Universitat
Autònoma
de Barcelona


# PROTECCIÓ DE DADES CATEGÒRIQUES I ANÀLISI DE LA PÈRDUA D'INFORMACIÓ

Memòria del Projecte Fi de Carrera
d'Enginyeria en Informàtica
realitzat per
Jordi Marés Soler
i dirigit per
Josep Puyol Gruart

Bellaterra, 17 de Juny de 2010

Universitat Autònoma de Barcelona

Escola Tècnica Superior d'Enginyeria

El sotasignat, Josep Puyol Gruart

Professor/a de l'Escola Tècnica Superior d'Enginyeria de la UAB,

**CERTIFICA**:

Que el treball a què correspon aquesta memòria ha estat realitzat sota la seva direcció per en Jordi Marés Soler

I per tal que consti firma la present.

Signat: Josep Puyol Gruart

Bellaterra, 17 de juny de 2010

## *Agraïments*

*M'agradria agrair especialment la col·laboració i la confiança dipositada en el projecte tant del Dr. Josep Puyol Gruart, director del projecte, com del Dr. Vicenç Torra Reventós ja que sense ells aquest projecte no hauria estat possible.*

*També agrair a totes les persones que han cregut i donat suport a aquest projecte.*

# ÍNDEX

# 0.- INTRODUCCIÓ

Aquest projecte és una petita part d'un altre projecte a nivell estatal anomenat ARES (Advanced Research on Information Security and Privacy) el qual està finançat pel Ministeri de Ciència i Innovació, i com el seu nom indica està centrat en la investigació sobre seguretat de la informació.

L'objectiu d'aquest projecte és ajudar a millorar l'important camp de la privadesa de dades. Concretament es centra en les bases de dades estadístiques ja que son una gran font d'informació privada de moltes persones a tot el món i requereix un gran control per tal de garantir la privacitat. Per aclarir l'objectiu del projecte es defineixen a continuació els conceptes bàsics de privacitat de dades i de base de dades estadística.

*Què és la privacitat de dades?*

La privacitat de dades pot ser definida com la relació entre la recol·lecció i difusió de dades, tecnologia, l'expectació publica de privacitat, i els temes legals i polítics que els envolten. Tot i que les motivacions de la privacitat i els tipus de dades a tractar poden ser molt diferents, totes les aplicacions ténen en comú que es tracta d'informació confidencial.

Des del punt de vista del propietari de les dades trobem que no es desitja que certa informació privada sigui difosa per tal d'evitar, per exemple, discriminacions, danys a la seva reputació,...

Pel que fa al punt de vista financer podria ser molt perillós i perjudicial si es difonguessin dades sobre transaccions financeres ja que seria molt fàcil ser victima de fraus o robatori d'identitat.

El punt de vista mèdic ens diu que no es poden difondre expedients mèdics de persones simplement perquè és informació molt privada i perquè potser podria afectar la seva feina i cobertura d'assegurança. Així mateix, aquest tipus d'informació també podria revelar informació de la vida personal com per exemple l'activitat sexual.

Així doncs és obvi que es tracta d'un tema molt important.

*Què és una base de dades estadística?*

Una base de dades estadística és aquella que emmagatzema informació extreta a persones per tal de realitzar estudis sobre les dades i extreu-re'n resultats.

Per exemple, una base de dades estadística en el cas mèdic seria tota aquella que guarda les relacions de cada pacient amb la seva patologia i el seu tractament, i se'n poden extreure percentatges de cada patologia, exits de tracaments...

Aquestes bases de dades contenen habitualment informació personal i la seva publicació pot atemptar contra els drets de privacitat de les persones. Per a això cal que siguin protegides abans de ser divulgades o passades a terceres persones per al seu estudi.

Aquestes bases de dades estadístiques poden contenir dades de dos tipus: dades contínues i dades categòriques.

Es consideren dades contínues aquelles que són numèriques i s'hi poden realitzar operacions aritmètiques. Per exemple el salari d'una persona.

Les dades contínues per contra són aquelles que prenen valors dins un conjunt finit de possiblitats i no s'hi poden realitzar operacions aritmètiques. Per exemple un codi postal.

Aquest projecte es centrarà només en dades de caire categòric fent una comparació dels diferents mètodes existents i presentant un estudi detallat de cada un d'ells sobre com trobar els seus paràmetres òptims.

L'estructura d'aquesta memòria serà la següent:

- El primer capítol es centrarà en la descripció detallada de cadascún dels mètodes de protecció de dades categòriques actuals.

- En el segon capítol es descriurà el programa d'anàlisi de pèrdua d'informació disenyat per a aquest projecte, i el programa d'anàlisi del risc de revelació utilitzat.

- El tercer capítol mostrarà els resultats dels anàlisis de pèrdua d'informació i risc de revelació de cadascún dels mètodes descrits en el capítol primer.

- El quart capítol està basat en explicar el mètode d'optimització desenvolupat per al mètode de protecció PRAM explicat en el segon capítol.

- Finalment hi haurà les conclusions que se'n poden extreure i quin serà el treball futur a realitzar.

# 1.- ELS MÈTODES DE PROTECCIÓ ACTUALS

Actualment existeixen 6 mètodes de protecció de dades categòriques: la Microagregació, el Top-Coding, el Bottom-Coding, el Global Recoding, el Post Randomization Method (PRAM), i el Rank Swapping.

En aquest projecte s'analitzaràn tots ells, però abans es donarà una descripció detallada de cada un.

## *1.1 – MÈTODES PERTURBATIUS*

Els mètodes perturbatius són aquells que distorsionen les dades abans de ser publicades, d'aquesta manera s'aconsegueix que combinacions úniques de valors al dataset original desapareixin i n'apareguin de noves creant una confusió beneficiària per a la preservació de la confidencialitat estadística.

La idea bàsica d'aquests mètodes és que es canvien alguns valors de les variables a protegir per altres categories existents a les descripcions d'aquestes.

Existèixen 3 mètodes perturbatius per a la protecció de dades categòriques: la Microagregació, el Post Randomization Method (PRAM), i el Rank Swapping.

### *1.1.1 – MICROAGREGACIÓ*

La Microagregació és un mètode perturbatiu que depèn de 2 paràmetres (*K* i *N*) i d'un conjunt de variables a protegir.

La idea d'aquest mètode és la següent: Partint el conjunt de variables a protegir en grups de *N* o menys, es creen clústers de mida com a mínim *K* agregant registres del fitxer de dades, i es calcula el valor mig de cada variable del grup amb els valors dels registres del cluster. Aquests valors mitjos són els valors que es publicaran per als registres continguts al clúster.

Aquesta idea ens porta a parlar de la *K*-anonimitat que és un dels pilars en que es basa la microagregació. Un conjunt de dades compleixen la *K*-anonimitat (per $K>1$) si per cada combinació de valors quasi-identificadors (conjunt de camps/variables que s'utilitzen per enllaçar base de dades), existeixen almenys *K* registres amb la mateixa combinació. Per tant si un intrús intenta enllaçar les dades protegides amb dades externes trobarà a les dades protegides com a mínim *K* possibles registres que coincideixen amb els quasi-identificadors utilitzats per a l'enllaçament.

Així doncs la microagregació aconsegueix la *K*-anonimitat posant els mateixos valors mitjos a *K* registres dins un clúster.

Per tal de minimitzar la pèrdua d'informació, els clústers han de ser el més homogènis possible, és a dir, els valors de cada registre per les variables que conté el grup han de ser el més semblants possibles.

Exemple 1.1

Tenim una base de dades on hi ha emmagatzemada informació sobre persones tal com el nom, edat, sexe, pès, ... Volem protegir la variable pès (*N*=1) aplicant una microagregació univariant amb *K*=3 , és a dir, agafant 3 registres per clúster (aplicant 3-anonimitat). És univariant perquè a l'hora de fer els clústers es tracta cada variable de forma independent.

Així en el gràfic de l'esquerra de la figura 1.1 es poden observar els diferents valors originals de la variable 'pès' als registres del 1 al 12. Els clústers estàn representats per colors i com es pot observar inclouen els *K* valors mes semblants del conjunt.

Llavors per cada clúster es mesuraria el valor mig i aquest seria el valor publicat per a tots els registres del clúster. Com es pot observar al gràfic de la dreta, tots els registres dins un clúster ténen el mateix valor.

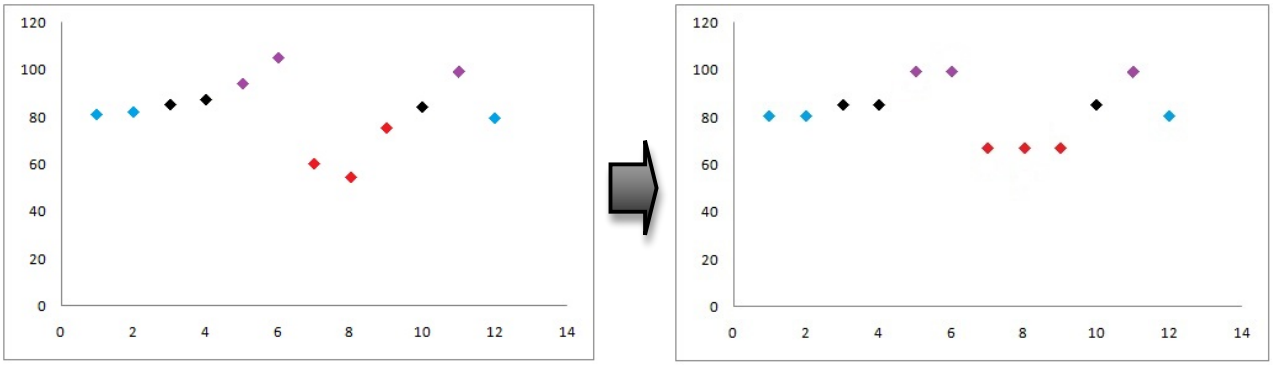



**Figura 1.1 – Exemple de l'efecte de la microagregació d'una variable amb k=3**

Existeixen diferents implementacions d'aquest mètode però en aquest projecte s'ha implementat la versió MDAV-generic (Maximum Distance to Average Vector) desenvolupada l'any 2003.

Aquesta versió es basa en generar un vector amb els valors mitjos de totes les variables a protegir respecte tot el dataset original, trobar el registre amb valors més distants als del vector calculat, trobar el registre amb valors més distants a

aquest registre trobat i formar un clúster al voltant de cada un dels dos registres amb els *K*-1 registres més semblants.
Per tal de realitzar el càlcul de la distància entre categories i de valors mitjos es necessiten uns operadors que ho calculin. Aquests operadors depènen del tipus de variables categòriques que es vulguin tractar: ordinals o nominals. Es diu que una variable categòrica és de caire ordinal si els seus valors determinen un ordre, i es dirà que una variable categòrica és de caire nominal en cas contrari.
Com a operadors de distància i de càlcul de mitjana s'han implementat els següents:

- Variables ordinals

  La distància entre dues categories (*a*,*b*) d'una variable ordinal és el nombre de categories que hi ha entre *a* i *b*, dividit entre el nombre total de categories que la variable pot prendre.

$$d_{ORD}(a,b) = \frac{\left|\left\{i \middle| a \le i \le b\right\}\right|}{\left|D(V_i)\right|}$$

  Pel que fà al càlcul dels valors mitjos en variables ordinals es fan servir el median i el convex median.

  El median diu que donada una llista ordenada creixent de categories, el valor mig és la categoria que ocupa la posició central. En termes de freqüències, el median seria la categoria la qual tant els seus predecessors com els seus successors ténen aproximadament la mateixa freqüència.

  Per exemple, el median de *S={1,2,2,5,6}* és 2.
  Si apliquem el median sobre les freqüències resultants de la transformació de la funció de freqüència a convexa, tenim el convex median. Tenint que *f* és la funció de freqüència original i *f'* la convexa:

$$f'(c_i) = \min(\max_{c_j \le c_i}(f(c_j)), \max_{c_j \ge c_i}(f(c_j)))$$

- Variables nominals

  La distància entre dues categories (*a*,*b*) d'una variable nominal és 0 si són iguals o 1 si són diferents.

$$d_{NOM}(a,b) = \begin{cases} 0, & a = b \\ 1, & a \ne b \end{cases}$$

Pel que fà al càlcul dels valors mitjos en variables nominals es fa servir la regla de pluralitat.

Aquesta regla diu que per un conjunt de categories es selecciona com a valor mig el valor més freqüent.

Així si tenim el conjunt de valors *S={1,2,4,2,7,8,1,2,9,2}*, el valor mig seria 2.

Diagrama de flux:

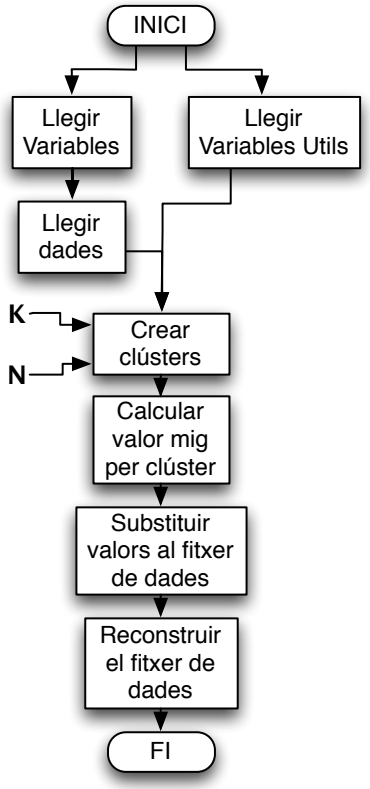



**Figura 1.2 – Diagrama de flux del mètode de la Microagregació**

## *1.1.2 – POST RANDOMIZATION METHOD (PRAM)*

El Post Randomization Method (PRAM) és un mètode perturbatiu que depèn d'un sol paràmetre *P* i d'un conjunt de variables a protegir.

La idea d'aquest mètode és que cada variable té associada una matriu de Markov indicant la probabilitat d'intercanvi d'una categoria per una altra de totes les categories possibles. Així per cada registre del dataset original es miren els valors de les variables a protegir especificades i depenent del mecanisme probabilistic es canvien els seus valors.

En la implementació realitzada per a aquest projecte les matrius de markov es caracteritzen per tenir un valor únic en la posició de cada fila corresponent a la diagonal, és a dir a la probabilitat de mantenir la categoria (no canviar), i una probabilitat de canvi igual per a cada categoria restant a la fila. Per tal de calcular els valors de la diagonal es fa servir la següent fórmula:

$$p_{ii} = 1 - \theta T_v(K) / T_v(i)$$

on,

- $T_v(x)$ és la freqüència de la categoria x en la variable actual
- $K$ és la categoria de la variable actual amb menor freqüència més gran que 0
- $i$ és la categoria actual, cumplint sempre que $T_v(i) \geq T_v(K)$.
- $\theta$ és un paràmetre tal que $0 < \theta < 1$, el qual és calculat mitjançant el paràmetre d'entrada $P$ del mètode. Així limitant $P$ als valors 1...9, tenim que $P=10\theta$. Així doncs amb el paràmetre d'entrada $P$ es regula la intensitat de la protecció que s'aplicarà.

Un cop tenim calculat l'element de la fila corresponent a la diagonal, dividim la diferència entre 1 i el valor de l'element calculat és dividit entre totes les posicions restants a la fila. Aquest valor serà la probabilitat de canvi de categoria.

Exemple 1.2

Tenim $P = 2$ i el següent fitxer de dades amb 3 variables i 6 files de dades.

| V1 | V2 | V3 |
|---|---|---|
| 1 | 3 | 09 |
| 8 | 3 | 12 |
| 7 | 8 | 00 |
| 1 | 3 | 01 |
| 7 | 8 | 22 |
| 7 | 2 | 04 |

**Taula 1.1 – Fitxer original de l'exemple amb 3 variables i 6 files**

Suposant que V2 pot agafar valors entre 1 i 9, trindriem les següents freqüències,

| Categoria | Freqüència |
|---|---|
| 1 | 0 |
| 2 | 1 |
| 3 | 3 |
| 4 | 0 |
| 5 | 0 |
| 6 | 0 |
| 7 | 0 |
| 8 | 2 |
| 9 | 0 |

**Taula 1.2 – Freqüències de cada categoria corresponents a la variable V2 de l'exemple**

Amb aquestes freqüències tindriem la següent matriu de Markov,

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| **1** | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **2** | 0.025 | 0.8 | 0.025 | 0.025 | 0.025 | 0.025 | 0.025 | 0.025 | 0.025 |
| **3** | 0.0088 | 0.0088 | 0.93 | 0.0088 | 0.0088 | 0.0088 | 0.0088 | 0.0088 | 0.0088 |
| **4** | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| **5** | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| **6** | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| **7** | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| **8** | 0.0125 | 0.0125 | 0.0125 | 0.0125 | 0.0125 | 0.0125 | 0.0125 | 0.9 | 0.0125 |
| **9** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

**Taula 1.3 – Matriu de Markov corresponent a la variable V2 de l'exemple**

Com es pot observar els valors de la diagonal són molt mes alts que la resta ja que no es poden fer excessius canvis perquè hi hauria molta pèrdua d'informació.

També es pot observar que les categories amb menys aparicions són les que ténen la probabilitat de canvi més altes. Això és degut a que com més freqüència hi ha d'una categoria, menys probabilitat de ser descobert un individu i per tant no és tant important com una categoria que aparegui només una vegada la qual identifica inequivocament a un sol individu.

Diagrama de flux:

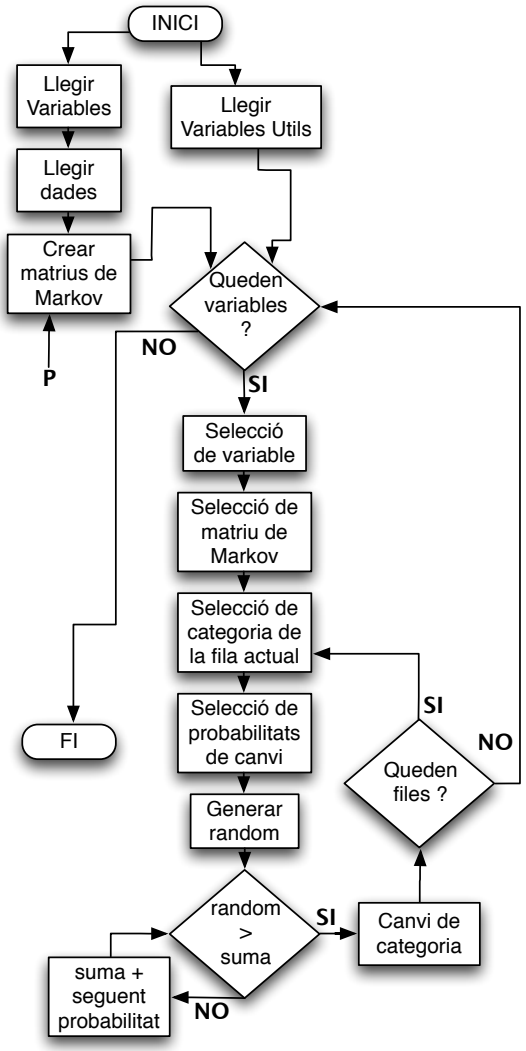



**Figura 1.3 – Diagrama de flux del mètode PRAM**

## *1.1.3 – RANK SWAPPING*

El Rank Swapping és un mètode de protecció perturbatiu que només depèn d'un sol paràmetre *P* i d'un conjunt de variables a protegir. A més, aquest mètode només es pot aplicar a variables categòriques ordinals ja que es requereix que sigui possible una ordenació de les categories.

La idea bàsica del Rank Swapping és la d'intercanviar categories dins d'un rang limitat pel valor de *P*.

Més extensament:

Tenim un fitxer de dades categòriques del qual volem protegir algunes variables. Suposem que volem protegir la variable 'V1', i que aquesta pot prendre 'n' valors/categories diferents. Agafant la categoria corresponent a V1 a cada fila del fitxer iterativament:

- S'ordenen les categories de V1 que apareixen al fitxer original
- Es busca la posició de la categoria original (*index*)
- Es calcula la mida del rang (*r*): *r = P % de les categories totals*
- Així tenim un rang [index-rang, index+rang] en el qual hi ha totes les categories candidates a substituïr la original
- Es selecciona aleatòriament una de les categories que conté el rang i es substitueix al fitxer de dades.

