2425	-0,2055	0,2470	-0,0048	-0,0246	-0,1237	-0,1942
4_15	-1,2425	-0,2055	0,2470	-0,0048	-0,0246	-0,1237	-0,1942
4_16	-1,2425	-0,2055	0,2470	-0,0048	-0,0246	-0,1237	-0,1942
4_24	-1,2425	-0,2055	0,2470	-0,0048	-0,0246	-0,1237	-0,1942
4_25	-1,2425	-0,2055	0,2470	-0,0048	-0,0246	-0,1237	-0,1942
6_14	-1,0999	0,0320	0,0639	-0,0275	-0,0959	-0,1040	-0,1859
3_14	-1,0999	0,0320	0,0639	-0,0275	-0,0959	-0,1040	-0,1859
3_24	-0,9405		0,2903	0,0411	-0,0422	-0,1640	-0,3109
3_25	-0,9405		0,2903	0,0411	-0,0422	-0,1640	-0,3109
10_6	-0,9374			-0,1132	-0,1135	-0,0178	-0,0520
6_6	-0,8938		-0,2763	-0,0337	-0,0797	-0,0008	0,0061
5_26	-0,8541	-0,3354	-0,3281	-0,0924	-0,0416	0,0974	0,3045
22_5	-0,7468	-0,2055	-0,2944	-0,0438	-0,0311	0,0692	0,1490
5_2	-0,7411		0,2763	-0,0275	-0,0389	-0,1157	-0,2251
3_4	-0,6997		0,2763	0,0264	-0,0517	-0,1447	-0,2103
3_13	-0,5756	-0,2219	-0,4036	0,0691	0,0626	0,0784	0,0491
5_35	-0,5466	-0,2582	-0,1013	0,0381	-0,0382	-0,0028	0,1677
5_20	-0,5292	-0,1876	-0,3451	-0,0315	-0,0353	0,0570	0,2567
10_21	-0,5109	0,2993	0,3054	-0,0196	-0,0802	-0,1969	-0,3297
10_22	-0,5109	0,2993	0,3054	-0,0196	-0,0802	-0,1969	-0,3297
3_27	-0,4743		0,2635	0,0072	-0,0118	-0,1144	-0,1914
3_28	-0,4743		0,2635	0,0072	-0,0118	-0,1144	-0,1914
10_7	-0,4655		0,2763	0,0016	-0,0254	-0,1128	-0,1759
10_5	-0,2924		-0,2763	-0,0300	-0,0270	0,0581	0,0722
6_5	-0,2924		-0,2763	-0,0300	-0,0270	0,0581	0,0722
5_4	-0,2715		0,2763	0,0324	-0,0115	-0,0972	-0,1788
3_8	-0,2306	-0,1050	-0,2028	-0,0385	-0,0018	0,0818	0,1014
10_30	-0,0665	0,2618	0,4075	-0,0651	-0,1323	-0,0986	-0,2222
10_31	-0,0665	0,2618	0,4075	-0,0651	-0,1323	-0,0986	-0,2222
10_32	0,1653	-0,2618	-0,4075	-0,0358	0,0938	0,1733	0,3112
6_9	0,2306	0,1050	0,2028	0,0385	0,0018	-0,0818	-0,1014
4_2	0,2924		0,2763	0,0300	0,0270	-0,0581	-0,0722
4_6	0,2924		0,2763	0,0300	0,0270	-0,0581	-0,0722
3_5	0,2924		0,2763	0,0300	0,0270	-0,0581	-0,0722
4_5	0,2924		0,2763	0,0300	0,0270	-0,0581	-0,0722
5_5	0,2924		0,2763	0,0300	0,0270	-0,0581	-0,0722
4_4	0,2924		0,2763	0,0300	0,0270	-0,0581	-0,0722
4_1	0,2924		0,2763	0,0300	0,0270	-0,0581	-0,0722
25_3	0,3748	-0,1557	-0,2261	0,0249	0,0167	0,1094	0,2664

Taula B.5: Suma estandarditzada i correlacions entre els diversos estadístics i les mesures generals obtingudes en aplicar les diverses versions de microagregació multivariant amb k=10 (2/4).