Exemple 1.3

Seguint amb l'exemple 1.2 de l'apartat anterior, tenim el següent fitxer de dades categòriques:

| V1 | V2 | V3 |
|---|---|---|
| 1 | 3 | 09 |
| 8 | 3 | 12 |
| 7 | 8 | 00 |
| 1 | 3 | 01 |
| 7 | 8 | 22 |
| 7 | 2 | 04 |

**Taula 1.4 – Fitxer de dades original a protegir**

Volem protegir la variable V3 mitjançant Rank Swapping i tenim que la ordenació dels valors que prèn la variable és:

*{00,01,04,09,12,22}*

Suposem també que tenim *P =10*, així el càlcul del rang sería:

$$r = \frac{Num\_categories(V3) * p}{100} = \frac{10 * 10}{100} = 1$$

Així, tenim que els conjunts de valors candidats a substituir l'original contindràn el valor anterior a l'original, el posterior, i el mateix valor original.

En els casos dels valors existents al fitxer en la variable V3 tindriem els següents conjunts de valors candidats per a cada posició:

| Valor original | Conjunt de candidats |
|---|---|
| 09 | {04, 09, 12} |
| 12 | {12, 22} |
| 00 | {00, 01} |
| 01 | {00, 01, 04} |
| 22 | {12, 22} |
| 04 | {01, 04, 09} |

**Taula 1.5 – Conjunts de categories candidates a substituïr cada una de les originals**

A partir d'aquí la selecció del valor que ha de substituïr a l'original sería de forma aleatòria d'entre els valors que formen cada conjunt.

Diagrama de flux:

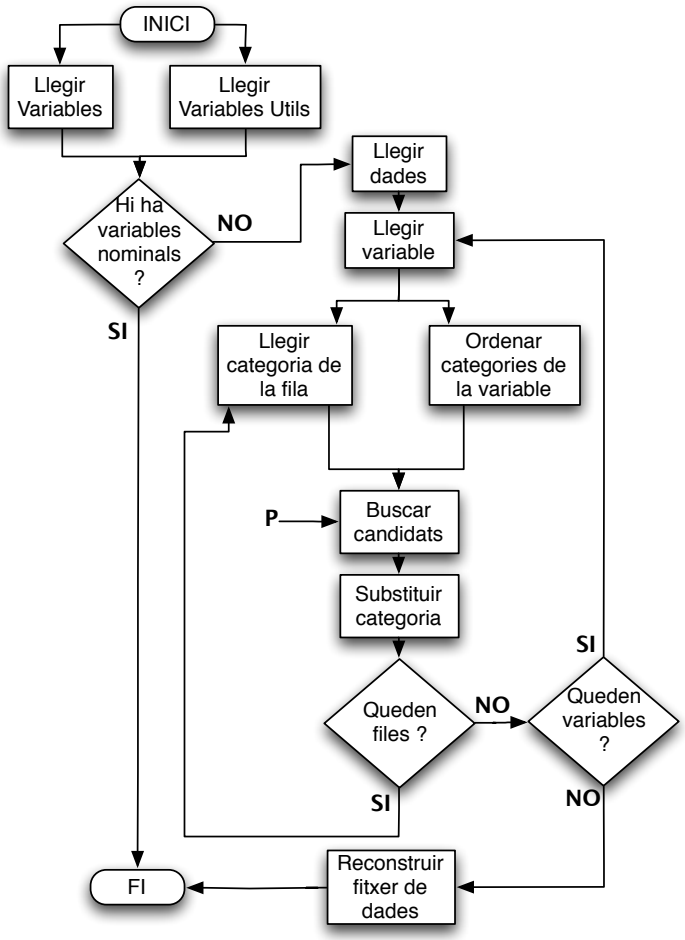



**Figura 1.4 – Diagrama de flux del mètode Rank Swapping**

## *1.2 – MÈTODES NO-PERTURBATIUS*

Els mètodes no-perturbatius no alteren les dades com els mètodes perturbatius, sinó que produeixen una supressió o reducció del detall del dataset original.

Un exemple típic per explicar aquest tipus de mètodes és el següent:

Exemple 1.4

Tenim dues bases de dades estadístiques $B_1$ i $B_2$.
$B_1$ conté informació mèdica sobre uns pacients però per ajudar amb la privacitat de dades només guarden la data de naixement, el sexe i el codi postal de cada pacient.

D'aquesta manera sembla que no es poden identificar exactament els pacients, però ara imaginem que $B_2$ conté les dades dels votants en unes eleccions. Aquesta base de dades també conté els camps anteriors que té guardats $B_1$ i a més també conté el nom, adreça,... de tots els votants.
Així complementant la informació de $B_1$ amb la de $B_2$ es poden a reidentificar molts pacients i per tant la seva privacitat de dades quedaria violada.
Si es generalitzessin els atributs que té $B_1$ susceptibles a ser utilitzats per enllaçar $B_2$ ja no seria tant trivial la reidentificació de cada individu.

Considerem que tenim els seguents codis postals relacionats amb diverses entrades a la base de dades: 08081, 08250, 08032, 08211, 08057.
Aplicant un mètode de generalització no-perturbatiu podem agrupar-los en les categories 08080 i 08200. O si volem encara més generalització podem deixar una sola categoria 08000 el qual englobaria encara moltes més entrades dificultant aixi la reidentificació de cada individu.

S'ha de tenir en compte també que com més es generalitzen les dades, més pèrdua d'informació hi haurà.

| Sexe | Data de naixement | Codi Postal |
|---|---|---|
| Masculí | 10/7/1985 | 08081 |
| Femení | 9/11/1984 | 08250 |
| Femení | 1982 | 08032 |
| Masculí | 2001 | 08211 |
| Femení | 2005 | 08057 |

| Sexe | Data de naixement | Codi Postal |
|---|---|---|
| Masculí | 1980 | 08000 |
| Femení | 1980 | 08200 |
| Femení | 1980 | 08000 |
| Masculí | 2000 | 08200 |
| Femení | 2000 | 08000 |

**Figura 1.5 – Exemple de l'efecte d'un mètode no-perturbat**

Existèixen 3 mètodes no-perturbatius per a la protecció de dades categòriques: el Global Recoding, el Top Coding, i el Bottom Coding.

### *1.2.1 – GLOBAL RECODING*

El Global Recoding és un mètode no-perturbatiu que depèn només d'un paràmetre *P* i d'un conjunt de variables a protegir.

Es pot mirar com una funció F sobre una variable categòrica V que dóna com a resultat una variable V', que compleix que |D(V')|>|D(V)|, on D(V) és el domini de la variable V i |·| és l'operador de cardinalitat.

En aquest mètode la selecció de quines categories s'han de recodificar es determina mitjançant una funció que en el cas d'aquest projecte selecciona les *P* categories de la variable V amb menys freqüència dins el fitxer de dades original.

A la implementació realitzada en aquest projecte s'utilitza un esquema de recodificació per cada variable que conté tota la informació de com s'ha de recodificar cada categoria. Aquest esquema té la seguent estructura:

```
3  2
9
1.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0
pr      1.0  3  01  02  06
st      1.0  2  09  10
to      1.0  2  pr  st
```

**Figura 1.6 – Exemple de l'estructura d'un fitxer de recodificació**

A la primera fila trobem el nombre de files recodificadores que conté el fitxer seguit del nombre de files recodificadores que modifiquen categories del fitxer original.

A la segona fila s'indica el nombre de probabilitats d'intercanvi de cada categoria susceptible a esser canviada.

La tercera fila conté totes les probabilitats d'intercanvi. En el cas d'aquest projecte suposem que s'han de canviar sempre i per tant fem servir en tot cas probabilitats de 1.0 (100%).

A partir de la quarta fila trobem les files de recodificadores les quals están formades per diferents camps: una etiqueta corresponent a la nova categoria que es crea, la probabilitat d'intercanvi associada a aquesta nova etiqueta, el nombre de categories que es recodificarán, i una llista de les categories a substituïr per la categoria creada.

Com es pot observar a l'última fila recodificadora de l'esquema de la figura 1.6, també es poden recodificar etiquetes noves creades dins el mateix esquema, és a dir, les etiquetes *pr* i *st* són creades per l'esquema per tal de substituïr categories inicials, però les dues són alhora recodificades per l'etiqueta *to*.

Exemple 1.5

Considerem el següent fitxer de dades amb les corresponents variables,

| | V1 | V2 | V3 | V4 | V5 |
|---|---|---|---|---|---|
| r1 | 10 | 1 | 00 | 01 | 1 |
| r2 | 23 | 3 | 20 | 04 | 6 |
| r3 | 02 | 4 | 10 | 00 | 7 |
| r4 | 10 | 3 | 20 | 02 | 4 |
| r5 | 02 | 1 | 30 | 02 | 3 |

| | |
|---|---|
| V1 : nominal | Domini = {02, 10, 23, 30} |
| V2 : ordinal | Domini = {1, 2, 3, 4} |
| V3 : nominal | Domini = {00, 10, 20, 30} |
| V4 : ordinal | Domini = {00, 01, 02, 03, 04} |
| V5 : nominal | Domini = {1, 2, 3, 4, 5, 6, 7} |

**Taula 1.6 – Fitxer de dades original a protegir**

Volem protegir la variable V2 amb el següent esquema de recodificació,

```
2   2
4
1.0  1.0  1.0  1.0
C1  1.0  2   1  2
C2  1.0  2   3  4
```

Suposant que tenim $P = 2$ busquem les $P$ categories menys freqüents a V2,

$$\text{Cats\_menys\_freq} = \{2, 4\}$$

Així doncs, el fitxer final amb la variable V2 protegida , seguint l'esquema de recodificació, quedaria de la següent manera,

| | V1 | V2 | V3 | V4 | V5 |
|---|---|---|---|---|---|
| r1 | 10 | 1 | 00 | 01 | 1 |
| r2 | 23 | 3 | 20 | 04 | 6 |
| r3 | 02 | C2 | 10 | 00 | 7 |
| r4 | 10 | 3 | 20 | 02 | 4 |
| r5 | 02 | 1 | 30 | 02 | 3 |

**Taula 1.7 – Fitxer de dades protegit**

Diagrama de flux:

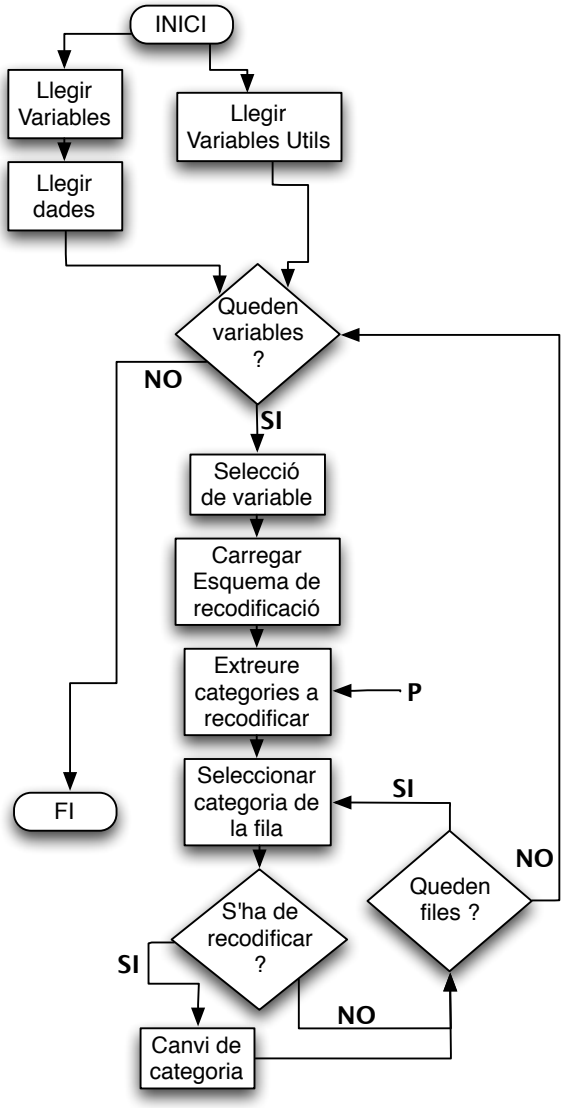



**Figura 1.7 – Diagrama de flux del mètode Global Recoding**

### *1.2.2 – TOP CODING*

El Top Coding es un mètode no-perturbatiu que només depèn d'un parametre *P* i d'un conjunt de variables a protegir. A més és un mètode que només s'aplica a variables ordinals ja que es necessita ordenar el conjunt de categories que pot prendre la variable per a fer-ne la selecció.

Es pot dir que es tracta d'una variació del Global Recoding ja que té el mateix principi però el que canvia es a l'hora de seleccionar quines categories han de ser protegides.

Aquest mètode no fa servir fitxers de recodificació sinó que simplement utilitza el paràmetre *P* d'entrada (un enter) per seleccionar les *P* primeres categories del conjunt ordenat que pot prendre la variable a protegir.

A la implementació realitzada en aquest projecte les categories són declarades en el fitxer descriptor de variables. Així aquest mètode agafa les categories que apareixen en aquest fitxer, les ordena i n'agafa les *P* primeres. L'estructura del fitxer és la següent:

```
...          ...          ...          ...
DEGREE   10      1    2      8    1 2 3 4 5 6 9
...          ...          ...          ...
```

**Figura 1.8 – Estructura del fitxer descriptor de variables**

A la primera columna trobem el nom de la variable, que en aquest exemple és 'DEGREE'.

A la segona columna hi ha la posició (caràcter) de cada fila del fitxer de dades on comença el valor de la variable.

En tercer lloc tenim la longitud en caràcters dels valors que pren la variable.

A la quarta posició s'indica el tipus de la variable: 0 – no categorica, 1 - nominal, 2 - ordinal.

A la cinquèna columna hi ha el nombre de categories que pot prendre la variable.

Finalment, a la última columna trobem la llista de totes les categories que pot prendre la variable. D'aquí és d'on agafa el Top Coding les categories per ordenar-les i fer-ne la selecció per recodificar-ne les *P* primeres en una de nova. En aquest projecte s'ha triat la categoria '9'.

Exemple 1.6

Seguint amb l'exemple 1.5, ara volem protegir la variable V2 amb $P = 2$. Recordem que tenim el fitxer original següent:

| | V1 | V2 | V3 | V4 | V5 |
|---|---|---|---|---|---|
| r1 | 10 | 1 | 00 | 01 | 1 |
| r2 | 23 | 3 | 20 | 04 | 6 |
| r3 | 02 | 4 | 10 | 00 | 7 |
| r4 | 10 | 3 | 20 | 02 | 4 |
| r5 | 02 | 1 | 30 | 02 | 3 |

**Taula 1.8 – Fitxer de dades original a protegir**

V1 : nominal Domini = {02, 10, 23, 30}
V2 : ordinal Domini = {1, 2, 3, 4}
V3 : nominal Domini = {00, 10, 20, 30}
V4 : ordinal Domini = {00, 01, 02, 03, 04}
V5 : nominal Domini = {1, 2, 3, 4, 5, 6, 7}

Les *P* categoríes de V2 a recodificar serien les següents,

$$\text{Cats\_recod} = \{1, 2\}$$

Així el fitxer protegit (utilitzant '9' com a nova categoria) quedaria de la següent forma,

| | V1 | V2 | V3 | V4 | V5 |
|---|---|---|---|---|---|
| r1 | 10 | 9 | 00 | 01 | 1 |
| r2 | 23 | 3 | 20 | 04 | 6 |
| r3 | 02 | 4 | 10 | 00 | 7 |
| r4 | 10 | 3 | 20 | 02 | 4 |
| r5 | 02 | 9 | 30 | 02 | 3 |

**Taula 1.9 – Fitxer de dades protegit**

Diagrama de flux:

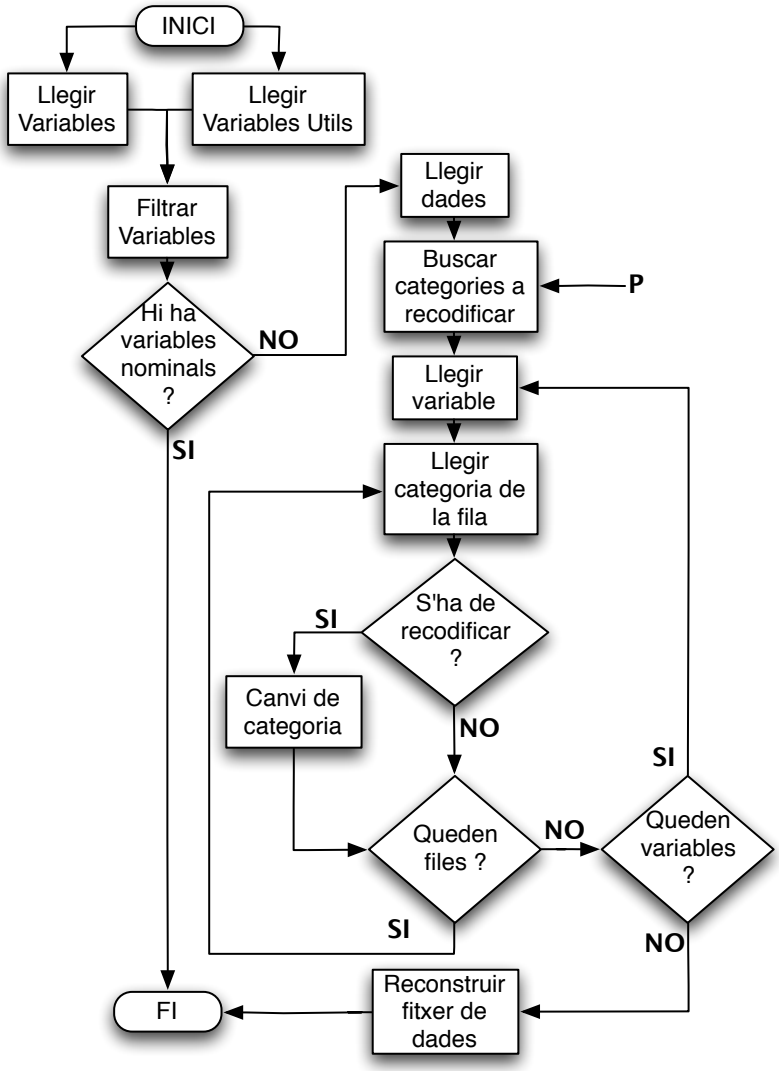



**Figura 1.9 – Diagrama de flux del mètode Top Coding**

## *1.2.3 – BOTTOM CODING*

El Bottom Coding es un mètode no-perturbatiu que només depèn d'un parametre *P* i d'un conjunt de variables a protegir. A més és un mètode que només s'aplica a variables ordinals ja que es necessita ordenar el conjunt de categories que pot prendre la variable per a fer-ne la selecció.

Es pot dir que aquest mètode també es tracta d'una variació del Global Recoding ja que té el mateix principi però el que canvia es a l'hora de seleccionar quines categories han de ser protegides.

Aquest mètode tampoc fa servir fitxers de recodificació sinó que simplement utilitza el paràmetre *P* d'entrada per seleccionar les *P* últimes categories del conjunt ordenat que pot prendre la variable a protegir.

A la implementació realitzada en aquest projecte les categories són declarades en el fitxer descriptor de variables. Així aquest mètode agafa les categories que apareixen en aquest fitxer, les ordena i n'agafa les *P* últimes. L'estructura del fitxer està descrita al mètode anterior Top Coding (*veure figura 1.7*).