Estad.	Suma rel.	Mic1-12mul	Mic2-11mul	Mic3-10mul	Mic4-9mul	Mic5-8mul	Mic6-7mul
25_7	0,3748	-0,1557	-0,2261	0,0249	0,0167	0,1094	0,2664
22_4	0,3893	-0,1928	-0,2593	0,0122	0,0230	0,1353	0,3050
6_27	0,4743		-0,2635	-0,0072	0,0118	0,1144	0,1914
6_28	0,4743		-0,2635	-0,0072	0,0118	0,1144	0,1914
5_15	0,4993	-0,2336	0,2322	0,0839	0,0004	-0,0087	0,1993
5_16	0,4993	-0,2336	0,2322	0,0839	0,0004	-0,0087	0,1993
10_15	0,5109	-0,2993	-0,3054	0,0196	0,0802	0,1969	0,3297
10_16	0,5109	-0,2993	-0,3054	0,0196	0,0802	0,1969	0,3297
10_24	0,5109	-0,2993	-0,3054	0,0196	0,0802	0,1969	0,3297
10_25	0,5109	-0,2993	-0,3054	0,0196	0,0802	0,1969	0,3297
6_13	0,5756	0,2219	0,4036	-0,0691	-0,0626	-0,0784	-0,0491
5_6	0,6667		0,2763	0,0217	0,0623	-0,0418	0,0079
6_4	0,6997		-0,2763	-0,0264	0,0517	0,1447	0,2103
3_6	0,8938		0,2763	0,0337	0,0797	0,0008	-0,0061
5_23	0,9028	-0,0238	0,1086	0,1152	0,0040	0,0217	0,1205
6_24	0,9405		-0,2903	-0,0411	0,0422	0,1640	0,3109
6_25	0,9405		-0,2903	-0,0411	0,0422	0,1640	0,3109
5_13	0,9528	-0,0428	-0,1251	0,0712	0,0759	0,1191	0,1413
10_29	1,0878	-0,2055	-0,2912	0,0306	0,0889	0,2186	0,3752
13_4	1,2425	0,2055	-0,2470	0,0048	0,0246	0,1237	0,1942
13_5	1,2425	0,2055	-0,2470	0,0048	0,0246	0,1237	0,1942
13_6	1,2425	0,2055	-0,2470	0,0048	0,0246	0,1237	0,1942
6_1	1,2702		-0,2763	0,0385	0,0914	0,1812	0,2463
10_1	1,2702		-0,2763	0,0385	0,0914	0,1812	0,2463
10_4	1,2702		-0,2763	0,0385	0,0914	0,1812	0,2463
5_32	1,3065	-0,0428	-0,1942	0,0393	0,1608	0,1104	0,2630
5_1	1,3203		0,2763	0,1306	0,0557	0,0041	0,0715
10_13	1,3387	0,2618	0,4075	0,0947	0,0487	-0,0832	-0,1305
3_1	1,3512		0,2763	0,1132	0,0548	0,0351	0,0577
10_27	1,3928	0,2055	-0,2270	0,0169	0,0370	0,1310	0,1957
10_28	1,3928	0,2055	-0,2270	0,0169	0,0370	0,1310	0,1957
6_15	1,4033		-0,3054	0,0196	0,0802	0,1969	0,3297
6_16	1,4033		-0,3054	0,0196	0,0802	0,1969	0,3297
4_20	1,4379	-0,2055	-0,3002	0,0915	0,1396	0,2215	0,3751
13_7	1,4379	-0,2055	-0,3002	0,0915	0,1396	0,2215	0,3751
4_29	1,4379	-0,2055	-0,3002	0,0915	0,1396	0,2215	0,3751
5_29	1,4379	-0,2055	-0,3002	0,0915	0,1396	0,2215	0,3751
4_32	1,4379	-0,2055	-0,3002	0,0915	0,1396	0,2215	0,3751
14_4	1,4379	-0,2055	-0,3002	0,0915	0,1396	0,2215	0,3751
4_17	1,4379	-0,2055	-0,3002	0,0915	0,1396	0,2215	0,3751
13_3	1,4379	-0,2055	-0,3002	0,0915	0,1396	0,2215	0,3751
14_1	1,4379	-0,2055	-0,3002	0,0915	0,1396	0,2215	0,3751
14_2	1,4379	-0,2055	-0,3002	0,0915	0,1396	0,2215	0,3751
14_5	1,4379	-0,2055	-0,3002	0,0915	0,1396	0,2215	0,3751
14_6	1,4379	-0,2055	-0,3002	0,0915	0,1396	0,2215	0,3751
4_26	1,4379	-0,2055	-0,3002	0,0915	0,1396	0,2215	0,3751
6_10	1,4700	0,3019	0,2023	0,1349	0,0655	-0,0186	-0,1403
3_10	1,4700	0,3019	0,2023	0,1349	0,0655	-0,0186	-0,1403
3_9	1,5286	0,2993	0,1856	0,1301	0,0647	-0,0051	-0,1170
10_10	1,5286	0,2993	0,1856	0,1301	0,0647	-0,0051	-0,1170
3_15	1,5674		0,2751	0,1127	0,0553	0,0531	0,1129
3_16	1,5674		0,2751	0,1127	0,0553	0,0531	0,1129
3_11	1,5833	0,2967	0,1681	0,1241	0,0629	0,0078	-0,0911
16_1	1,6356	-0,0967	-0,3030	0,0236	0,1439	0,2430	0,3681
3_20	1,6363		-0,2717	0,0334	0,0935	0,2058	0,3470

Taula B.5: Suma estandarditzada i correlacions entre els diversos estadístics i les mesures generals obtingudes en aplicar les diverses versions de microagregació multivariant amb k=10 (3/4).

Estad.	Suma rel.	Mic1-12mul	Mic2-11mul	Mic3-10mul	Mic4-9mul	Mic5-8mul	Mic6-7mul
6_17	1,6363		-0,2717	0,0334	0,0935	0,2058	0,3470
6_20	1,6607		0,2775	0,1122	0,0591	0,0601	0,1315
3_17	1,6607		0,2775	0,1122	0,0591	0,0601	0,1315
10_12	1,6683	0,2055	0,2998	0,0867	0,0795	-0,0075	-0,0091
6_12	1,6683	0,2055	0,2998	0,0867	0,0795	-0,0075	-0,0091
13_2	1,7736		-0,2838	0,0581	0,1054	0,2104	0,3572
22_2	1,7736		-0,2838	0,0581	0,1054	0,2104	0,3572
25_2	1,7834		-0,2915	0,0624	0,1089	0,2104	0,3572
5_30	1,9643	-0,0428	0,3012	0,0691	0,1422	0,0633	0,2127
5_31	1,9643	-0,0428	0,3012	0,0691	0,1422	0,0633	0,2127
3_35	2,0265		0,0639	0,1211	0,0809	0,1332	0,2765
6_35	2,0265		0,0639	0,1211	0,0809	0,1332	0,2765
3_26	2,0446		-0,1864	0,1092	0,1039	0,2075	0,3376
6_26	2,0446		-0,1864	0,1092	0,1039	0,2075	0,3376
10_35	2,0731	-0,0837	-0,2399	0,1527	0,1664	0,2121	0,3443
5_17	2,0902	-0,2336	0,1989	0,1803	0,1125	0,1463	0,3376
27_1	2,0977	-0,0870	-0,2344	0,1330	0,1624	0,2174	0,3726
3_30	2,3043		0,2062	0,1274	0,2699	0,0974	0,0801
3_31	2,3043		0,2062	0,1274	0,2699	0,0974	0,0801
10_26	2,5286	0,2993	-0,2717	0,0334	0,0935	0,2058	0,3470
10_17	2,5286	0,2993	-0,2717	0,0334	0,0935	0,2058	0,3470
10_23	2,5286	0,2993	-0,2717	0,0334	0,0935	0,2058	0,3470
25_4	2,5363	0,1557	-0,3502	0,0899	0,1587	0,2398	0,3691
25_5	2,5363	0,1557	-0,3502	0,0899	0,1587	0,2398	0,3691
25_6	2,5363	0,1557	-0,3502	0,0899	0,1587	0,2398	0,3691
25_1	2,6871	0,1557	-0,5195	0,1547	0,2291	0,2530	0,3649
22_3	2,8099	0,1928	-0,2462	0,0972	0,1537	0,2323	0,3647
16_5	2,8166	0,1897	-0,2584	0,0939	0,1614	0,2357	0,3682
16_6	2,8166	0,1897	-0,2584	0,0939	0,1614	0,2357	0,3682
16_4	2,8166	0,1897	-0,2584	0,0939	0,1614	0,2357	0,3682
16_7	2,8328	0,1884	0,0503	0,1808	0,1154	0,1307	0,2704
16_3	2,8328	0,1884	0,0503	0,1808	0,1154	0,1307	0,2704
3_29	2,8851		-0,1168	0,1652	0,2156	0,2232	0,3725
6_29	2,8851		-0,1168	0,1652	0,2156	0,2232	0,3725
6_23	2,9043		0,2023	0,1558	0,1264	0,1877	0,3275
3_23	2,9043		0,2023	0,1558	0,1264	0,1877	0,3275
22_7	2,9850	0,2081	-0,2142	0,1157	0,1628	0,2299	0,3613
16_2	3,0048	0,2055	-0,2018	0,1348	0,1618	0,2202	0,3568
28_1	3,0817	0,2055	0,2282	0,1488	0,1545	0,1391	0,1891
26_1	3,0817	0,2055	0,2282	0,1488	0,1545	0,1391	0,1891
26_5	3,0817	0,2055	0,2282	0,1488	0,1545	0,1391	0,1891
26_6	3,0817	0,2055	0,2282	0,1488	0,1545	0,1391	0,1891
26_2	3,0817	0,2055	0,2282	0,1488	0,1545	0,1391	0,1891
15_4	3,0817	0,2055	0,2282	0,1488	0,1545	0,1391	0,1891
26_3	3,0817	0,2055	0,2282	0,1488	0,1545	0,1391	0,1891
15_1	3,0817	0,2055	0,2282	0,1488	0,1545	0,1391	0,1891
15_2	3,0817	0,2055	0,2282	0,1488	0,1545	0,1391	0,1891
15_5	3,0817	0,2055	0,2282	0,1488	0,1545	0,1391	0,1891
15_6	3,0817	0,2055	0,2282	0,1488	0,1545	0,1391	0,1891
26_4	3,0817	0,2055	0,2282	0,1488	0,1545	0,1391	0,1891
26_7	3,0817	0,2055	0,2282	0,1488	0,1545	0,1391	0,1891
6_32	3,1373		-0,0628	0,1792	0,2777	0,1953	0,3737
3_32	3,1373		-0,0628	0,1792	0,2777	0,1953	0,3737