Exemple 1.7

Com en l'exemple 1.6 volem protegir la variable V2 amb $P = 2$, però en aquest cas utilitzant el mètode Bottom Coding. Recordem el fixer original del que disposem:

| | V1 | V2 | V3 | V4 | V5 |
|---|---|---|---|---|---|
| r1 | 10 | 1 | 00 | 01 | 1 |
| r2 | 23 | 3 | 20 | 04 | 6 |
| r3 | 02 | 4 | 10 | 00 | 7 |
| r4 | 10 | 3 | 20 | 02 | 4 |
| r5 | 02 | 1 | 30 | 02 | 3 |

**Taula 1.10 – Fitxer de dades original a protegir**

V1 : nominal Domini = {02, 10, 23, 30}
V2 : ordinal Domini = {1, 2, 3, 4}
V3 : nominal Domini = {00, 10, 20, 30}
V4 : ordinal Domini = {00, 01, 02, 03, 04}
V5 : nominal Domini = {1, 2, 3, 4, 5, 6, 7}

Les categoríes a recodificar serien les següents,

$$Cats\_recod = \{3, 4\}$$

Així el fitxer protegit (utilitzant '9' com a nova categoria) quedaria de la següent forma,

| | V1 | V2 | V3 | V4 | V5 |
|---|---|---|---|---|---|
| r1 | 10 | 1 | 00 | 01 | 1 |
| r2 | 23 | 9 | 20 | 04 | 6 |
| r3 | 02 | 9 | 10 | 00 | 7 |
| r4 | 10 | 9 | 20 | 02 | 4 |
| r5 | 02 | 1 | 30 | 02 | 3 |

**Taula 1.11 – Fitxer de dades protegit**

## Diagrama de flux:

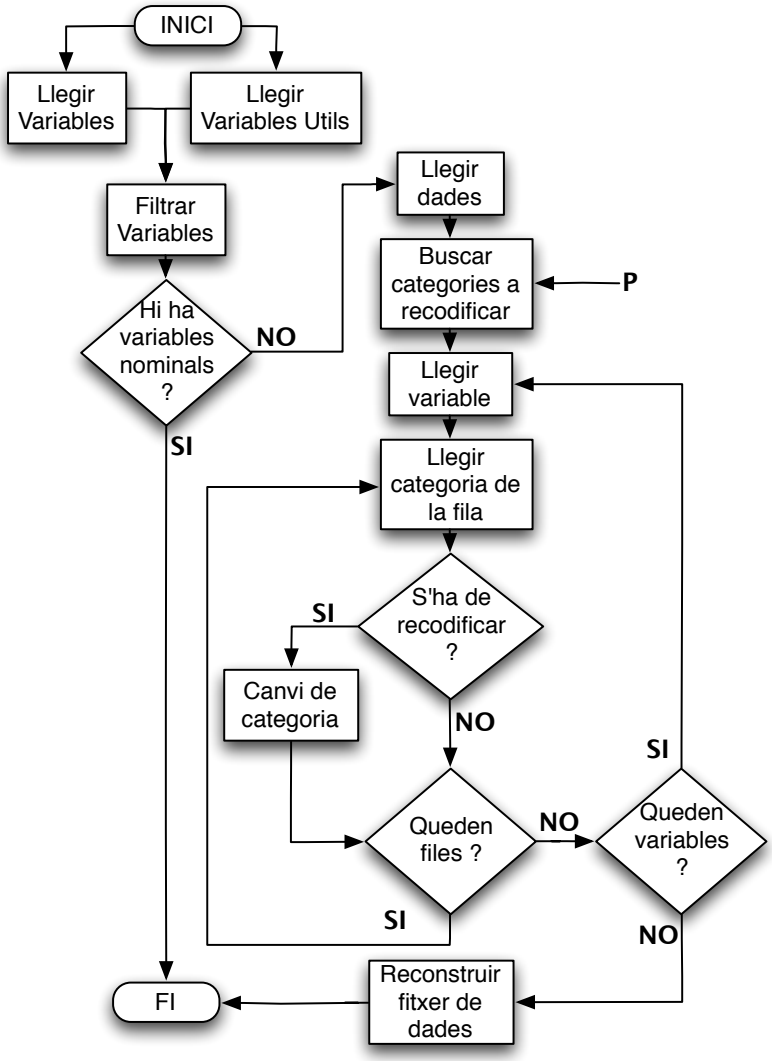



**Figura 1.10 – Diagrama de flux del mètode Bottom Coding**

# 2.- PROGRAMES AVALUADORS DE PÈRDUA D'INFORMACIÓ I RISC DE REVELACIÓ

Les mesures que determinen la qualitat d'una protecció son la pèrdua d'informació i el risc de revelació que generen.

*Què és la pèrdua d'informació?*

La pèrdua d'informació sorgeix quan es distorsionen massa les dades i és impossible extreu-re'n certa informació original. A més és una mesura que depen dels usos que tindràn les dades. No obstant, existeixen diversos usos de les dades i pot ser complicat identificar-los tots en el moment que una oficina estadística publiqui unes dades. La pèrdua d'informació pot ser molt negativa ja que si unes dades en ténen molta, no ens servirien de res de cara a estudis estadístics ja que haurien perdut totes les seves propietats originals. Així es necessita una manera genèrica que reflexi la quantitat de dany que ha provocat el mètode de protecció a les dades.
Per tant la pèrdua d'informació és un aspecte a minimitzar i es necessita un avaluador per quantificar la pèrdua en cada cas.

*Què és el risc de revelació?*

Pel què fa al risc de revelació, sorgeix quan es distorsionen poc les dades i són massa fàcil de trobar els valors originals per part d'un intrús. Això pot ser molt negatiu ja que si unes dades ténen un risc de revelació elevat s'estaria permetent la revelació de dades confidencials. Tal com passa amb la pèrdua d'informació necessitem una manera genèrica que reflexi com son de segures les dades després de protegir-les amb un cert mètode.
Per tant el risc de revelació és un aspecte a minimitzar i també es necessita un avaluador per quantificar la quantitat de dades originals que es poden obtenir en cada cas.

*Quin es el problema?*

El problema de minimitzar les dues mesures és que la pèrdua d'informació i el risc de revelació están inversament relacionats. Així que si insertem una protecció forta dins un conjunt de dades tindrem molt poc risc de revelació però molta pèrdua d'informació, i si en canvi insertem poca protecció tindrem molt risc de revelació i poca pèrdua d'informació.
Així doncs cal trobar un equilibri òptim entre aquestes dues mesures.

## *2.1 – AVALUADOR DE PÈRDUA D'INFORMACIÓ*

L'avaluador de pèrdua d'informació ha estat implementat expressament per a aquest projecte ja que fins ara només es disposava de la implementació per a dades contínues, no categòriques.

Aquesta implementació consta de tres tipus de càlculs per tal d'obtenir la quantitat de pèrdua d'informació que pateixen unes dades protegides: comparació directa entre valors categòrics, comparació de taules de contingència, i mesures basades en l'entropia. Tot i haver-hi tres tipus de càlculs, es descriuen quatre mètodes ja que l'ultim és una modificació del mètode basat en l'entropia. Tot seguit es descriuen detalladament aquests tipus de càlculs.

### *2.1.1 – DISTANCE-BASED INFORMATION LOSS (DBIL)*

Quan un fitxer de dades és protegit, la informació associada als objectes és en certa mesura modificada. Una manera per mesurar la distorisió de la informació és mesurant la diferència entre els valors inicials del fitxer original i els valors finals del fitxer protegit.

Així la distància entre valors es pot definir com la suma de les distàncies entre registres de cada fitxer, on cada distància entre registres és la suma de les distàncies entre els valors que prèn cada variable en els registres.

Més formalment,

$$DBIL(F,G) = \sum_{r \in F} d(r^F, r^G)$$

$$d(r^1, r^2) = \sum_{V_i \in W} dist_{V_i}(V_i(r^1), V_i(r^2))$$

On, F i G són els fitxers original i protegit, *r* correspon a un registre d'un fitxer, W és el conjunt de variables.

S'ha de tenir en compte que la funció *dist* depèn de cada tipus de variable. Tot seguit es descriu com és aquesta funció per cada cas:

· Variables ordinals:

Per a variables ordinals, la distància es descriu com el numero de categories que hi ha entre els dos valors, dividit pel nombre de categories que pot prendre la variable. Mes formalment:

$$dist_{V_i}(a,b) = \frac{\left|\left\{i \middle| a \le i \le b\right\}\right|}{\left|D(V_i)\right|}$$

On, *a* i *b* són els valors a comparar, D(V) és el domini de la variable *i*, i $|\cdot|$ és l'operador de cardinalitat.

· Variables nominals:

Per a variables nominals, la distància entre valors serà 0 si els valors són iguals, o 1 si els valors són diferents. Mes formalment:

$$dist_{V_i}(a,b) = \begin{cases} 0, & a = b \\ 1, & a \neq b \end{cases}$$

On, *a* i *b* són els valors a comparar.

Posem un exemple per il·lustrar-ho millor.

Exemple 2.1

Variables:

- V1 : ordinal domini = [1, 2, 3, 4, 5]
- V2 : nominal domini = [04, 32, 50]
- V3 : nominal domini = [00, 10, 20, 30]
- V4 : ordinal domini = [1, 2, 3, 4, 5, 6, 7, 8]
- V5 : nominal domini = [1, 6, 8, 9]

*FITXER ORIGINAL (F)*

| | V1 | V2 | V3 | V4 | V5 |
|---|---|---|---|---|---|
| $r_1$ | 2 | 04 | 00 | 8 | 6 |
| $r_2$ | 1 | 50 | 20 | 5 | 6 |
| $r_3$ | 4 | 50 | 10 | 1 | 9 |
| $r_4$ | 2 | 04 | 20 | 2 | 1 |
| $r_5$ | 5 | 32 | 10 | 7 | 8 |

*FITXER PROTEGIT (G)*

| | V1 | V2 | V3 | V4 | V5 |
|---|---|---|---|---|---|
| $r'_1$ | 2 | 04 | 00 | 2 | 6 |
| $r'_2$ | 3 | 04 | 20 | 4 | 9 |
| $r'_3$ | 4 | 50 | 10 | 1 | 1 |
| $r'_4$ | 2 | 32 | 20 | 2 | 9 |
| $r'_5$ | 1 | 32 | 30 | 6 | 8 |

**Taula 2.1 – Fitxers de dades original (esquerra) i protegit (dreta)**

-Distàncies entre registres:

$d(r_1,r'_1) = 0 + 0 + 0 + 7/8 + 0 = 7/8$

$d(r_2,r'_2) = 3/5 + 1 + 0 + 2/8 + 1 = 57/20$

$d(r_3,r'_3) = 0 + 0 + 0 + 0 + 1 = 1$

$d(r_4,r'_4) = 0 + 1 + 0 + 0 + 1 = 2$

$d(r_5,r'_5) = 1 + 0 + 1 + 2/8 + 0 = 9/4$

- Pèrdua d'informació total:

**DBIL(F,G) =** $7/8 + 57/20 + 1 + 2 + 9/4 =$ **8,9750**

## *2.1.2 – CONTINGENCY TABLES-BASED INFORMATION LOSS (CTBIL)*

Una altra mesura seria la comparació de taules de contingència. Aquestes taules mostren la relació entre dues o més variables, indicant la freqüencia en què aparèix cada túpla de valors original i protegit.

L'estructura d'una taula de contingència de dues dimensions seria,

| | | V' | | | | |
|---|---|---|---|---|---|---|
| | | $s_1'$ | $s_2'$ | ... | $s_N'$ | |
| V | $s_1$ | $n_{1,1}$ | $n_{1,2}$ | ... | $n_{1,N}$ | $m_1$ |
| | $s_2$ | $n_{2,1}$ | $n_{2,2}$ | ... | $n_{2,N}$ | $m_2$ |
| | ... | ... | ... | ... | ... | ... |
| | $s_N$ | $n_{N,1}$ | $n_{N,2}$ | ... | $n_{N,N}$ | $m_N$ |
| | | $m_1'$ | $m_2'$ | ... | $m_N'$ | C |

**Figura 2.1 – Estructura d'una taula de contingència de dues dimensions**

On V és la variable original, V' és la variable protegida, $s_x$ són les categories que prèn la variable V, $s_x$' són les categories que pot prendre la variable V', $n_{i,j}$ són les freqüències en què aparèixen les tuples ($s_i$ , $s_j$'), $m_i$ és la freqüència absoluta de la categoria $s_i$, $m_i$' és la freqüència absoluta de la categoria $s_i$', i C és la suma de totes les freqüències absolutes.

Així per al càlcul d'aquesta mesura es necessita un paràmetre *K* el qual limita les dimensions que tindràn les taules de contingència creades. Així donats dos fitxers de dades F i G (original i protegit respectivament) i un subconjunt de variables a tenir en compte W, es generaràn tantes taules de contingència com combinacións de fins a *K* elements es puguin fer amb W.

Per tant podem definir la pèrdua d'informació basada en taules de contingència com la suma de les diferències dels valors cel·la a cel·la entre les taules dels dos fitxers corresponents a la mateixa combinació de variables.

Mes formalment sería,

$$CTBIL(F,G;W,K) = \sum_{\substack{\{V_{j1}\dots V_{jt}\}\subseteq W \\ |\{V_{j1}\dots V_{jt}\}|\le K}} \sum_{i_1\dots i_t} \left| x^F_{i_1\dots i_t} - x^G_{i_1\dots i_t} \right|$$

On, $x^{fitxer}_{indexs}$ és l'entrada de la taula *fitxer* a la posició donada pels *índexs*, i $|\cdot|$ és l'operador de valor absolut.

El resultat té com a desaventatge que el nombre de taules de contingència depèn del nombre de variables W, i la seva dimensió. A més la grandària de les taules depèn també del nombre de categories que té cada variable. Per tant, no es podria comparar un resultat obtingut amb 4 taules de contingència petites, amb un resultat obtingut amb 10 taules de contingència grans.

Així doncs és necessari algun tipus de normalització que ens determini un resultat indicant la diferència per cel·la que es té. Això s'aconsegueix dividint el resultat entre el nombre total de cel·les que conténen totes les taules utilitzades.

Més formalment,

$$ACTBIL(F,G;W,K) = \frac{CTBIL(F,G;W,K)}{\sum_{\substack{\{V_{j1}\dots V_{jt}\}\subseteq W \\ |\{V_{j1}\dots V_{jt}\}|\le K}} |D(V_{j1})| \quad \cdots \quad |D(V_{jt})|}$$

On, D(V) és el domini de la variable V, i $|\cdot|$ és l'operador de cardinalitat.

A aquest resultat normalitzat se l'anomena Average Contingency Table-Based Information Loss (ACTBIL).

Posem un exemple del càlcul d'aquesta mesura:

Exemple 2.2

Seguint amb l'exemple de la secció anterior tindriem les següents taules de contingència relacionant V2 i V5.

*FITXER ORIGINAL (F)*

| V2\V5 | **1** | **6** | **8** | **9** |
|---|---|---|---|---|
| 04 | 0 | 1 | 0 | 1 |
| 32 | 0 | 0 | 1 | 1 |
| 50 | 1 | 0 | 0 | 0 |

*FITXER PROTEGIT (G)*

| V2\V5 | **1** | **6** | **8** | **9** |
|---|---|---|---|---|
| 04 | 1 | 1 | 0 | 0 |
| 32 | 0 | 0 | 1 | 0 |
| 50 | 0 | 1 | 0 | 1 |

**Taula 2.2 – Fitxers de dades original (esquerra) i protegit (dreta)**

- Càlcul de la diferència entre taules amb W = {V2, V5}, i K = 2:

**CTBIL (F,G;W,K) =** 1 + 0 + 0 + 1 + 0 + 0 + 0 + 1 + 1 + 1 + 0 + 1 = **6**

- Càlcul de la mesura normalitzada:

**ACTBIL(F,G;W,K) =** 6 / 12 = **0,5**

### *2.1.3 – ENTROPY-BASED INFORMATION LOSS (EBIL)*

La mesura de pèrdua d'informació basada en l'entropia interpreta el mètode de protecció com un canal amb soroll per on es transmet la informació.

Seguint aquesta interpretació es pot calcular la pèrdua d'informació interpretant-ho com la quantitat de soroll o desordre que hi ha entre el fitxer original i el protegit.

Suposem que en el fitxer original tenim la variable V, i en el fitxer protegit tenim la variable V'. També suposem que $P_{V,V'}$= { p( V' = *i* | V = *j* ) } és la matriu de probabilitats condicionades. També tenim S que és el conjunt de categories de la variable V, llavors la pèrdua d'informació de V en el registre *r* seria:

$$H(V \mid V' = j) = -\sum_{i,j \in S} p(V = i \mid V' = j) \log p(V = i \mid V' = j)$$

On p( V = *i* | V' = *j* ) és la probabilitat condicional de què el valor original de V en el registre *r* sigui *i* si tenim que el valor de la variable protegida V' és *j*.

Aquestes probabilitats condicionades poden calcular-se mitjançant la següent expressió:

$$p(V = i = V' = j) = \frac{\left|\left\{ j \mid r_j = i \right\}\right| * p_{i,j}}{\sum_{k \in S}}$$

On $p_{i,j}$ és la probabilitat de que el valor original de V en el registre *r* sigui *i*. Això es pot traduïr com la freqüència en que apareix la tupla de valors (*i*, *j*) entre el fitxer original i el protegit, dividit entre el nombre total de registres del fitxer.

Més formalment,

$$p_{i,j} = \frac{\left|\left\{ j \mid r_j = i,\ r_j' = j \right\}\right|}{\left|\left\{ r_1, \ldots, r_n \right\}\right|}$$

On $r_x$ és el registre x-èssim del fitxer, i $|\cdot|$ és l'operador de cardinalitat.

Un cop tenim clar com es pot calcular la pèrdua d'informació en una variable dins un registre del fitxer, es pot extendre a tot el fitxer en general tan sols sumant tots els resultats de cada registre dins el fitxer. Això és:

$$EBIL(P_{V,V'}, G) = \sum_{r \in G} H(V \mid V' = j_r)$$

On G és el fitxer protegit.

Aquest càlcul només ens dóna el resultat per a una variable, així si volem saber la pèrdua total del fitxer hem de realitzar la operació per a cada variable i sumar tots els resultats.

Considerem a continuació un exemple.

Exemple 2.3

Continuant amb l'exemple dels mètodes de mesura anteriors, calcularem la pèrdua d'informació basada en l'entropía de la variable V2.

Les probabilitats $p_{i,j}$ de la matriu de Markov associada serien,

$p_{04,04} = 1/5$ $p_{04,32} = 1/5$ $p_{04,50} = 0$

$p_{32,04} = 0$ $p_{32,32} = 1/5$ $p_{32,50} = 0$

$p_{50,04} = 1/5$ $p_{50,32} = 0$ $p_{50,50} = 1/5$

Així la matriu de Markov associada a V2 queda de la següent manera,

| $p_{i,j}$ | 04 | 32 | 50 |
|---|---|---|---|
| 04 | 1/5 | 1/5 | 0 |
| 32 | 0 | 1/5 | 0 |
| 50 | 1/5 | 0 | 1/5 |

**Taula 2.3 – Matriu de Markov associada a V2**

Ara doncs podem calcular les probabilitats condicionades,

$$P(V2=04 \mid V2'=04) = (2\cdot 1/5) / (2\cdot 1/5 + 1\cdot 0 + 2\cdot 1/5) = 1/2$$
$$P(V2=04 \mid V2'=32) = (2\cdot 1/5) / (2\cdot 1/5 + 1\cdot 1/5 + 2\cdot 0) = 2/3$$
$$P(V2=04 \mid V2'=50) = (2\cdot 0) / (2\cdot 0 + 1\cdot 0 + 2\cdot 1/5) = 0$$
$$P(V2=32 \mid V2'=04) = (1\cdot 0) / (2\cdot 1/5 + 1\cdot 0 + 2\cdot 1/5) = 0$$
$$P(V2=32 \mid V2'=32) = (1\cdot 1/5) / (2\cdot 1/5 + 1\cdot 1/5 + 2\cdot 0) = 1/3$$
$$P(V2=32 \mid V2'=50) = (1\cdot 0) / (2\cdot 0 + 1\cdot 0 + 2\cdot 1/5) = 0$$
$$P(V2=50 \mid V2'=04) = (2\cdot 1/5) / (2\cdot 1/5 + 1\cdot 0 + 2\cdot 1/5) = 1/2$$
$$P(V2=50 \mid V2'=32) = (2\cdot 0) / (2\cdot 1/5 + 1\cdot 1/5 + 2\cdot 0) = 0$$
$$P(V2=50 \mid V2'=50) = (2\cdot 1/5) / (2\cdot 0 + 1\cdot 0 + 2\cdot 1/5) = 1$$

Formant una taula com la següent,

| $P_{V2,V2'}$ | 04 | 32 | 50 |
|---|---|---|---|
| **04** | ½ | 2/3 | 0 |
| **32** | 0 | 1/3 | 0 |
| **50** | ½ | 0 | 1 |

**Taula 2.4 – Matriu de probabiltats de canvi entre categories**

Per tant, les incerteses corresponents a cada categoria de la variable són,

$$H(V2 \mid V2'=04) = (-1/2 \cdot \log(1/2)) + 0 + (-1/2 \cdot \log(1/2)) = 0{,}6931$$

$$H(V2 \mid V2'=32) = (-2/3 \cdot \log(2/3)) + (-1/3 \cdot \log(1/3)) + 0 = 0{,}6365$$

$$H(V2 \mid V2'=50) = 0 + 0 + (-1 \cdot \log(1)) = 0$$

Així el càlcul de la pèrdua d'informació per a la variable V2 és,

$$\mathbf{EBIL(P_{V2,V2'}, G)} = H(V2 \mid V2'=04) + H(V2 \mid V2'=04) + H(V2 \mid V2'=04) + H(V2 \mid V2'=04) + H(V2 \mid V2'=04) = 0{,}6931 + 0{,}6931 + 0 + 0{,}6365 + 0{,}6365 = \mathbf{2{,}6592}$$

Acumulant els resultats d'aquest càlcul per cadascuna de les variables, obtenim la pèrdua d'informació de tot el fitxer.

### *2.1.4 – ALTERNATIVE INFORMATION LOSS MEASURE (IL)*

Finalment la última mesura per determinar la pèrdua d'informació que genera un mètode de protecció en un fitxer és una modificació del mètode basat en l'entropia.

Segons aquest mètode alternatiu el mètode basat en l'entropia té un inconvenient: la mesura de la pèrdua d'informació és una funció del fitxer protegit G però no depèn del fitxer original F.

Tenint en compte la relació de com més petita és la probabilitat condicional d'un canvi de categoría en una variable més gran és la pèrdua d'informació que pot generar, es pot definir la pèrdua d'informació com una funcío de tres elements: la probabilitat condicional, la categoria inicial *i* i la categoria protegida *L*.

Si utilitzem menys el logaritme de la probabilitat P( V = *i* | V' = *j* ), la resultant pèrdua d'informació conserva la relació descrita anteriorment de què si augmenta la probabilitat condicional, disminueix la pèrdua d'informació (i a l'inrevés).

Així la pèrdua d'informació per registre quan V=*i* i V'=*L* és:

$$PRIL(P_{V,V'},i,j) = -log\ P(V'=i \mid V=L)$$

Com que no cal calcular el PRIL per categories amb probabilitat de canvi 0 ja que no es produirà mai aquest canvi, es pot dir que el PRIL està ben definit.

Per acabar, la pèrdua d'informació per als fitxers sencers és:

$$IL(P_{V,V'},F,G) = \sum_{r \in G} PRIL(P_{V,V'},i_r,j_r)$$

On $i_r$ és el valor que pren V al registre *r* dins el fitxer original F, i $j_r$ és el valor que prèn V' al registre *r* dins el fitxer protegit G.

Exemple 2.4

Seguint amb l'exemple dels mètodes anteriors, recordem que teniem la següent matriu de probabilitats,

| $P_{V2,V2'}$ | **04** | **32** | **50** |
|---|---|---|---|
| **04** | ½ | 2/3 | 0 |
| **32** | 0 | 1/3 | 0 |
| **50** | ½ | 0 | 1 |

**Taula 2.5 – Matriu de probabilitats associades a V2**

La pèrdua d'informació de cada fila és,

$$PRIL_{r1}(P_{V2,V2'}, 04, 04) = -\log(1/2) = 0{,}6931$$
$$PRIL_{r2}(P_{V2,V2'}, 50, 04) = -\log(1/2) = 0{,}6931$$
$$PRIL_{r3}(P_{V2,V2'}, 50, 50) = -\log(1) = 0$$
$$PRIL_{r4}(P_{V2,V2'}, 04, 32) = -\log(2/3) = 0{,}4055$$
$$PRIL_{r5}(P_{V2,V2'}, 32, 32) = -\log(1/3) = 1{,}0986$$

I finalment la pèrdua d'informació total,

$$\mathbf{IL(P_{V2,V2'}, F, G)} = PRIL_{r1} + PRIL_{r2} + PRIL_{r3} + PRIL_{r4} + PRIL_{r5} =$$
$$0{,}6931 + 0{,}6931 + 0 + 0{,}4055 + 1{,}0986 = \mathbf{2{,}8903}$$

### *2.1.5 – PONDERACIÓ DE LES MESURES*

Per tal d'identificar la pèrdua d'informació general que té cada mètode s'ha realitzat una ponderació de les principals mesures intermitges. Com a tal s'han seleccionat tres mesures: la comparació entre taules de contingència normalitzada (ACTBIL), el mètode basat en distància (DBIL), i el mètode basat en l'entropia (EBIL).

El problema és que cap mètode d'avaluació dels descrits anteriorment ens dóna un resultat objectiu, ja que no es té un valor màxim acotat per a cada mesura.

No obstant per a la realització d'aquest projecte s'ha dut a terme una normalització dels resultats dividint les mesures entre els màxims valors de cada una obtinguts entre tots els resultats de les proves realitzades. D'aquesta manera tenim uns valors entre 0 i 1 (simulant un tant per cent) que ens mostra quin mètode té més pèrdua que un altre i ens permeten comparar-los de manera bastant fiable.

Un cop normalitzades les mesures, s'ha realitzat una ponderació equitativa per a totes tres mesures de la següent manera:

$$AIL(x) = \frac{ACTBIL'(x) + DBIL'(x) + EBIL'(x)}{3}$$

On ACTBIL'(x), DBIL'(x), i EBIL'(x), són les mesures normalitzades.

## *2.2 – AVALUADOR DE RISC DE REVELACIÓ*

Com ja s'ha comentat a l'introducció de la secció, el programa avaluador de risc de revelació no s'ha implementat expressament per a aquest projecte sinó que s'han afegit modificacions a un ja existent que aplica els mètodes Distance-Based Record Linkage (DBRL), Probabilistic Record Linkage (PRL), Interval Disclosure (ID), i Rank Swapping Record Linkage (RSRL).

Tot seguit es fa una breu descripció de cada mètode.

### *2.2.1 – DISTANCE-BASED RECORD LINKAGE (DBRL)*

El mètode basat en distàncies, mesura el perill de que una entrada sencera del fitxer pugui ser descoberta.

La idea bàsica d'aquest mètode és mesurar la distància entre el registre protegit i tots els registres originals disponibles, assignant la condició d'enllaçat al registre original que presenti menor distància.

Per tal de calcular la distància entre dos registres simplement s'ha d'anar calculant les distàncies valor a valor i acumulant els resultats.

Com ja s'ha fet servir en algun mètode de protecció en el capítol 1, s'han fet servir diferents maneres de calcular la distància entre un parell de valors (original, protegit) deperent de si es tracten de valors corresponents a una variable de tipus ordinal o a una de tipus nominal.

Per a variables de tipus ordinal tenim que la distància entre dues categories ($a$,$b$) és el nombre de categories que hi ha entre $a$ i $b$, dividit entre el nombre total de categories que la variable pot prendre. Més formalment sería:

$$d_{ORD}(a,b) = \frac{\left|\left\{i \mid a \le i \le b\right\}\right|}{\left|D(V_i)\right|}$$

Pel què fa a variables de tipus nominal, la distància entre dues categories ($a$,$b$) és 0 si són iguals o 1 si són diferents.

$$d_{NOM}(a,b) = \begin{cases} 0, & a = b \\ 1, & a \neq b \end{cases}$$

Un cop tenim un registre enllaçat, es fa la comprovació de si la correspondència és correcte utilitzant una numeració introduida als registres.

Així doncs, el resultat final serà el nombre de registres correctament enllaçats dividit entre el nombre total de registres que conté el fitxer, multiplicat per 100. Més formalment,

$$DBRL = \frac{reg\_linkats}{reg\_total} * 100$$

### *2.2.2 – PROBABILISTIC RECORD LINKAGE (PRL)*

El mètode Probabilistic Record Linkage intenta enllaçar cada registre protegit amb el seu corresponent original, a partir d'un index calculat per a cada parell de registres ($r_a$, $r_b$), classificant-los en tres grups: Linked Pairs (LP), Non-Linked Pairs (NP), i Clerical Pairs (CP) .

Per a cada parell de registres ($a$,$b$) es calcula un índex R($a$,$b$) a partir de les probabilitats:

P(coincidència|M) de coincidència de valors condicionada a la correspondència de registres
P(coincidència|U) de coincidència de valors condicionada a la no correspondència de registres.

Així els parells de registres es classifiquen a partir d'aquest índex i dos llindars (LlindarEnllaç i LlindarNoEnllaç) com segueix:

- Si R(a,b) > LlindarEnllaç , el parell es classifica com a LP
- Si R(a,b) < LlindarNoEnllaç , el parell es classifica com a NP
- En qualsevol altre cas el parell es classifica com a CP els quals necessiten ser revistats.

El resultat final serà el nombre de registres enllaçats correctament dividit entre el nombre total de registres.

### *2.2.3 – INTERVAL DISCLOSURE (ID)*

El mètode de Interval Disclosure calcula quants valors originals de la matriu de dades d'un fitxer protegit poden ser descoberts individualment.

La idea bàsica és la de crear uns conjunts del 1% i del 2% de la quantitat de categories que pot prendre la variable, i anar mirant quants dels valors protegits tenen el seu corresponent valor original que caigui entre aquests llindars. Tot seguit s'explica el mètode amb més detall.

⇒ Algorisme:

En la primera iteració tenim $P = 1$, i en la segona iteració incrementem el seu valor en 1 per a realitzar l'execució però amb $P = 2$

Per a cada element de la matriu de dades:

Agafem l'element original ($X_o$) i l'element protegit ($X_m$) corresponents a la mateixa posició.

Calculem la mida del conjunt de valors considerats correctes amb la següent fórmula:

Calculem el límit inferior de l'interval agafant l'índex de la categoria protegida dins la llista de categories i restant-li la mida calculada en el pas anterior.
Calculem el llindar superior de l'interval agafant l'índex de la categoria protegida dins la llista de categories i sumant-li la mida calculada.

Si l'índex de la categoria original està entre els dos llindars calculats, considerem el valor com a enllaçat.

El resultat d'una iteració serà el nombre de valors enllaçats dividit entre el nombre total d'elements en la matriu de dades.

El resultat final del mètode serà la suma dels resultats de totes les iteracions, dividit entre el nombre d'iteracions realitzades i multiplicat per 100.

### *2.2.4 – RANK SWAPPING RECORD LINKAGE (RSRL)*

L'enllaçament mitjançant Rank Swapping es tracta d'un mètode bastant nou el qual és molt efectiu per al mètode Rank Swapping, mètode de protecció el qual fins ara es resistia als atacs amb els mètode d'avaluació de risc descrits anteriorment.

Aquest mètode s'ha hagut d'implementar per a la realització del projecte ja que, com passava amb els mètodes de protecció de dades, només es disposava de la versió per a dades contínues, no categòriques.

El Rank Swapping estàndard intercanvia un valor original per un altre entre els *P* següents valors de la taula ordenada de categories. Per tant, per cada valor d'un atribut original $x_{ij}$ es pot calcular un conjunt $B(x_{ij})$ de '$2P$' registres protegits els quals poden ser el resultat de la transformació del registre original. Així, si es realitza el càlcul dels conjunts per a més atributs i es fa la seva intersecció, es va aconseguint reduïr el nombre de registres quedant-nos només amb els registres candidats comuns a tots els atributs utilitzats, que si agafem un nombre suficient d'atributs molt probablement arribarem a quedar-nos amb només un registre del qual podem estar 100% segurs que és correcte.

A la següent figura es pot observar un exemple il·lustratiu:

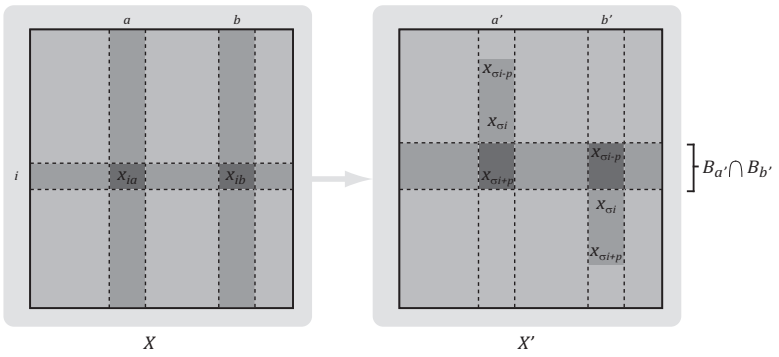



**Figura 2.2 – Exemple de l'efecte del mètode Rank Swapping Record Linkage (RSRL)**

A l'exemple de la figura 2.2 hi ha dos fitxers: X com a fitxer original i X' com a fitxer protegit. Volem enllaçar el registre *i* del fitxer X, i en coneixem dos atributs *a* i *b* amb valors $x_{ia}$ i $x_{ib}$.

Si sabem que el valor no es pot haver alterat mes de *p* categories, tenim que els registres candidats a ser enllaçats per l'atribut *a* són el conjunt $B_{a'}$ que conté els registres amb les següents categories corresponents a l'atribut *a'* :

$$\{x_{ia-p}, \cdots, x_{ia}, \cdots, x_{ia+p}\}$$

Aquest conjunt el representa la columna gris fosc de l'atribut *a'* del fitxer X'. Així veiem que és un conjunt que comprèn una important quantitat de registres, però si realitzem el mateix càlcul ara per l'atribut *b'* tenim un altre conjunt $B_{b'}$ que també comprèn bastants registres. Ara bé, si utilitzem els conjunts per fer la intersecció $B_{a'} \cap B_{b'}$, veiem que s'eliminen molts registres i ens quedem amb una petita porció que són els candidats comúns als dos conjunts, que formen un conjunt molt més reduït i proper a la solució.

### *2.2.5 – PONDERACIÓ DE LES MESURES*

Per tal d'identificar el risc de revelació general que té cada mètode s'ha realitzat una ponderació de totes les mesures intermitges.

En aquest cas no hi ha el problema que ens hem trobat amb les mesures de la pèrdua d'informació. La raó és que tota mesura té un màxim acotat el qual és el descobriment dels valors originals a tots els registres (en el cas dels mètodes de Record Linkage) o a tots els valors (en el cas del Interval Disclosure). Per tant els resultats ja es dónen sempre en un tant per cent equivalent a qualsevol conjunt de dades.

Així doncs, la ponderació utilitzada és la següent:

$$ADR = \frac{ID + \max(PRL, DBRL, RSRL)}{2}$$

D'aquesta manera sempre utilitzem el resultat més gran dels mètodes de Record Linkage i el resultat del Interval Disclosure equitativament.

# 3.- ANÀLISI DELS MÈTODES

En aquest capítol s'analitzaràn els mètodes de protecció presentats en el capítol 1 mitjançant els mètodes d'avaluació explicats en el capítol 2.

El conjunt de dades utilitzat per a aquests experiments és el U.S. Housing Survey del 1993 on hi ha informació sobre la mida i la composició de l'inventàri d'habitatges als Estats Units en aquell any. El fitxer consta doncs de 1000 registres amb 14 variables, on 11 de les quals són de tipus categòric.

Per a la realització dels experiments s'han protegit tres de les onze variables del fitxer de dades original les quals són les úniques de tipus ordinal i per tant podien ser protegides per tots els mètodes (recordem que tant el Bottom Coding com el Top Coding, com el Rank Swapping només es poden aplicar a variables ordinals). D'aquesta manera es pot obtenir una comparació mes equitativa que no hauriem aconseguit si per cada mètode haguessim protegit variables diferents.

A continuació es mostren les característiques de les variables utilitzades:

| Nom | Tipus | # Categories | Categories |
|---|---|---|---|
| BUILT | Ordinal | 25 | 01 02 03 04 05 06 07 08 09 80 81 82 83 84 85 86 87 88 89 90 91 92 93 99 -- |
| DEGREE | Ordinal | 8 | 1 2 3 4 5 6 9 - |
| GRADE1 | Ordinal | 21 | 00 01 10 11 12 02 21 22 23 24 25 26 03 04 05 06 07 08 09 99 -- |

**Taula 3.1 – Descripció de les variables a protegir**

## *3.1 – PÈRDUA D'INFORMACIÓ*

En aquest anàlisi els fitxers protegits mitjançant tots els mètodes de protecció seràn sotmesos a l'avaluador de pèrdua d'informació per determinar quina quantitat d'informació continguda en les dades originals es perd durant el procés de protecció.

### 3.1.1 – MICROAGREGACIÓ

En referència a la Microagregació s'han realitzat dos tipus d'experiments. Ja que el mètode depèn de dos paràmetres es volia veure com reacciona el mètode fixant el valor d'un paràmetre i variant el valor de l'altre, en ambdós casos. Per tant s'han obtingut dos resultats per a cada mesura.

En els experiment on s'ha fixat el valor del paràmetre *K* (nombre de registres dins un clúster) i s'ha variat *N* (nombre de variables a tenir en compte conjuntament) només es ténen tres resultats ja que *N* havia d'estar dins l'interval de valors [1,3] al treballar amb només tres variables. Per contra el paràmetre K agafa valors fins a 16 el qual ja es considera un valor molt gran.

Com a paràmetres de fixació en l'experiment de *K* fixada teniem que *K*=6, mentre que en l'experiment de *N* fixada teniem que *N* = 3.

A la figura 3.1 trobem els gràfics corresponents als resultats de pèrdua d'informació en els dos casos.

En el cas de la K fixada s'observa que l'augment de la pèrdua d'informació quan s'incrementa el paràmetre N és de tipus exponencial, per tant, per a cada increment del paràmentre hi ha un augment molt important de la pèrdua d'informació. Aquest doncs, és l'efecte esperat ja que com més variables agafem, tindrem túples més divserses i per tant el seu valor mig serà cada cop més distant de l'original.

Pel què fa al cas de la N fixada s'observa que el creixement és més suavitzat i de tipus logarítmitc, no obstant, per a valors alts del paràmetre K, el mètode acaba arribant a valors de pèrdua d'informació superiors al cas de K fixada. Aquest també és l'efecte esperat ja que com més registres per clúster s'agafin més possibilitats hi ha que hi hagi valors mes diferents, tot i que degut a la quantitat limitada de categories hi ha moltes repeticions de cadascuna.

Tot i això també cal remarcar que la Microagregació és un mètode que depèn bastant del conjunt de dades, ja que si es ténen moltes combinacions de categories semblants es produirà poca pèrdua d'informació, però si hi ha moltes combinacions diferents de categories hi haurà molta pèrdua d'informació.

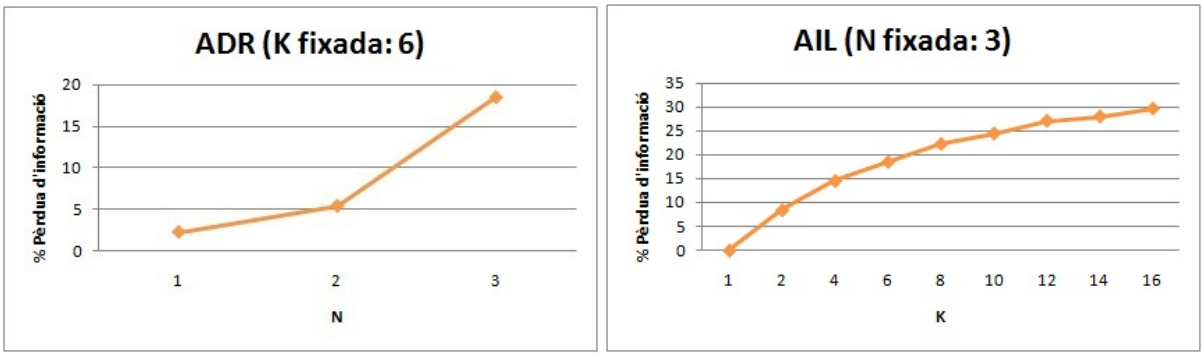



**Figura 3.1 – Gràfics amb els resultats de la pèrdua d'informació total mitjana per cada valor de *N* (esquèrra) i *K* (dreta)**

A l'annex I es poden observar els resultats parcials i totals (tant numèrics com gràfics) de tots els experiments realitzats per a aquest anàlisi.