Taula B.5: Suma estandarditzada i correlacions entre els diversos estadístics i les mesures generals obtingudes en aplicar les diverses versions de microagregació multivariant amb k=10 (4/4).

Bibliografia

Contribucions

Josep Domingo-Ferrer i Àngel Torres-Aragó, "An additive and multiplicative privacy homomorphism allowing inverse computation", *IV Catalan Days of Applied Mathematics*, Tarragona, Feb. 1998.

Josep Domingo-Ferrer, Josep M. Mateo-Sanz, Anna Oganian i Àngel Torres (2002), "On the security of microaggregation with individual ranking: analytical attacks", *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 477-492. ISSN 0218-4885.

Josep Domingo-Ferrer, Josep M. Mateo-Sanz i Àngel Torres, "Concepts for the evaluation of anonymized data", *Anonymisierung wirtschaftsstatistischer Einzeldaten*, Wiesbaden: Statistisches Bundesamt, 2003 (en premsa).

Josep M. Mateo-Sanz, Josep Domingo-Ferrer, Francesc Sebé, Antoni Martínez, Àngel Torres i Narcís Macià (2002), *Microaggregation algorithms (software, documentation and related papers)*, Deliverables 1-1.D5 i 1-1.D6 del projecte europeu CASC (IST-2000-25069).