### 3.1.2 – POST-RANDOMIZATION METHOD (PRAM)

Per al mètode de protecció PRAM també s'han realitzat diversos experiments per a cada valor del seu paràmetre *P* (cada matriu P). Això s'ha fet perquè el mètode conté una decisió aleatòria dins el procés de protecció i no sempre dóna el mateix resultat, és a dir, un mateix fitxer protegit diferents cops amb el mateix valor de *P* té resultats lleugerament diferents. Així, com a valor representatiu de cada mesura per a cada *P*, s'ha agafat la mitjana dels valors obtinguts en tots els seus experiments.

Com es pot observar a la figura 3.2, el mètode PRAM no genera una excessiva pèrdua d'informació (té un màxim de un 11% en el valor màxim admès per al seu paràmetre *P*). Això és degút als poc canvis que aquest mètode genera dins el fitxer de dades.

Recordant l'estructura de la taula de canvis PRAM de l'exemple 1.2 podem observar que les probabilitats corresponents a conservar la categoria (els de la diagonal de la matriu) són probabilitats molt altes, així molt poques categories són substituïdes.

També podem observar com a mesura que es va incrementant el valor de *P* el gràfic experimenta un creixement bastant lineal, el qual no es el tipus de creixement més desitjat però al treballar amb pèrdues d'informació relativament petites, no és tant problema.


**Figura 3.2 – Gràfic amb els resultats de la pèrdua d'informació total mitjana per cada valor de *P***

A l'annex II es poden observar els resultats parcials i totals (tant numèrics com gràfics) de tots els experiments realitzats per a aquest anàlisi.

### 3.1.3 – RANK SWAPPING

En el càs del mètode de protecció Rank Swapping s'han realitzat diversos experiments per a cada valor del seu paràmetre *P*. Això s'ha fet perquè, com passava amb el mètode PRAM, aquest mètode conté una decisió aleatòria dins el procés de protecció i no sempre dóna el mateix resultat, és a dir, un mateix fitxer protegit diferents cops amb el mateix valor de *P* té resultats lleugerament diferents. Així, com a valor representatiu de cada mesura per a cada *P*, s'ha agafat la mitjana dels valors obtinguts en tots els seus experiments.

La figura 3.3 ens mostra que aquest mètode té un comportament bastant lineal respecte la variació del valor del seu paràmetre *P* partint d'una pèrdua d'informació inferior al 10% en *P*=1, fins arribar a un 45% en *P*=20.

Aquest mètode ens permet arribar fins a valors alts del seu paràmetre sense perdre excessiva informació (a *P*=8 encara estem per sota del 30%) a causa de que les dades si que canvien però sempre són les mateixes en posicions diferents, per tant la informació estadística com el median, la mitjana o la desviació, es mantenen.

Tot i que en general és un mètode que genera bastant pèrdua d'informació sobretot per a valors de *P* més grans de 8, a l'anàlisi conjunt combinant la pèrdua d'informació i el risc de revelació de l'apartat 3.3 veurem que el Rank Swapping es convertirà en un dels millors mètodes de protecció.

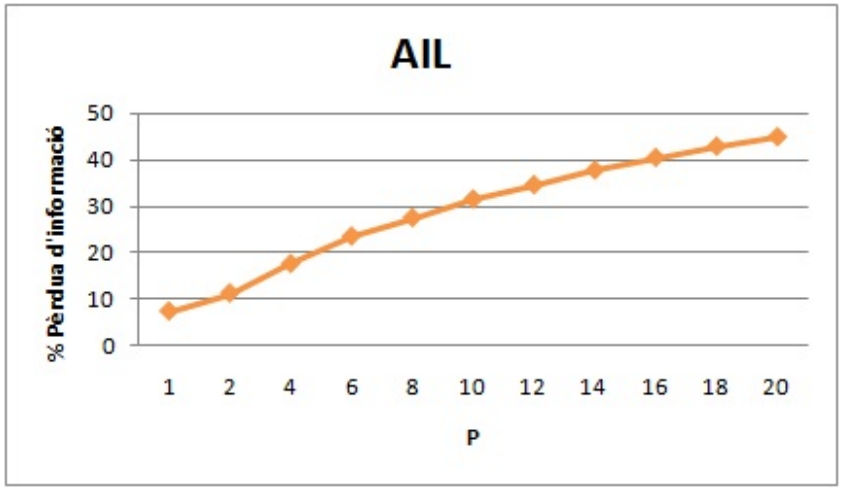



**Figura 3.3 – Gràfic amb els resultats de la pèrdua d'informació total mitjana per cada valor de *P***

A l'annex III es poden observar els resultats parcials i totals (tant numèrics com gràfics) de tots els experiments realitzats per a aquest anàlisi.

### 3.1.4 – GLOBAL RECODING

Per a l'avaluació de la pèrdua d'informació del mètode Global Recoding s'han realitzat experiments per a valors de *P* entre 1 i 6. No s'ha arribat al valor *P*=9, que és el límit permès en aquest mètode, perquè entre les variables protegides n'hi ha una la qual té només 8 categories d'on dues són "reservades" per a altres mètodes quedant-ne així 6 de disponibles. Per tant com que el paràmetre *P* indica el nombre de categories menys freqüents a codificar, per tal de fer un estudi equitatiu per a totes les variables, com a màxim es poden agafar 6 categories.

Respecte al gràfic de la figura 3.4 podem observar un creixement de tipus lleugerament exponencial causat per culpa del criteri de selecció de les categories a protegir que té aquest mètode. Recordem que sempre es protegeixen les *P* categories menys freqüents de cada variable, això vol dir que cada nova categoria que s'agafa per protegir tindrà com a mínim tantes aparicions a les dades com la última categoria protegida, però probablement en tindrà més. Així tenim un increment de categories protegides que no és lineal, sinó superior al lineal.

En quant als percentatges de pèrdua d'informació aquest mètode té unes pèrdues baixes per a valors inferiors a 4 però arriba a pèrdues ja importants si es va augmentant aquest valor.


**Figura 3.4 – Gràfic amb els resultats de la pèrdua d'informació total mitjana per cada valor de *P***

A l'annex IV es poden observar els resultats parcials i totals (tant numèrics com gràfics) de tots els experiments realitzats per a aquest anàlisi.

### 3.1.5 – TOP CODING

Per a l'avaluació de la pèrdua d'informació del mètode Top Coding s'han realitzat experiments per a valors de *P* entre 1 i 6. No s'ha arribat al valor *P*=9, que és el límit permès en aquest mètode, perquè entre les variables protegides n'hi ha una la qual té només 8 categories d'on dues són "reservades" per a altres mètodes quedant-ne així 6 de disponibles. Per tant com que s'han de codificar les *P* primeres categories, com a màxim es poden agafar 6 categories.

El comportament d'aquest mètode depèn molt de cada fitxer a protegir ja que només té en compte la protecció de les *P* primeres categories que apareixen al fitxer descriptor de variables. Per tant la quantitat d'entrades que es protegiràn al fitxer dependrà de la freqüència que tingui cada categoria dins d'aquest, cosa que varia a cada fitxer.

La figura 3.5 mostra els resultats obtinguts per al fitxer utilitzat en aquest projecte. Es pot observar que el creixement no segueix un patró uniforme degut a que cada categoria nova té una freqüència diferent i per tant la pèrdua d'informació augmenta més bruscament (càs de l'interval [4,5]) o més lentament que en l'anterior pas (càs de l'interval [5,6]).


**Figura 3.5 – Gràfic amb els resultats de la pèrdua d'informació total mitjana per cada valor de *P***

A l'annex V es poden observar els resultats parcials i totals (tant numèrics com gràfics) de tots els experiments realitzats per a aquest anàlisi.

### 3.1.6 – BOTTOM CODING

Per a l'avaluació de la pèrdua d'informació del mètode Bottom Coding s'han realitzat experiments per a valors de *P* entre 1 i 6. No s'ha arribat al valor *P*=9,

que és el límit permès en aquest mètode, perquè entre les variables protegides n'hi ha una la qual té només 8 categories d'on dues són "reservades" per a altres mètodes quedant-ne així 6 de disponibles. Per tant com que s'han de codificar les *P* últimes categories, com a màxim es poden agafar 6 categories.

El comportament d'aquest mètode (com en el mètode Top Coding) depèn molt de cada fitxer a protegir ja que només té en compte la protecció de les *P* últimes categories que apareixen al fitxer descriptor de variables. Per tant la quantitat d'entrades que es protegiràn al fitxer dependrà de la freqüència que tingui cada categoria dins d'aquest, cosa que varia a cada fitxer.

No obstant el fitxer utilitzat per a la realització d'experiments en aquest projecte, ha resultat tenir una freqüència molt similar en les *P* últimes categories, cosa que ha generat el gràfic de la figura 3.6 on s'observa un augment de la pèrdua d'informació bastant lineal. Així doncs tenim que el mètode tolera bastant bé la pèrdua d'informació per als primers tres valors de *P*, però genera massa pèrdua per als següents valors.


**Figura 3.6 – Gràfic amb els resultats de la pèrdua d'informació total mitjana per cada valor de *P***

A l'annex VI es poden observar els resultats parcials i totals (tant numèrics com gràfics) de tots els experiments realitzats per a aquest anàlisi.

## *3.2 – RISC DE REVELACIÓ*

En aquest anàlisi els fitxers protegits mitjançant tots els mètodes de protecció seràn sotmesos a l'avaluador de risc de revelació per determinar quina quantitat de dades originals poden ser obtingudes de cada fitxer protegit.

### 3.2.1 – MICROAGREGACIÓ

Com passa a l'anàlisi de la pèrdua d'informació, per a la Microagregació s'han realitzat dos tipus d'experiments. Ja que el mètode depèn de dos paràmetres es volia veure com reacciona el mètode fixant el valor d'un paràmetre i variant el valor de l'altre, en ambdós casos. Per tant s'han obtingut dos resultats per a cada mesura.

En els experiment on s'utilitza un valor fixe del paràmetre *K* (nombre de registres dins un clúster) i es varia el del paràmetre *N* (nombre de variables a tenir en compte conjuntament) només es ténen tres resultats ja que, al treballar amb només tres variables, *N* havia d'estar dins l'interval de valors [1,3]. Per contra, en el càs de l'experiment amb el paràmetre N fixat, el paràmetre *K* agafa valors fins a 16 el qual ja es considera un valor molt gran. A la figura 3.7 trobem els gràfics corresponents als resultats del risc de revelació en els dos casos.

En el cas de la *K* fixada s'observa que el decrement del risc de revelació quan s'incrementa el paràmetre *N* és de tipus exponencial, per tant, per a cada increment del paràmentre hi ha un decrement molt important del risc de revelació. Aquest doncs, és l'efecte esperat ja que com més variables agafem, tindrem túples més divserses i per tant el seu valor mig serà cada cop més distant de l'original.

Pel què fa al cas de la *N* fixada s'observa que el decrement és més suavitzat i de tipus logarítmitc, així com més augmentem el valor del paràmetre *K* més petit és el decrement que experimenta el risc de revelació. Aquest també és l'efecte esperat ja que com més registres per clúster s'agafin més possibilitats hi ha que hi hagi valors mes diferents, tot i que degut a la quantitat limitada de categories hi ha moltes repeticions de cadascuna.

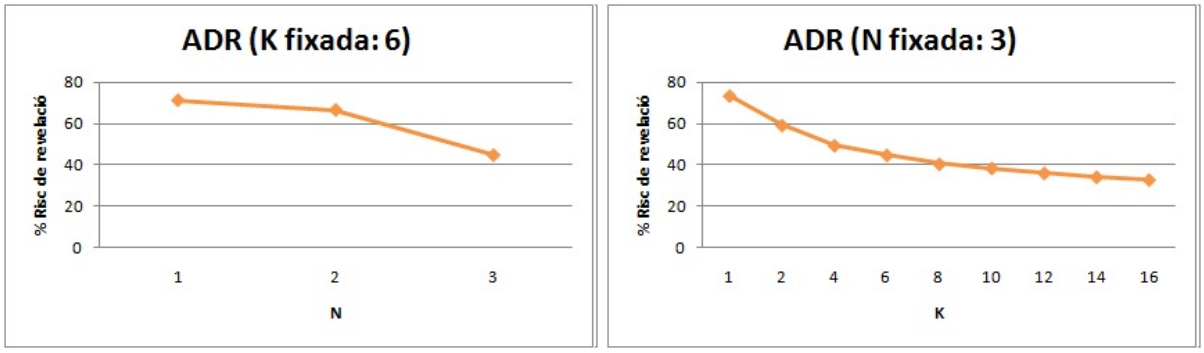



**Figura 3.7 – Gràfics amb els resultats del risc de revelació total mitjà per a cada valor de *N* (esquèrra) i *K* (dreta)**

Tot i això també cal remarcar que la Microagregació és un mètode que depèn bastant del conjunt de dades, ja que si es ténen moltes combinacions de

categories semblants es produirà molt risc de revelació, però si hi ha moltes combinacions diferents de categories hi haurà poc risc de revelació.

A l'annex 1 es poden observar els resultats parcials i totals (tant numèrics com gràfics) de tots els experiments realitzats per a aquest anàlisi.

### 3.2.2 – POST-RANDOMIZATION METHOD (PRAM)

Com en el càs de la pèrdua d'informació, aquí també s'han realitzat diversos experiments per a cada valor del seu paràmetre *P* degut a què el mètode conté una decisió aleatòria dins el procés de protecció i no sempre dóna el mateix resultat, per tant, un mateix fitxer protegit diferents cops amb el mateix valor de *P* obté resultats lleugerament diferents. Així, com a valor representatiu de cada mesura per a cada *P*, s'ha agafat la mitjana dels valors obtinguts en tots els seus experiments.

Tal com s'ha comentat en l'anàlisi d'aquest mètode corresponent a la pèrdua d'informació, el punt feble d'aquest mètode és que no modifica gaire les dades durant el procés de protecció. Per aquesta raó es pot observar a la figura 3.8 que, tot i tenir un decreixement bastant lineal, aquest mètode té un alt risc de revelació ja que té un màxim d'un 73% i un mínim d'un 63% aproximadament degut als pocs canvis introduïts a les dades originals.

Per tant aquest mètode no funciona garie bé a l'hora de garantir la protecció de les dades perquè una gran quantitat de dades originals poden ser descobertes.

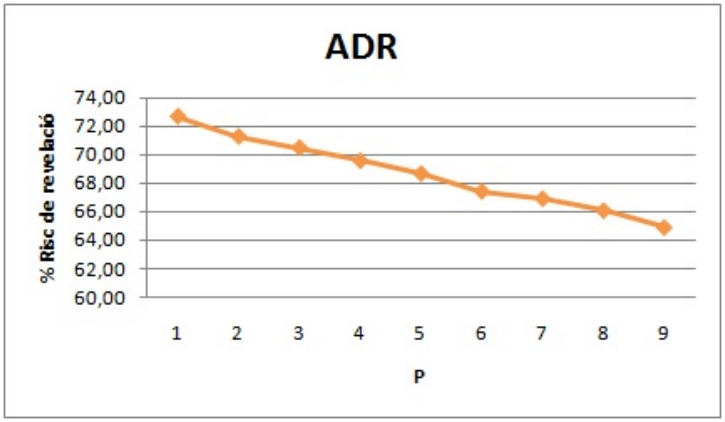



**Figura 3.8 – Gràfic amb els resultats del risc de revelació total mitjà per a cada valor de *P***

A l'annex 2 es poden observar els resultats parcials i totals (tant numèrics com gràfics) de tots els experiments realitzats per a aquest anàlisi.

### 3.2.3 – RANK SWAPPING

Respecte al mètode de protecció Rank Swapping també s'han realitzat diversos experiments per a cada valor del seu paràmetre *P* com en els experiments de la pèrdua d'informació. Els motius són els mateixos, és a dir, el mètode conté una decisió aleatòria dins el procés de protecció i no sempre dóna el mateix resultat, és a dir, un mateix fitxer protegit diferents cops amb el mateix valor de *P* té resultats lleugerament diferents. Així, com a valor representatiu de cada mesura per a cada *P*, s'ha agafat la mitjana dels valors obtinguts en tots els seus experiments.

Com es pot observar a la figura 3.9, en aquest anàlisi ens trobem amb un decreixement molt pronunciat al principi fins a P=8 i amb un decreixement més suavitzat fins a P=20. Això és degut a que cada cop els valors poden intercanviar-se amb valors mes llunyans i per tant tot queda més desordenat disminuint ràpidament el risc de revelació dels valors originals, però com més s'apropa al valor limit que pot aconseguir, el decreixement es va fent cada vegada més lent.

Tal com es comenta en l'anàlisi de la pèrdua d'informació, es podrà comprovar a l'anàlisi conjunt de pèrdua d'informació i risc de revelació de l'apartat 3.3 que aquest mètode és un dels millors gràcies a la gran disminució del risc de revelació i el lent augment de la pèrdua d'informació.

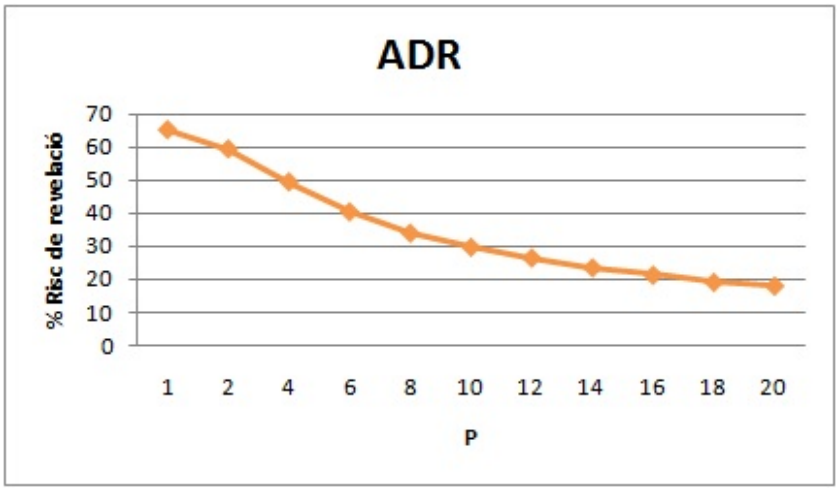



**Figura 3.9 – Gràfic amb els resultats del risc de revelació total mitjà per a cada valor de *P***

A l'annex 3 es poden observar els resultats parcials i totals (tant numèrics com gràfics) de tots els experiments realitzats per a aquest anàlisi.

### 3.2.4 – GLOBAL RECODING

Com en el càs de la pèrdua d'informació corresponent al Global Recoding, aquí també s'han realitzat experiments per a valors de *P* entre 1 i 6. Tampoc s'ha arribat al valor *P*=9, que és el límit permès en aquest mètode, perquè entre les variables protegides n'hi ha una la qual té només 8 categories d'on dues són "reservades" per a altres mètodes quedant-ne així 6 de disponibles. Per tant com que el paràmetre *P* indica el nombre de categories menys freqüents a codificar, per tal de fer un estudi equitatiu per a totes les variables, com a màxim s'han agafat 6 categories.

La figura 3.10 mostra el gràfic corresponent als resultats de l'anàlisi del risc de revelació. Es pot observar que el gràfic té un decreixement lleugerament exponencial. Això és degut, com en el cas de la pèrdua d'informació, a que cada vegada que s'augmenta el valor del paràmetre, s'està protegint una categoria que tindrà com a mínim la mateixa freqüència que l'anterior (cosa que ens donaria un comportament lineal) però probablement tindrà una freqüència superior, per tant es protegiran més categories que en el pas anterior i així donarà el comportament lleugerament exponencial que hem experimentat.

No obstant els valors de risc de revelació en els que ens movem són bastant alts i van des d'un 70% fins a un 40%, valors massa elevats per tal de garantir la seguretat de les dades. Això passa perquè aquest mètode protegeix poques categories al triar les *P* categories amb menys freqüència.

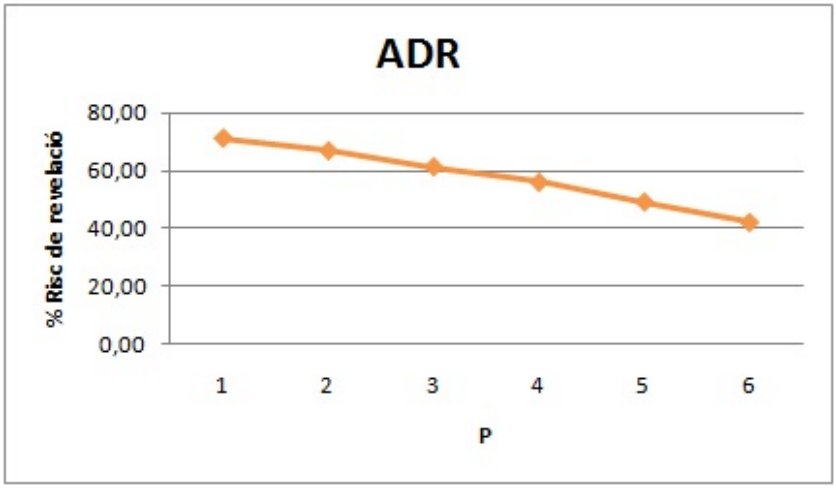



**Figura 3.10 – Gràfic amb els resultats del risc de revelació total mitjà per a cada valor de *P***

A l'annex 4 es poden observar els resultats parcials i totals (tant numèrics com gràfics) de tots els experiments realitzats per a aquest anàlisi.

### 3.2.5 – TOP CODING

En el càs del mètode Top Coding també s'han realitzat experiments per a valors de *P* entre 1 i 6. Aquí tampoc s'ha arribat al valor *P*=9, que és el límit permès en aquest mètode, perquè entre les variables protegides n'hi ha una la qual té només 8 categories d'on dues són "reservades" per a altres mètodes quedant-ne així 6 de disponibles. Per tant com que s'han de codificar les *P* primeres categories, com a màxim es poden agafar 6 categories.

Com ja s'ha comentat a l'anàlisi de la pèrdua d'informació, el comportament d'aquest mètode varia en cada fitxer de dades diferent depenent de les freqüències en què apareixen les categories.

A la figura 3.11 hi ha els resultats corresponents al fitxer de dades utilitzat en aquest projecte on aquest fet s'observa amb canvis lleus en el percentatge de risc de revelació quan es protegeixen categories amb poca freqüència (càs de l'interval [5,6]) i canvis bruscs quan es s'afegeix una categoria amb gran freqüència (càs de l'interval [4,5]).

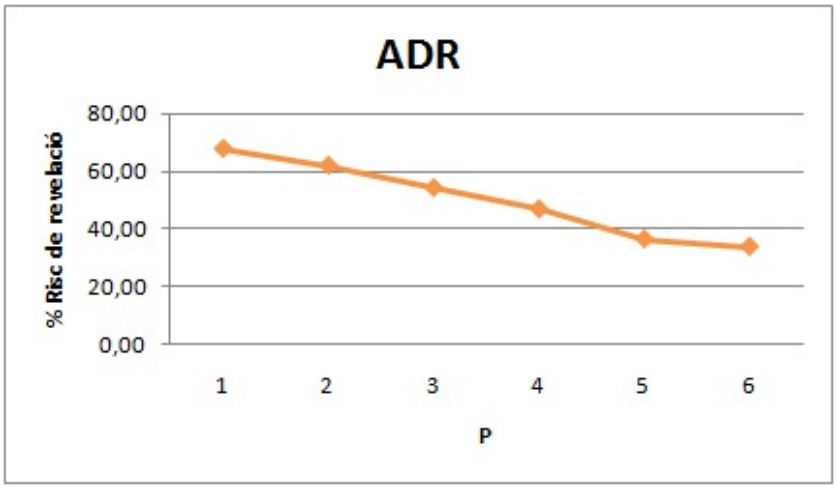



**Figura 3.11 – Gràfic amb els resultats del risc de revelació total mitjà per a cada valor de *P***

A l'annex 5 es poden observar els resultats parcials i totals (tant numèrics com gràfics) de tots els experiments realitzats per a aquest anàlisi.

### 3.2.6 – BOTTOM CODING

Finalment, com en el càs de la pèrdua d'informació, per al Bottom Coding s'han realitzat experiments per a valors de *P* entre 1 i 6. Com ja s'ha explicat, no s'ha arribat al valor *P*=9, que és límit permès en aquest mètode, perquè entre les

variables protegides n'hi ha una la qual té només 8 categories d'on dues són "reservades" per a altres mètodes quedant-ne així 6 de disponibles. Per tant com que s'han de codificar les *P* últimes categories, com a màxim es poden agafar 6 categories.

Igual que en el càs del mètode de Top Coding, el comportament d'aquest mètode varia en cada fitxer de dades diferent depenent de les freqüències en què apareixen les categories.

No obstant en aquest càs ha resultat que les *P* últimes categoríes tenien una freqüència molt similar, cosa que ha donat com a resultat el gràfic de la figura 3.12 on es veu un decreixement del risc de revelació bastant lineal el qual s'inicia en alts valors prop del 70% de risc en *P*=1, fins a valors no gaire baixos de fins a un 35% de risc en *P*=6.

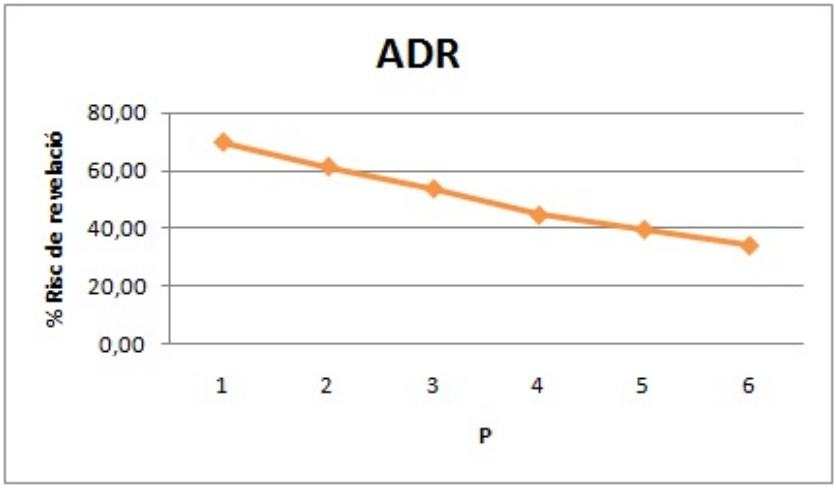



**Figura 3.12 – Gràfic amb els resultats del risc de revelació total mitjà per a cada valor de *P***

A l'annex 6 es poden observar els resultats parcials i totals (tant numèrics com gràfics) de tots els experiments realitzats per a aquest anàlisi.

## *3.3 – RESULTATS CONJUNTS*

Com s'ha pogut observar en els anàlisis individuals, no tots els mètodes ténen la mateixa quantitat d'efecte sobre les dades però si que hi ha un patró lògic clar, el comportaments de la pèrdua d'informació i del risc de revelació están inversament relacionats quan es varia el valor del paràmetre de cada mètode. Així la pèrdua d'informació sempre tenedeix a pujar quan s'augmenta el valor del paràmetre, mentre que el risc de revelació sempre tendeix a baixar.

D'aquesta manera ens cal veure quins mètodes combinen millor aquests dos valors per tal de tenir un equilibri raonable entre tots dos.

Així en aquesta secció es fa la comparació dels mètodes utilitzant el seu score total en cada valor dels mètodes fins a 6, ja que els superiors a 6 donaven ja bastant mal resultats. Aquest score ha estat calculat com la mitjana dels valors de la pèrdua d'informació i el risc de revelació, és a dir, amb la següent fórmula:

$$Score(x) = \frac{IL(x) + DR(x)}{2}$$

On $x$ és el valor del paràmetre, IL(x) és la funció avaluadora de la pèrdua d'informació i DR(x) és la funció avaluadora del risc de revelació.
A la figura 3.13 es mostra el gràfic correponent als valors de score que prèn cada mètode per a cada valor del seu corresponent paràmetre.

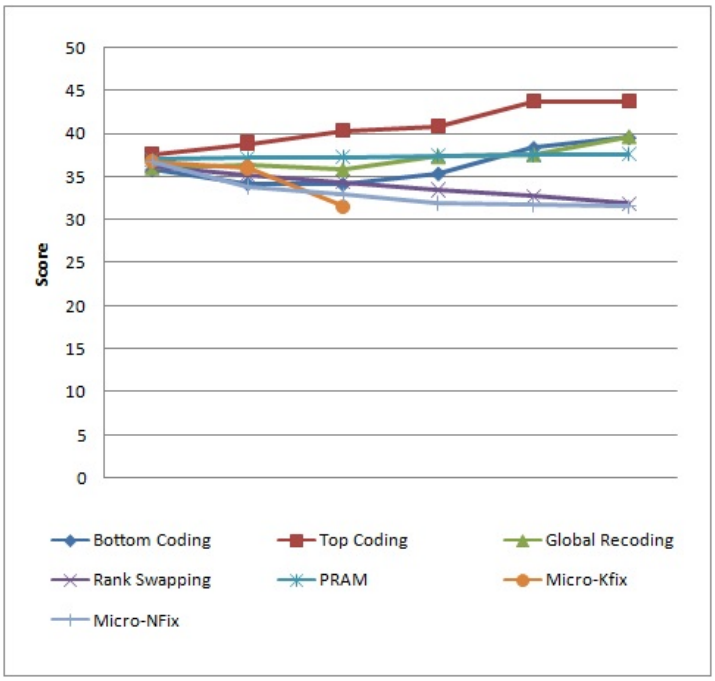



**Figura 3.13 – Gràfic amb els resultats del score per a cada mètode i els seus paràmetres**

Es pot observar que el mètode amb millor score és la Microagregació. A més en la majoria de tots els possibles valors que pot prendre el seu paràmetre, no hi ha gairebé cap altre mètode que el superi.

També es veu que l'altre mètode que millor suporta la variació del seu paràmetre i, per tant, millor combina l'efecte de la pèrdua d'informació i l'efecte del risc de revelació és el Rank Swapping ja que en tot el seu gràfic hi ha una variació màxima de nomes 5 punts aconseguint un molt bon score.

Per altra banda tenim que el Top Coding ha donat scores molt dolents i és el pitjor en score en tots els valors del paràmetre. Aquest fet és destacable en aquest cas però no es pot tenir en compte en altres casos ja que és un mètode que depèn molt de les dades i les freqüències de les categories. En aquest càs s'ha trobat que les P primeres categories tenien molta freqüència o molt poca i per tant el score s'ha disparat, no obstant en casos on la freqüència fos diferent podriem obtenir resultats molt bons amb aquest mètode.

Per últim, cal comentar també els resultats del mètode PRAM. Segons el gràfic es pot observar que és un mètode que també conserva bastant bé el seu score ja que gairebé no té variació en el seu score a mesura que augmentem el valor del paràmetre. No obstant, l'inconvenient que té és que durant tots els valors del seu paràmetre el score és gairebé sempre el pitjor, només superat pel Top Coding i pels pitjors scores del Bottom Coding i Global Recoding. Aquest fet, fa que el PRAM sigui un mètode poc útil per a la protecció de dades.

Per últim tenim el gràfic de la figura 3.14 on tenim el score desglossat per a cada mètode i els seus paràmetres.

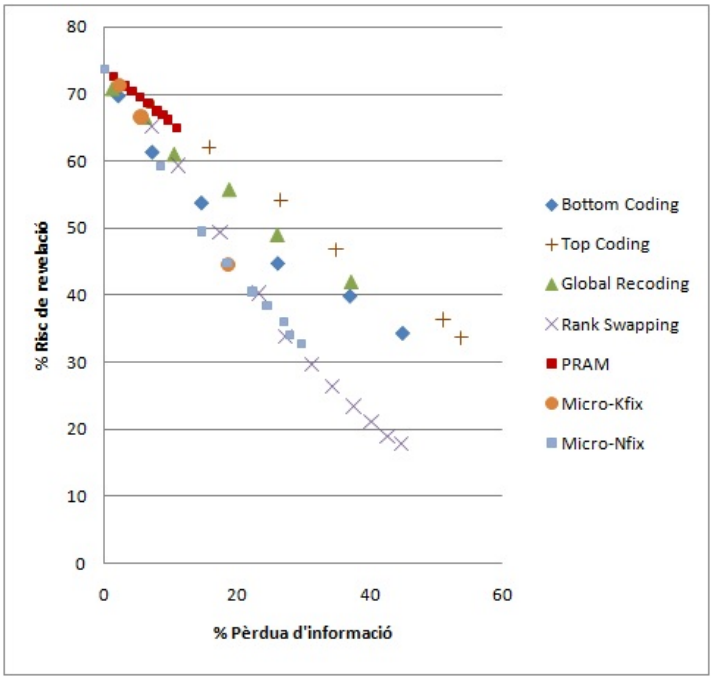



**Figura 3.14 – Gràfic de dispersió amb els score desglossats per a cada mètode i els seus paràmetres**

El punt òptim (però impossible) sería la cantonada inferior esquerra del gràfic corresponent al punt (0,0) on voldria dir que no es perd gens d'informació i no s'en revela gens tampoc.

Tenint com a referència aquest punt es veu que els punts més propers corresponen als mètodes de protecció Rank Swapping i Microagregació.

Pel què fa al PRAM veiem que tots els seus punts estàn molt allunyats i per tant podriem dir que és un dels mètodes que pitjor funciona. Per aquesta raó en aquest projecte s'ha dissenyat també un mètode optimitzador per al PRAM. Aquest mètode s'explicarà en el següent capítol.

# 4.- OPTIMITZACIÓ DESENVOLUPADA

Tal com s'ha vist en el capítol anterior, el mètode de protecció PRAM és el que donava resultats, en general, més dolents. Per tal de millorar aquests resultats, en la realització d'aquest projecte s'ha desenvolupat un mètode evolutiu el qual té la funció d'optimitzar la matriu de probabilitats d'intercanvi de categories del PRAM per tal que al fer-la servir per a la protecció d'una variable en un arxiu dóni el millor valor possible tant en pèrdua d'informació com en risc de revelació.

## *4.1 – DESCRIPCIÓ DEL MÈTODE D'OPTIMITZACIÓ*

El mètode d'optimització desenvolupat es basa en un algorisme genètic.

Els algorismes genètics són mètodes adaptatius que poden utilitzar-se per a resoldre problemes de recerca i optimització. Aquests algorismes estàn basats en el procés genètic dels organismes vius. Al llarg de les generacions, les poblacions evolucionen a la natura d'acord amb els principis de selecció natural i la supervivència dels més forts, postulats per Darwin. Per imitació d'aquest procés, els algorismes genètics són capaços d'anar creant solucions per a problemes del món real. L'evolució d'aquestes solucions cap a valors òptims del problema depèn en bona mesura d'una adequada codificació de les mateixes.

```
Entrada: P(0) = X  matriu de probabilitats inicial
Sortida: P(t) = X'  matriu de probabilitats final

t=0;
inicialitzar(P);

MENTRE no es s'aturi el programa FER
        alter =  valor aleatòri 0 o 1 per triar mutació o creuament

        SI alter == 0 LLAVORS
                Y = mutar(X);
        SINÓ
                Y = creuament(X);
        FI SI;

        avaluar(Y,X);
        t = t+1;
FI MENTRE;

RETORNA P(t);
```

**Algorisme 4.1 – Esquelet de l'algorisme genètic desenvolupat**

Els operadors bàsics del algorismes genètics són la mutació i el creuament, i són els passos a dissenyar més importants ja que si estàn bén dissenyats poden generar més bones solucións i més ràpidament. També es necessita una funció d'avaluació per comprovar si les solucions trobades després de cada mutació i creuament són millors o no que les originals.

Tot seguit es descriu l'algorisme genètic desenvolupat per a aquest mètode optimitzador, així com els seus operadors.

La població inicial de l'algorisme desenvolupat constarà d'un sol individu el qual s'anirà substituïnt per solucions obtingudes amb millor adaptació, és a dir, amb millor resultat de la funció d'avaluació. Aquest individu serà la matriu de probabilitats d'intercanvi utilitzada pel mètode PRAM a l'hora de protegir una variable d'un fitxer.

La representació utilitzada per als valors de la matriu ha estat en la conversió a codis Gray dels valors de les parts enteres dels valors decimals multiplicats per 100. Així treballarem amb valors enters entre 0 i 1000. Els codis Gray ténen la gran característica de què cada valor només difereix d'un sol bit amb el seu successor i el seu antecessor. Això s'ha utilitzat per tal de suavitzar el comportament de la mutació, que com s'explica més endavant, tracta amb l'alteració de bits de valors, i prevenir canvis molt bruscs de valor. Tot seguit es mostra un exemple de la codificació dels valors de la matriu.

Exemple 4.1

Aquest exemple mostra un exemple de codificació d'un valor decimal, que podria correspondre a un valor de la matriu, en una representació en codi Gray.

Valor original de la matriu = *0,25*
Valor original de la matriu multiplicat per 100 = *250,0*
Part sencera del valor original de la matriu multiplicat per 100 = *250*
Codificació en binari (10 bits) = *0011111010*
Transformació a codi Gray = *0010000111*

La funció encarregada de fer una valoració de com de bona (o adaptada) és una solució serà la funció d'avaluació. Aquesta funció restaura els valors de la matriu a nombres decimals dins el rang [0,1], ja que són probabilitats, utilitza la matriu restaurada per a realitzar una protecció del fitxer de dades original utilitzant el mètode PRAM. Tot seguit el fitxer protegit obtingut es crida desde el programa avaluador de pèrdua d'informació i el programa avaluador de risc de revelació i s'avalua. La mitjana dels dos valors obtinguts per part dels programes avaluadors serà l'indicador de l'adaptació de la matriu solució.

Així, per comparar una matriu solució i la consequent matriu solució trobada després d'una mutació o un creuament, simplement cal comparar els valors d'adaptació de les dues matrius i si la nova matriu té un valor més baix substituirà a l'altra com a matriu solució de la generació *t*, sinó es descartarà i tot continuarà com estava.

Pel que fa a la funció de mutació el què es fa és simplement triar un valor de la matriu en codi Gray de forma aleatòria, triar un bit del valor també de forma aleatòria, i invertir-lo, és a dir, si el bit és 1 passarà a ser 0, i viceversa. Així substituïnt el valor a la matriu, tenim un nou individu a avaluar.

Per últim, la funció de creuament agafa aleatòriament dos punts de la matriu i escull també de forma aleatòria un nombre d'elements a ser substituïts. Així partint dels dos punts escollits es seleccionen tants nombres com indiqui el nombre trobat obtenint dos conjunts de valors. Així s'intercanvien els valors dels conjunts i s'obté un altre nou individu a avaluar.

## *4.2 – RESULTATS OBTINGUTS*

S'ha realitzat dos experiments protegint una variable diferent a cada un i comparant els valors obtinguts abans i després d'aplicar-hi el mètode optimitzador. S'han triat variables amb diferent nombre de categories ja que a l'hora de protegir pot influir tenir moltes opcions per substituir una categoria o tenir-ne poques. Cal dir que els resultats de pèrdua d'informació i risc de revelació diferèixen dels presentats als resultats del capítol anterior ja que aquí només s'està avaluant una sola variable.

- *Experiment 1*

En el primer experiment s'ha protegit una variable amb 8 categories disponibles.