Josep M. Mateo-Sanz, Ángel Torres-Aragó i Josep Domingo-Ferrer (2002), "Un nuevo algoritmo de microagregación: método de la distancia máxima modificado", en *XXVII Congreso Nacional de Estadística e Investigación Operativa*, Lleida, (en premsa)

Josep M. Mateo-Sanz, Ángel Torres-Aragó i Josep Domingo-Ferrer (2003), "Microagregación multivariante: estudio de las particiones del conjunto de variables", en *XXVII Congreso Nacional de Estadística e Investigación Operativa*, Lleida, (en premsa).

Francesc Sebé, Josep Domingo-Ferrer, Josep M. Mateo-Sanz, Antoni Martínez, Àngel Torres i Narcís Macià (2002), *The rankswapping algorithm (software, documentation and related papers)*, Deliverables 1-1.D5bis i 1-1.D6bis del projecte europeu CASC (IST-2000-25069).

Vicenç Torra, Josep Domingo-Ferrer i Àngel Torres (2003), "Data mining methods for linking data coming from several sources", *3rd UNECE/EUROSTAT Joint Worksession on Statistical Data Confidentiality*, Luxemburg.

Referències

N. A. Adam i J. C. Wortmann (1989), Security-control methods for statistical databases: a comparative study, *ACM Computing Surveys*, vol. 21: 515-556.

N. Anwar (1993), *Micro-aggregation - The Small Aggregates Method*, internal report, Eurostat.

B. C. Arnold, N. Balakrishnan i H. N. Nagaraja (1993), *A First Course in Order Statistics*. New York: Wiley.

J. Bacaria-Martrus (1993), "El secret estadístic: contingut jurídic", *Qüestiió*, vol. 17, pp. 405-410.

Y. Baeyens i D. Defays (1999), "Estimation of variance loss following microaggregation by the individual ranking method", in *Proceedings of Statistical Data Protection'98*. Luxembourg: Office for Official Publications of the European Communities, pp. 101-108.

J. G. Bethlehem, W. J. Keller i J. Pannekoek (1990), "Disclosure control of microdata", *Journal of the American Statistical Association*, vol. 85, pp. 38-45.

U. Blien, H. Wirth i M. Müller (1992), "Disclosure risk for microdata stemming from official Statistics", *Statistica Neerlandica*, vol. 46, pp. 69-82.

G. E. P. Box i G. C. Tiao (1992), *Bayesian Inference in Statistical Analysis*. New York: Wiley.

Crystal Ball, `http://www.cbpro.com`

T. Dalenius (1977), "Towards a methodology for statistical disclosure control", *Statistik Tidskrift*, vol. 15, pp. 429-444.

T. Dalenius i S. P. Reiss (1982), Data-swapping: a technique for disclosure control, *Journal of Statistical Planning and Inference*, vol. 6: 73-85.

T. Dalenius (1988), *Controlling Invasion of Privacy in Survey*, Department of Development and Research, Statistical Research Unit. Stockholm: Statistics Sweden.

R. A. Dandekar, J. Domingo-Ferrer i F. Sebé (2002), "LHS-Based Hybrid Microdata vs. Rank Swapping and Microaggregation for Numeric Microdata Protection", *Lecture Notes in Computer Science*, vol. 2316, pp. 153-162.

D. Defays i N. Anwar (1995), Micro-aggregation: a generic method", in *Proc. of the 2nd International Symposium on Statistical Confidentiality,* Luxembourg: Office for Official Publications of the European Communities, 69-78.

D. Defays i P. Nanopoulos (1993), Panels of enterprises and confidentiality: the small aggregates method, in *Proc. of 92 Symposium on Design and Analysis of Longitudinal Surveys*, Ottawa: Statistics Canada, 195-204.

A. G. DeWaal i L. C. R. J. Willenborg (1995), Global recodings and local suppressions in microdata sets, in *Proc. of Statistics Canada Symposium 95,* Ottawa: Statistics Canada, 121-132.

J. Domingo-Ferrer i J.M. Mateo-Sanz (1998), *Practical Data-Oriented Microaggregation for Statistical Disclosure Control,* Research Report DEI-RR-98-005, Department of Computer Science, Tarragona: Universitat Rovira i Virgili.

J. Domingo-Ferrer i J.M. Mateo-Sanz (1999), On resampling for statistical confidentiality in contingency tables, *Computers & Mathematics with Applications*, no. 38:13-32.