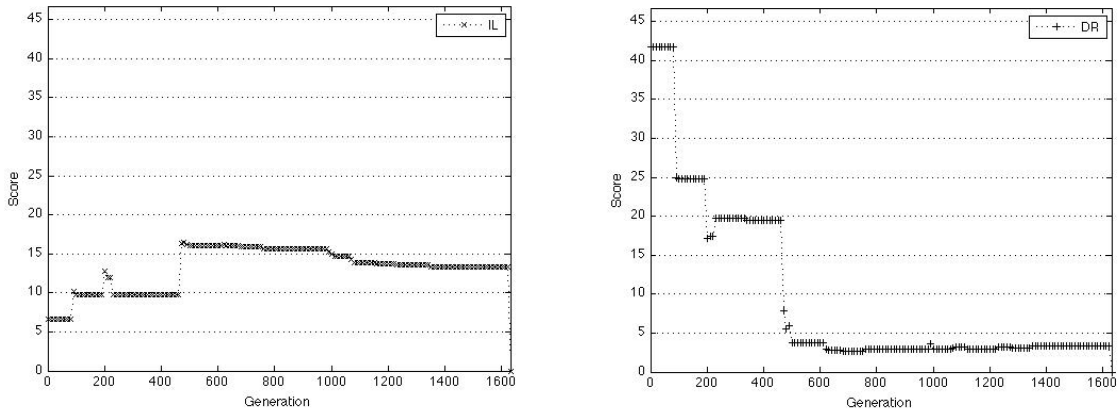



**Figura 4.1 – Gràfics indicant l'evolució de la pèrdua d'informació (IL) i del risc de revelació (DR) corresponents a cada generació en el primer experiment**

Com es pot observar a la figura 4.1 els valors de la pèrdua d'informació i del risc de revelació s'han anat ajustant per tal de conseguir un *score* menor, la qual cosa no vol dir que els dos valors hagin anat decrementant ja que la pèrdua d'informació ha augmentat, sinó que vol dir que s'ha anat ajustant la combinació dels dos valors per tal que la seva mitjana conjunta sí que anés decrementant.

En el cas de la pèrdua d'informació s'ha experimentat un increment aproximadament del doble del valor inicial. Tot i això encara s'està en un valor relativament baix (menys del 15%). Per contra, el risc de revelació experimenta un decrement de fins aproximadament una vuitena part del valor inicial.

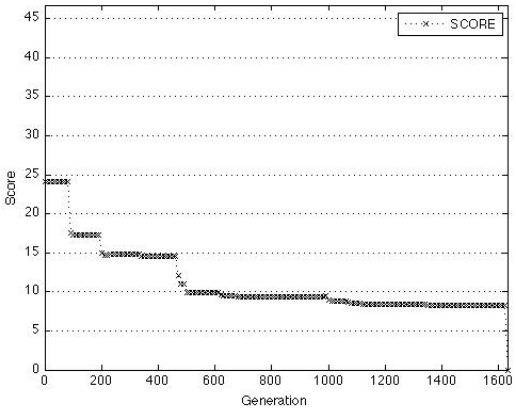



**Figura 4.2 – Gràfic indicant l'evolució del *score* total corresponent a cada generació en el primer experiment**

En el gràfic de la figura 4.2 corresponent a l'evolució del *score* total, podem observar que s'ha aconseguit una reducció de més del 65% en poc més de 1600 generacions.

- *Experiment 2*

En aquest segon experiment s'ha protegit una variable amb 25 categories disponibles.

Com es pot observar a la figura 4.3, en aquest segón experiment els valors de la pèrdua d'informació i del risc de revelació també s'han anat ajustant per tal de conseguir un *score* menor, la qual cosa no vol dir que els dos valors hagin anat decrementant ja que la pèrdua d'informació ha augmentat, sinó que vol dir que s'ha anat ajustant la combinació dels dos valors per tal que la seva mitjana conjunta sí que anés decrementant.

En el cas de la pèrdua d'informació s'ha experimentat un increment aproximadament del doble del valor inicial. Tot i que es té una pèrdua d'informació superior a l'experiment anterior també s'està en un valor relativament baix (poc més del 15%). Per contra, el risc de revelació experimenta un decrement de més d'un 90% respecte el valor inicial.

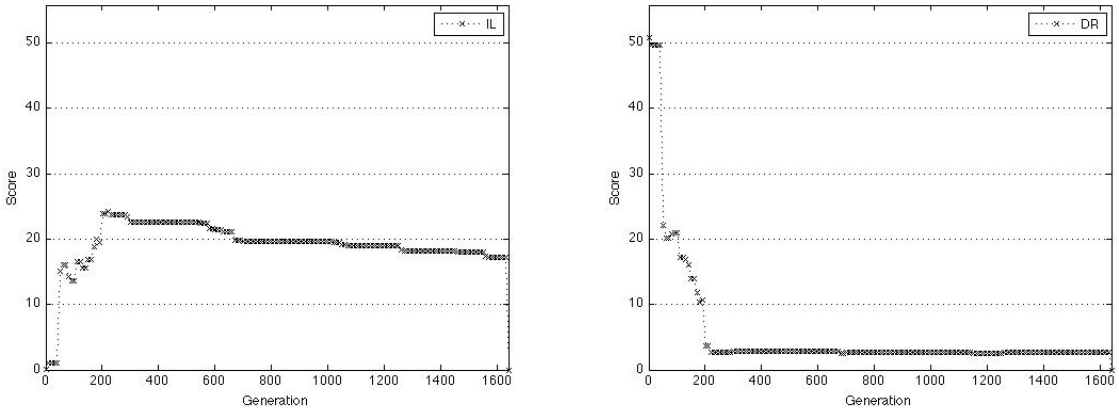



**Figura 4.3 – Gràfics indicant l'evolució de la pèrdua d'informació (IL) i del risc de revelació (DR) corresponents a cada generació en el segón experiment**

Finalment, en el gràfic de la figura 4.4 corresponent a l'evolució del *score* total, podem observar que aquí s'ha aconseguit una reducció de més del 70% en poc més de 1600 generacions.

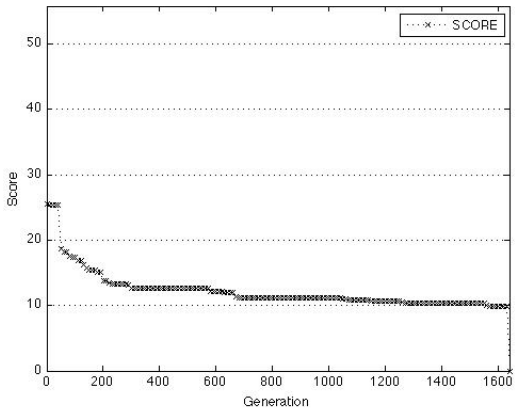



**Figura 4.4 – Gràfic indicant l'evolució del *score* total corresponent a cada generació en el segón experiment**

Per últim, la figura 4.5 mòstra desglossats tant el score inicial com el final per a cada experiment i els seus paràmetres.

S'hi pot observar com per als dos experiment el resultat final és significativament més proper a l'ideal (0,0) que l'inicial.

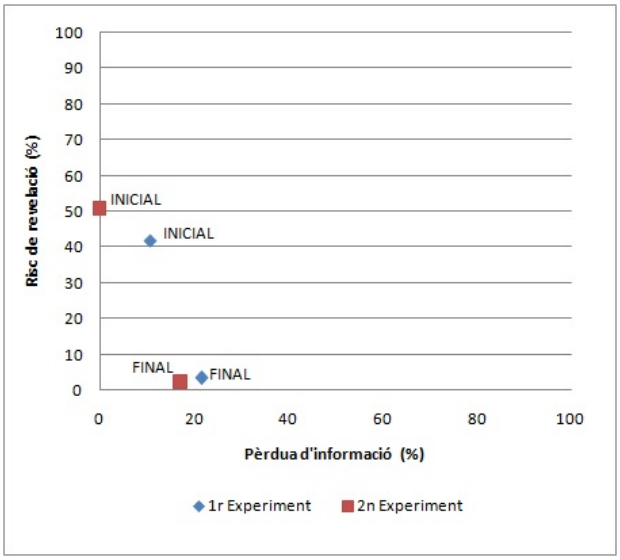



**Figura 4.5 – Gràfic amb els resultats del risc de revelació total mitjà per a cada valor de *P***

Així doncs queda demostrada la important millora que aporta aquest mètode optimitzador respecte els resultats originals del mètode de protecció PRAM.

# 5.- CONCLUSIONS

En aquest projecte s'ha vist que el camp de la privacitat de dades categòriques és un camp difícil ja que aquestes ténen un valor més semàntic que les dades contínues i per tant les operacions que s'hi poden realitzar a l'hora de protegir-les són més complexes (no es poden realitzar operacions aritmètiques directament sobre les dades).

Així, segons el treball realitzat en aquest projecte, es poden extreure les següents conclusions.

En primer lloc s'ha vist que donat un conjunt de dades a protegir, els mètodes de protecció tenen comportaments diferents sobre les dades ja que cada selecció de categories és diferent.

A més, també s'ha vist que tots els comportaments dels mètodes de protecció són altament depenents del conjunt de dades, sobretot de la quantitat i freqüència de les categories de les variables. Per tant no es pot parlar d'un mètode millor que la resta ja que en alguns casos un mètode donarà millors resultats perquè el conjunt de dades afavorirà el seu funcionament, i en altres casos serà un altre el mètode que en sortirà beneficiat.

No obstant, centrant-nos en el conjunt de dades utilitzat en aquest projecte, hi ha el Rank Swapping i la Microagregació com a mètodes que millor han funcionat ja que, com s'ha vist a l'apartat de resultats del capítol 3, han obtingut els valors de *score* més baixos en tot moment.

Contràriament el mètode que ha sortit més perjudicat ha estat el PRAM degut als escassos canvis que ha produït a les dades originals, obtenint els valors de *score* més alts.

Finalment, s'ha demostrat que es poden optimitzar els resultats del PRAM, el qual ha obtingut la pitjor qualificació de tots en els anàlisis, mitjançant un algorisme genètic que busca matrius de probabilitats de canvi de categoria com a solucions. Els resultats han estat bastant òptims i, tal com s'observa a l'apartat de resultats del capítol anterior, han permès rebaixar més d'un 65% el valor de *score* corresponent a una sola variable a l'experiment 1 (*veure figura 4.2*), i més del 70% a l'experiment 2 (*veure figura 4.4*).

# 6.- TREBALL FUTUR

De cara al treball a realitzar en el futur hi ha tres punts principals a tenir en compte:

En primer lloc s'hauria de realitzar proves amb més conjunts de dades categòriques ja que només n'ha utilitzat un i, com s'ha comentat, els resultats dels mètodes són depenents del conjunt de dades. Així obtenint els resultats de diversos conjunts de dades es podrien agafar les mitjanes dels resultats de cadascun d'ells i d'aquesta manera obtenir resultats més robustos.

Un altre punt a tractar és l'aplicació d'algorismes genètics per tal de realitzar un optimitzador de proteccions de dades categòriques en general, no només centrant-nos en el mètode PRAM. D'aquesta manera es podria definir la mutació com a el canvi d'una categoria per una altra de forma aleatòria, i el creuament com a l'intercanvi de valors entre files.

Per últim, caldria trobar una normalització dels valors corresponents a la pèrdua d'informació per tal d'obtenir un resultat més robust. Cal recordar que cada mesura de pèrdua té la seva escala. En aquest projecte s'ha realitzat una normalització dels valors respecte els màxims trobats entre els resultats de totes les proves.

# 7.- REFERÈNCIES

## - Publicacions:


- Domingo-Ferrer, J., Torra, V., (2001) '*Disclosure control methods and information loss for microdata*'
- Domingo-Ferrer, J., Torra, V., (2001) '*A quantitative comparison of disclosure control methods for microdata*'
- Torra, V., (2000), '*On information loss measures for categorical variables*'
- Torra, V., (2006), '*SW2: Record linkage software description*'
- Gouweleeuw, J.M., Kooiman, P., Willenborg, L.C.R.J., de Wolf, P., (1998) '*Post randomization for statistical disclosure control: theory and implementation*'
- Gouweleeuw, J.M., Kooiman, P., Willenborg, L.C.R.J., de Wolf, P., (1998), '*The post randomization method for protecting microdata*'
- Sweeney, L. (2002), '*Achieving k-anonymity privacy protection using generalization and suppression*'
- Sweeney, L. (2002) 'k-Anonymity for protecting privacy'
- Torra, V., (2008), '*Constrained microaggregation: adding constraints for data editing*'
- Domingo-Ferrer, J., Torra, V., (2005), '*Ordinal, continuous and heterogeneous k-anonymity through microaggregation*'


## - Llocs web:


- *Privacy Preserving Data Mining (PPDM)*
  Grup de recerca en privadesa a l'IIIA-CSIC
  http://www.ppdm.cat/

- *Modeling Decisions for Artificial Intelligence*
  Modelització de Decisions per a la Intel·ligència Artificial
  http://www.mdai.cat/

- *SpringerLink*
  Base de dades de publicacions científiques, tècniques i mèdiques
  http://www.springerlink.com

- *UCI Machine Learning Repository*
  Repositori de data sets per a la realització de proves
  http://archive.ics.uci.edu/

## - Llibres:

- *Privacy in Statistical Databases*
  CASC Project Final Conference, PSD 2004
  (2004) Springer-Verlag

- *Modeling decisions: Information fusion and aggregation operators*
  V. Torra, Y. Narukawa
  (2007) Springer-Verlag

- *Elements of Statistical Disclosure Control*
  L.Willenborg, T. de Waal
  (2001) Springer-Verlag

- *Information Fusion in Data Mining*
  V. Torra
  (2003) Springer-Verlag

- *Genetic Algorithms + Data Structures = Evolution Programs*
  Z. Michalewicz
  (1996) Springer-Verlag

# ANNEX I

## RESULTATS PARCIALS DELS EXPERIMENTS CORRESPONENTS A LA MICROAGREGACIÓ

- *PÈRDUA D'INFORMACIÓ (K FIXADA)*

  - *EXPERIMENTS*

**N=1 K=6**

| | CTBIL | ACTBIL | DIST | EBIL |
|---|---|---|---|---|
| 1 | 254 | 0,268 | 5,633 | 254,53 |
| 2 | 262 | 0,277 | 5,633 | 260,63 |
| 3 | 242 | 0,256 | 5,633 | 242,21 |
| 4 | 260 | 0,275 | 5,633 | 254,99 |
| 5 | 260 | 0,275 | 5,633 | 247,68 |

**N=2 K=6**

| | CTBIL | ACTBIL | DIST | EBIL |
|---|---|---|---|---|
| 1 | 566 | 0,598 | 10,6 | 627,58 |
| 2 | 598 | 0,631 | 10,6 | 626,74 |
| 3 | 628 | 0,663 | 10,6 | 634,74 |
| 4 | 592 | 0,625 | 10,6 | 624,49 |
| 5 | 600 | 0,634 | 10,6 | 628,29 |

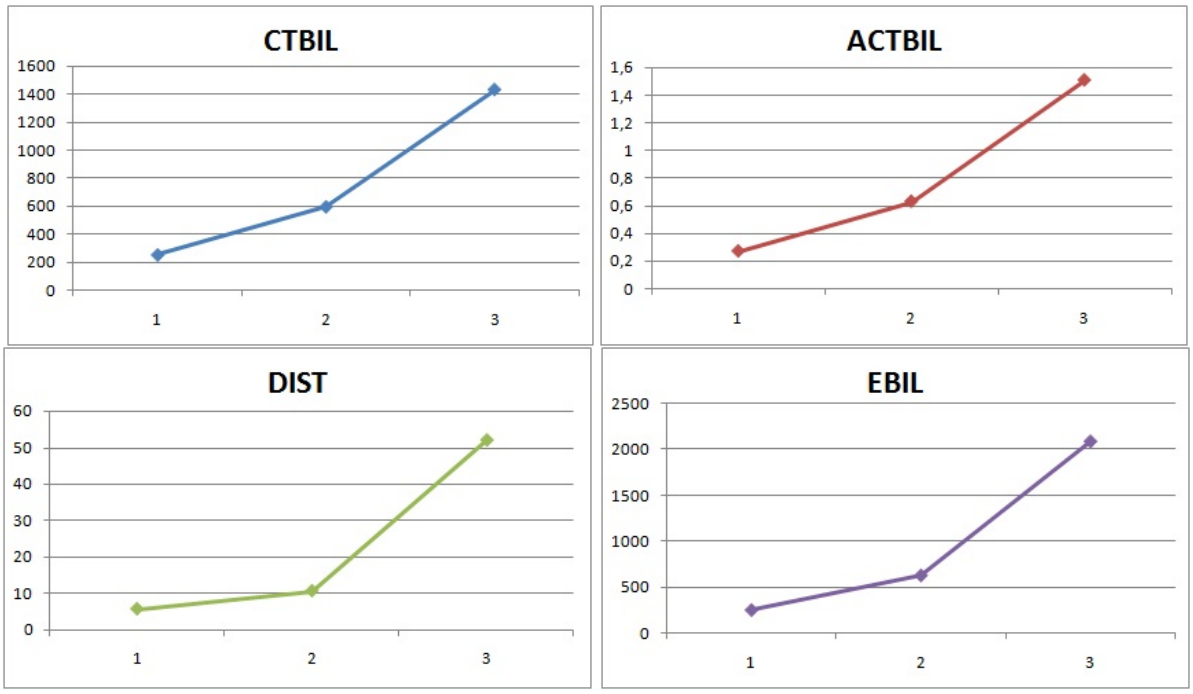
**N=3 K=6**

| | CTBIL | ACTBIL | DIST | EBIL |
|---|---|---|---|---|
| 1 | 1438 | 1,518 | 52,2 | 2082,33 |
| 2 | 1428 | 1,508 | 51,06 | 2084,47 |
| 3 | 1466 | 1,548 | 52,2 | 2080,95 |
| 4 | 1404 | 1,483 | 52,2 | 2071,94 |
| 5 | 1404 | 1,483 | 52,2 | 2093,21 |

  - *MITJANES*

| N | CTBIL | ACTBIL | DIST | EBIL | AIL |
|---|---|---|---|---|---|
| 1 | 255,6 | 0,2702 | 5,633 | 252,008 | 2,238779 |
| 2 | 596,8 | 0,6302 | 10,6 | 628,368 | 5,397774 |
| 3 | 1428 | 1,508 | 51,972 | 2082,58 | 18,52225 |

  - *GRÀFICS*



- *PÈRDUA D'INFORMACIÓ (N FIXADA)*

  - *EXPERIMENTS*

**N=3 K=1**

| | CTBIL | ACTBIL | DIST | EBIL |
|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 |

**N=3 K=2**

| | CTBIL | ACTBIL | DIST | EBIL |
|---|---|---|---|---|
| 1 | 594 | 0,628 | 16,28 | 977,13 |
| 2 | 598 | 0,631 | 16,28 | 1013,51 |
| 3 | 564 | 0,596 | 16,28 | 1001,54 |
| 4 | 576 | 0,608 | 16,28 | 988,9 |
| 5 | 568 | 0,599 | 16,28 | 1002,18 |

**N=3 K=4**

| | CTBIL | ACTBIL | DIST | EBIL |
|---|---|---|---|---|
| 1 | 1116 | 1,178 | 33,62 | 1678,9 |
| 2 | 1002 | 1,058 | 34,2 | 1671,28 |
| 3 | 1098 | 1,159 | 33,62 | 1683,19 |
| 4 | 1076 | 1,136 | 34,2 | 1671,99 |
| 5 | 1088 | 1,149 | 33,62 | 1700,91 |

**N=3 K=6**

| | CTBIL | ACTBIL | DIST | EBIL |
|---|---|---|---|---|
| 1 | 1398 | 1,476 | 52,2 | 2079 |
| 2 | 1354 | 1,429 | 52,2 | 2080,26 |
| 3 | 1318 | 1,392 | 52,2 | 2079,3 |
| 4 | 1492 | 1,576 | 51,06 | 2069,02 |
| 5 | 1484 | 1,567 | 52,2 | 2082,41 |

**N=3 K=8**

| | CTBIL | ACTBIL | DIST | EBIL |
|---|---|---|---|---|
| 1 | 1982 | 2,093 | 68,2 | 2479,37 |
| 2 | 1910 | 2,017 | 67,92 | 2453,29 |
| 3 | 1976 | 2,087 | 67,92 | 2434,14 |
| 4 | 1960 | 2,069 | 66,97 | 2436,32 |
| 5 | 1870 | 1,975 | 68,2 | 2459,68 |