J. Domingo-Ferrer i J.M. Mateo-Sanz (1999) Information loss in continuous data masking, in *Optimizing Data Utility Within Framework of Confidentiality*, Washington D.C.: U.S. Census Bureau.

J. Domingo-Ferrer i V. Torra (2001), "A quantitative comparison of disclosure control methods for microdata", in *Confidentiality, Disclosure and Data Access*, eds. P. Doyle, J. Lane, J. Theeuwes and L. Zayatz. Amsterdam: North-Holland, pp. 111-133.

J. Domingo-Ferrer i V. Torra (2001), "Disclosure protection methods and information loss for microdata", en *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, eds. L. Zayatz, P. Doyle, J. Theeuwes and J. Lane, Amsterdam: North-Holland, 2001, pp. 91-110. ISBN 0-444-50761-2.

J. Domingo-Ferrer i J.M. Mateo-Sanz (2002), "Practical data-oriented microaggregation for statistical disclosure control", *IEEE Transactions on Knowledge and Data Engineering,* vol. 14, pp. 189-201.

G. T. Duncan i D. Lambert (1986), "Disclosure-limited data dissemination", *Journal of the American Statistical Association*, vol. 81, pp. 10-28.

M. J. Elliot, C. J. Skinner i A. Dale (1999), Special uniques, random uniques and sticky populations: some counterintuitive effects of geographical detail on disclosure risk", *Research in Official Statistics*, vol. 1(2): 53-67.

A. D. Gordon i J. T. Henderson (1977), "An algorithm for Euclidean sum of squares classification", *Biometrics*, vol. 33, pp. 355-362.

S. L. Hansen i S. Mukherjee (2002), "A polynomial algorithm for optimal microaggregation", manuscript.

J. A. Hartigan (1975), *Clustering Algorithms*, Wiley.

G. R. Heer (1993), A bootstrap procedure to preserve statistical confidentiality in contingency tables, in *Proc. of the International Seminar on Statistical Confidentiality*, ed. D. Lievesley, Luxembourg: Office for Official Publications of the European Communities, pp. 261-271.

A. Hundepool, L. Willenborg, A. Wessels, L. van Gemerden, S. Tiourine i C. Hurkens (1998), *μ-Argus 3.0 User's Manual*. Voorburg: Statistics Netherlands.

A. Hundepool i L. Willenborg (1999), ARGUS: Software from the SDC project, in *Proc. of Joint UNECE-Eurostat Work Session on Statistical Data Confidentiality*, Luxembourg: UNECE-Eurostat, pp. 87-98.

M. A. Jaro (1989), Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida, *Journal of the American Statistical Association*, vol. 84:414-420.

Joint Photographic Experts Group, Standard IS 10918-1 (ITU-T T.81). `http://www.jpeg.org`

W. J. Keller i J. G. Bethlehem (1992), "Disclosure protection of microdata: problems and solutions", *Statistica Neerlandica*, vol. 46, pp. 5-19.

J. J. Kim (1986), A method for limiting disclosure in microdata based on random noise and transformation, in *Proc. of the ASA Sect. on Survey Res. Meth.*, pp. 303-308.

P. Kooiman, L. Willenborg i J. M. Gouweleeuw (1997), PRAM: A method for disclosure limitation of microdata, CBS research paper 9705. Available from `http://www.cbs.nl/research`

C. K. Liew, U. J. Choi i C. J. Liew (1985), A data distortion by probability distribution, *ACM Transactions on Database Systems*, vol. 10: 395-411.

R. J. A. Little (1993), Statistical analysis of masked data, *Journal of Official Statistics*, vol. 9:407-426.

J.M. Mateo-Sanz i J. Domingo-Ferrer (1998), "A comparative study of microaggregation methods", *Qüestiió*, vol. 22, no. 3, pp. 511-526.

J.M. Mateo-Sanz i J. Domingo-Ferrer (1999), "A method for data-oriented multivariate microaggregation", in *Proceedings of Statistical Data Protection'98*. Luxembourg: Office for Official Publications of the European Communities, pp. 89-99.

N. E. Matloff (1986), Another look at the use of noise addition for database security, in *Proc. of IEEE Symposium on Security and Privacy*, pp. 173-180.

R. J. Mokken, P. Kooiman, J. Pannekoek i L. Willenborg (1992), "Disclosure risks for microdata", *Statistica Neerlandica*, vol. 46, pp. 49-67.

R. Moore (1996), Controlled data swapping techniques for masking public use microdata sets, U. S. Bureau of the Census (unpublished manuscript).

A. Oganian i J. Domingo-Ferrer (2001), "On the complexity of optimal microaggregation for statistical disclosure control", *Statistical Journal of the United Nations Economic Commission for Europe*, vol. 18, no. 4.

A. Oganian, *Security and Information loss in Statistical Database Protection*, Tesi Doctoral, Departament de Matemàtica Aplicada 4, Universitat Politècnica de Catalunya, 2003.

D. Pagliuca i G. Seri (1999), *Some Results of Individual Ranking Method on the System of Enterprise Accounts Annual Survey*, Esprit SDC Project, Deliverable MI-3/D2.

S. P. Reiss (1984), Practical data-swapping, *ACM Transactions on Database Systems*, vol. 9:20-37.

S. M. Samuels (1998), "A bayesian, population- genetics-inspired approach to the unique problem in microdata disclosure risk assessment", a *Preproceedings of Statistical Data Protection'98*, Lisboa: Eurostat, 1998. A aparèixer a J. Domingo-Ferrer (ed.), *Statistical Disclosure Protection'98*, IOS Press.

G. Sande (2001), "Methods for data-directed microaggregation in one or more dimensions", in *Federal Committee on Statistical Methodology Research Conference*, Arlington VA, Nov. 14-16.

D. Schackis (1993), *Manual on Disclosure Control Methods*, Luxembourg: Eurostat.

F. Sebé, J. Domingo-Ferrer, J.M. Mateo-Sanz i V. Torra (2002), "Post-masking optimization of the tradeoff between information loss and disclosure risk in masked microdata sets", *Lecture Notes in Computer Science*, vol. 2316, pp. 163-171.

C. E. Shannon (1948), "A mathematical theory of communication", *Bell Systems Technical Journal*, vol. 27, pp. 379-423, 623-656.

C. J. Skinner, C. Marsh, S. Openshaw i C. Wymer (1990), "Disclosure avoidance for census microdata in Great Britain", in *Proceedings of the 1990 Annual Research Conference*, Washington D.C.: U. S. Bureau of the Census, pp. 131-143.

C. J. Skinner (1992), "On identification disclosure and prediction disclosure for microdata", *Statistica Neerlandica*, vol. 46, pp. 21-32.

C. Skinner, C. Marsh, S. Openshaw i C. Wymer (1994), Disclosure control for census microdata, *Journal of Official Statistics*, vol. 10:31-51.

G. Sullivan i W. A. Fuller (1989), The use of measurement error to avoid disclosure, in *Proc. of the ASA Sect. on Survey Res. Meth.*, pp. 802-807.

G. Sullivan i W. A. Fuller (1990), Construction of masking error for categorical variables, in *Proc. of the ASA Sect. on Survey Res. Meth.*, pp. 435-439.

J. H. Ward (1963), Hierarchical grouping to minimize an objective function, *Journal of the American Statistical Association*, vol. 58: 236-244.

L. Willenborg i T. De Waal (1996), *Statistical Disclosure Control in Practice*, Springer LNS 111.

L. Willenborg i T. de Waal (2001), *Elements of Statistical Disclosure Control*. New York: Springer-Verlag.

W. E. Winkler (1998), Re-identification methods for evaluating the confidentiality of analytically valid microdata, in *Proc. of Statistical Data Protection'98*, ed. J. Domingo-Ferrer, Luxembourg: Office for Official Publications of the European Communities, pp. 319-335.

P. De Wolf, J. M. Gouweleeuw, P. Kooiman i L. Willenborg (1999), Reflections on PRAM, in *Statistical Data Protection*, ed. J. Domingo-Ferrer, Luxembourg: Office for Official Publications of the European Communities, pp. 337-349.

L. V. Zayatz (1991), Estimation of the percent of unique population elements on a microdata file using the sample, Bureau of the Census, Statistical Research Division Report Series, Census/SRD/RR-91/08.