**N=3 K=10**

| | CTBIL | ACTBIL | DIST | EBIL |
|---|---|---|---|---|
| 1 | 2110 | 2,228 | 76,54 | 2666 |
| 2 | 2274 | 2,401 | 76,54 | 2662,63 |
| 3 | 2256 | 2,382 | 76,93 | 2697,9 |
| 4 | 2234 | 2,359 | 76,54 | 2671,35 |
| 5 | 2346 | 2,477 | 76,54 | 2682,7 |

**N=3 K=12**

| | CTBIL | ACTBIL | DIST | EBIL |
|---|---|---|---|---|
| 1 | 2430 | 2,566 | 94,73 | 2911,67 |
| 2 | 2396 | 2,53 | 93,68 | 2906,77 |
| 3 | 2438 | 2,574 | 94,16 | 2891,86 |
| 4 | 2564 | 2,707 | 94,73 | 2909,74 |
| 5 | 2344 | 2,475 | 93,68 | 2904,63 |

**N=3 K=14**

| | CTBIL | ACTBIL | DIST | EBIL |
|---|---|---|---|---|
| 1 | 3012 | 3,181 | 93,14 | 2965,47 |
| 2 | 2750 | 2,904 | 93,14 | 2956,8 |
| 3 | 2814 | 2,971 | 97,24 | 2999,05 |
| 4 | 2774 | 2,929 | 100,47 | 3032,5 |
| 5 | 2836 | 2,995 | 96,38 | 3013,95 |

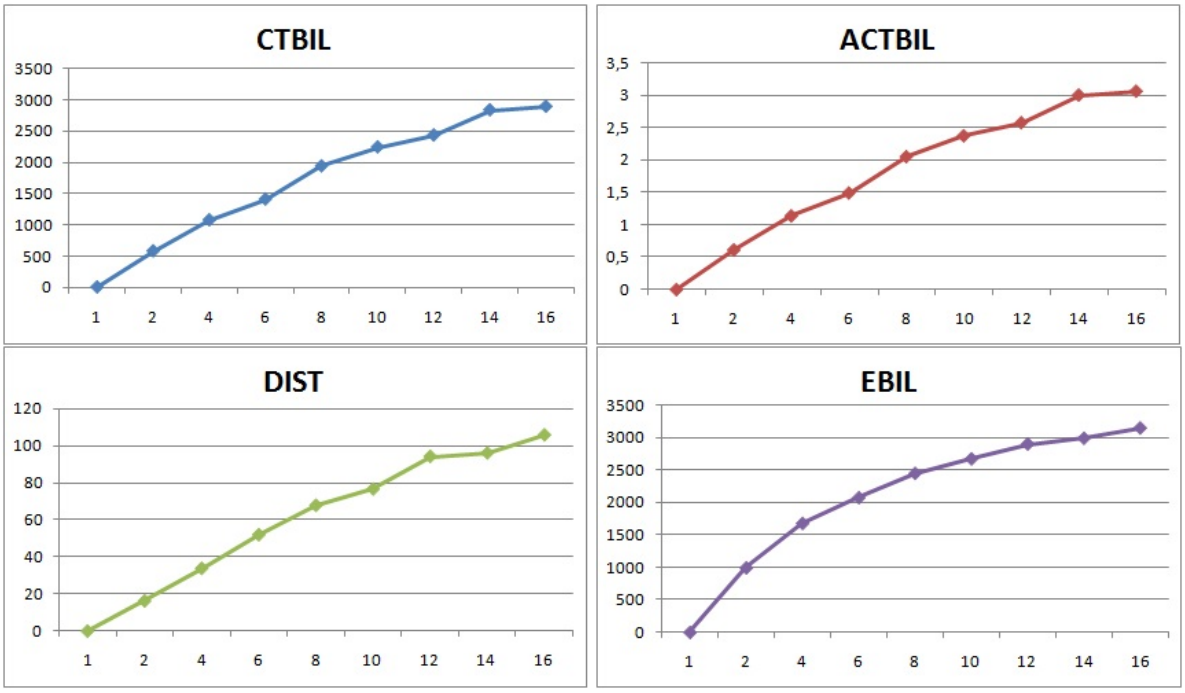
**N=3 K=16**

| | CTBIL | ACTBIL | DIST | EBIL |
|---|---|---|---|---|
| 1 | 2816 | 2,973 | 105,7 | 3150,24 |
| 2 | 3030 | 3,2 | 105,7 | 3166,35 |
| 3 | 2890 | 3,052 | 106,08 | 3146,48 |
| 4 | 2782 | 2,938 | 105,7 | 3163,72 |
| 5 | 2946 | 3,111 | 105,7 | 3131,76 |

  - *MITJANES*

| K | CTBIL | ACTBIL | DIST | EBIL | AIL |
|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 0,000 |
| 2 | 580 | 0,6124 | 16,28 | 996,652 | 8,404 |
| 4 | 1076 | 1,136 | 33,852 | 1681,254 | 14,526 |
| 6 | 1409 | 1,488 | 51,972 | 2077,998 | 18,481 |
| 8 | 1940 | 2,0482 | 67,842 | 2452,56 | 22,232 |
| 10 | 2244 | 2,3694 | 76,618 | 2676,116 | 24,432 |
| 12 | 2434 | 2,5704 | 94,196 | 2904,934 | 27,064 |
| 14 | 2837 | 2,996 | 96,074 | 2993,554 | 27,958 |
| 16 | 2893 | 3,0548 | 105,776 | 3151,71 | 29,629 |

- *GRÀFICS*




- *RISC DE REVELACIÓ (K FIXADA)*

  - *EXPERIMENTS*

**N=1 K=6**

| | ID | DBRL | PRL | RSRL |
|---|---|---|---|---|
| 1 | 98,1 | 44,2 | 43,9 | 10,1 |
| 2 | 98,067 | 44,2 | 44 | 10 |
| 3 | 98,1 | 44,3 | 44 | 9,9 |
| 4 | 98,067 | 44,5 | 44,4 | 9,7 |
| 5 | 98,067 | 44,4 | 44,3 | 9,6 |

**N=2 K=6**

| | ID | DBRL | PRL | RSRL |
|---|---|---|---|---|
| 1 | 94,233 | 40,4 | 39,7 | 6,8 |
| 2 | 94,233 | 40,1 | 39,8 | 7,2 |
| 3 | 94,2 | 39,8 | 39,3 | 7,2 |
| 4 | 94,233 | 40,2 | 39,4 | 7,6 |
| 5 | 94,233 | 40,3 | 39,8 | 7,2 |

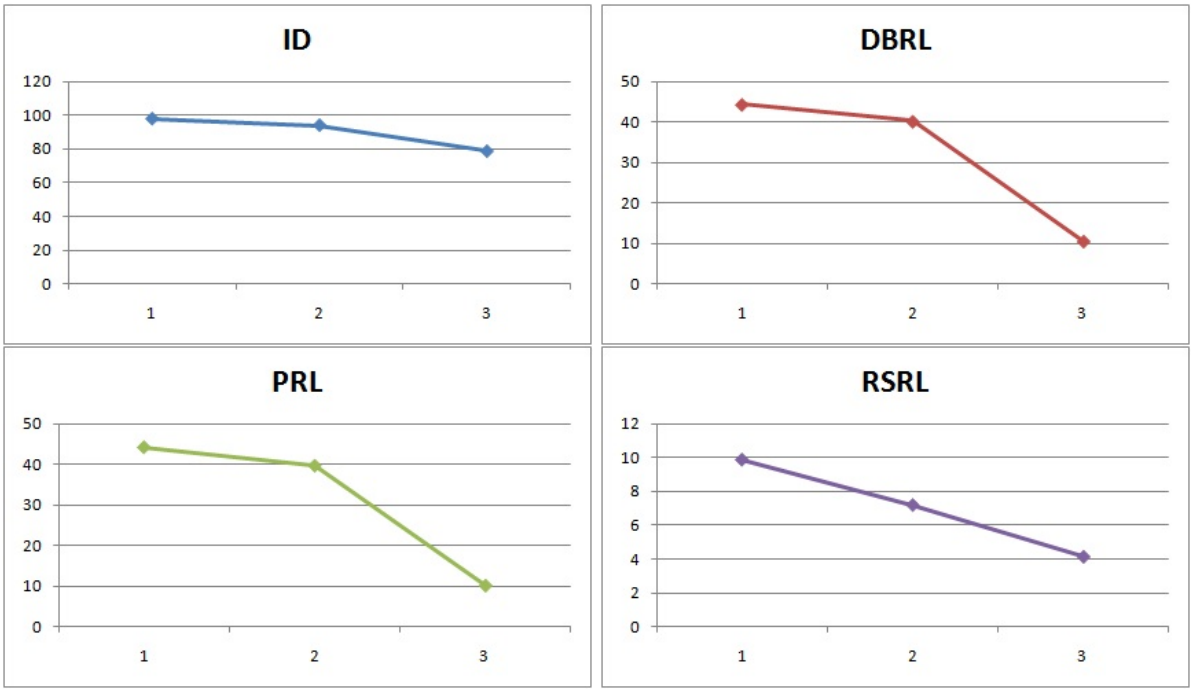
**N=3 K=6**

| | ID | DBRL | PRL | RSRL |
|---|---|---|---|---|
| 1 | 78,833 | 10,6 | 10,1 | 4,1 |
| 2 | 78,833 | 10,9 | 10,6 | 4,1 |
| 3 | 78,733 | 10,2 | 9,6 | 4,3 |
| 4 | 78,9 | 10,5 | 10,2 | 4,5 |
| 5 | 78,5 | 10,9 | 10,6 | 3,8 |

  - *MITJANES*

| N | ID | DBRL | PRL | RSRL | ADR |
|---|---|---|---|---|---|
| 1 | 98,0802 | 44,32 | 44,12 | 9,86 | 71,1998 |
| 2 | 94,2264 | 40,16 | 39,6 | 7,2 | 66,5936 |
| 3 | 78,7598 | 10,62 | 10,22 | 4,16 | 44,6902 |

- *GRÀFICS*




- *RISC DE REVELACIÓ (N FIXADA)*

  - *EXPERIMENTS*

**N=3 K=1**

| | ID | DBRL | PRL | RSRL |
|---|---|---|---|---|
| 1 | 100 | 47,1 | 47,1 | 11 |
| 2 | 100 | 47,1 | 47,1 | 11 |
| 3 | 100 | 47,1 | 47,1 | 11 |
| 4 | 100 | 47,1 | 47,1 | 11 |
| 5 | 100 | 47,1 | 47,1 | 11 |

**N=3 K=2**

| | ID | DBRL | PRL | RSRL |
|---|---|---|---|---|
| 1 | 91,1 | 27,1 | 26,6 | 6,5 |
| 2 | 91,1 | 27,7 | 26,8 | 7,2 |
| 3 | 91,1 | 27,3 | 26,7 | 7 |
| 4 | 91,1 | 27,8 | 26,5 | 6,3 |
| 5 | 91,1 | 26,7 | 26,3 | 7,3 |

**N=3 K=4**

| | ID | DBRL | PRL | RSRL |
|---|---|---|---|---|
| 1 | 83,167 | 15,7 | 15 | 5,1 |
| 2 | 83,167 | 15,5 | 15,2 | 5,5 |
| 3 | 83,2 | 16,1 | 15,2 | 4,7 |
| 4 | 83,267 | 14,8 | 14,2 | 5,2 |
| 5 | 83,233 | 15,8 | 14,9 | 5,4 |

**N=3 K=6**

| | ID | DBRL | PRL | RSRL |
|---|---|---|---|---|
| 1 | 78,633 | 10,8 | 10,6 | 4,3 |
| 2 | 78,733 | 10,9 | 10,3 | 4,3 |
| 3 | 78,767 | 10,7 | 10,4 | 4,2 |
| 4 | 78,5 | 10,9 | 10,2 | 4,3 |
| 5 | 78,567 | 10,2 | 10 | 4,1 |

**N=3 K=8**

| | ID | DBRL | PRL | RSRL |
|---|---|---|---|---|
| 1 | 72,9 | 8 | 7,5 | 2,8 |
| 2 | 72,933 | 8 | 7 | 2,9 |
| 3 | 73,267 | 7,9 | 7 | 2,7 |
| 4 | 73,167 | 7,5 | 7,2 | 3 |
| 5 | 72,7 | 7,9 | 7 | 2,8 |

**N=3 K=10**

| | ID | DBRL | PRL | RSRL |
|---|---|---|---|---|
| 1 | 70,467 | 6,5 | 4,3 | 2,4 |
| 2 | 70,467 | 6,1 | 3,8 | 2,8 |
| 3 | 70,367 | 6,4 | 4 | 2,4 |
| 4 | 70,433 | 6,5 | 5,4 | 2,5 |
| 5 | 70,4 | 6,2 | 3,9 | 2,3 |

**N=3 K=12**

| | ID | DBRL | PRL | RSRL |
|---|---|---|---|---|
| 1 | 66,767 | 5,4 | 4,8 | 2,2 |
| 2 | 66,9 | 5,2 | 4,8 | 1,9 |
| 3 | 66,633 | 5 | 4,3 | 2,2 |
| 4 | 66,433 | 5,1 | 4,6 | 2 |
| 5 | 67,033 | 5,1 | 4,6 | 2 |

**N=3 K=14**

| | ID | DBRL | PRL | RSRL |
|---|---|---|---|---|
| 1 | 64,1 | 4,3 | 3,4 | 2,1 |
| 2 | 63,933 | 4,3 | 3,9 | 2 |
| 3 | 63,5 | 4,6 | 3,8 | 1,7 |
| 4 | 63,4 | 4,3 | 3,5 | 1,9 |
| 5 | 63,8 | 4,3 | 3,4 | 1,8 |

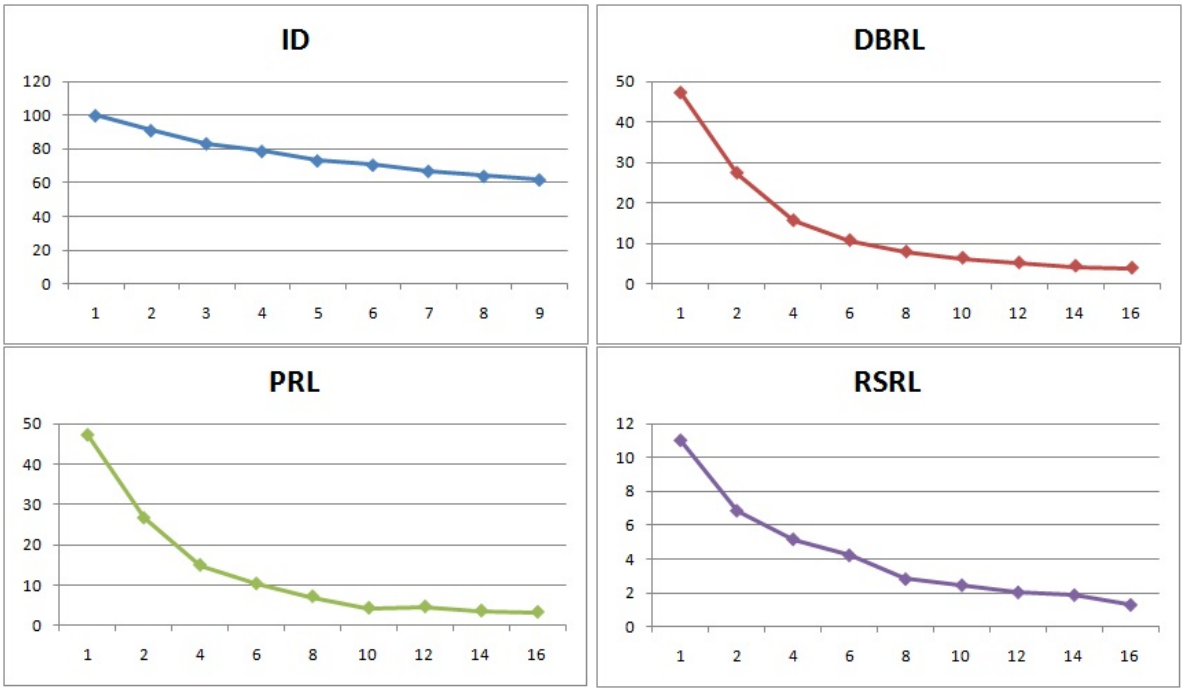
**N=3 K=16**

| | ID | DBRL | PRL | RSRL |
|---|---|---|---|---|
| 1 | 61,333 | 3,6 | 3,2 | 1,2 |
| 2 | 61,9 | 4 | 3,5 | 1,5 |
| 3 | 61,567 | 4,2 | 3,2 | 1,2 |
| 4 | 61,467 | 3,9 | 3,4 | 1,3 |
| 5 | 61,7 | 3,6 | 3 | 1,4 |

- *MITJANES*

| K | ID | DBRL | PRL | RSRL | ADR |
|---|---|---|---|---|---|
| 1 | 100 | 47,1 | 47,1 | 11 | 73,55 |
| 2 | 91,1 | 27,32 | 26,58 | 6,86 | 59,21 |
| 4 | 83,2068 | 15,58 | 14,9 | 5,18 | 49,3932 |
| 6 | 78,64 | 10,7 | 10,3 | 4,24 | 44,67 |
| 8 | 72,9934 | 7,86 | 7,14 | 2,84 | 40,4266 |
| 10 | 70,4268 | 6,34 | 4,28 | 2,48 | 38,3832 |
| 12 | 66,7532 | 5,16 | 4,62 | 2,06 | 35,9568 |
| 14 | 63,7466 | 4,36 | 3,6 | 1,9 | 34,0534 |
| 16 | 61,5934 | 3,86 | 3,26 | 1,32 | 32,7266 |

- *GRÀFICS*



# ANNEX II

## RESULTATS PARCIALS DELS EXPERIMENTS CORRESPONENTS AL PRAM

- PÈRDUA D'INFORMACIÓ

  - *EXPERIMENTS*

**P=1**

| | CTBIL | ACTBIL | DIST | EBIL |
|---|---|---|---|---|
| 1 | 98 | 0,103 | 9,52 | 141,44 |
| 2 | 98 | 0,103 | 6,99 | 124,26 |
| 3 | 108 | 0,114 | 9,32 | 139,65 |
| 4 | 102 | 0,108 | 7,68 | 123,59 |
| 5 | 92 | 0,097 | 8,01 | 137,85 |
| 6 | 92 | 0,097 | 8,52 | 113,61 |

**P=2**

| | CTBIL | ACTBIL | DIST | EBIL |
|---|---|---|---|---|
| 1 | 172 | 0,182 | 15,1 | 247,61 |
| 2 | 184 | 0,194 | 18,05 | 252,11 |
| 3 | 242 | 0,256 | 20,33 | 277,82 |
| 4 | 222 | 0,234 | 19,79 | 308,28 |
| 5 | 160 | 0,169 | 19,73 | 303,28 |
| 6 | 190 | 0,201 | 21,55 | 309,17 |

**P=3**

| | CTBIL | ACTBIL | DIST | EBIL |
|---|---|---|---|---|
| 1 | 208 | 0,22 | 23,12 | 318,76 |
| 2 | 182 | 0,192 | 18,11 | 316,69 |
| 3 | 310 | 0,327 | 35,53 | 443,32 |
| 4 | 232 | 0,245 | 26,74 | 374,82 |
| 5 | 248 | 0,262 | 27,02 | 390,71 |
| 6 | 280 | 0,296 | 28,84 | 394,12 |

**P=4**

| | CTBIL | ACTBIL | DIST | EBIL |
|---|---|---|---|---|
| 1 | 312 | 0,329 | 35,48 | 492,33 |
| 2 | 280 | 0,296 | 34,37 | 463,04 |
| 3 | 320 | 0,338 | 38,09 | 529,63 |
| 4 | 296 | 0,313 | 33,5 | 448,53 |
| 5 | 270 | 0,285 | 34,71 | 483,51 |
| 6 | 268 | 0,283 | 32,82 | 452,53 |

**P=5**

| | CTBIL | ACTBIL | DIST | EBIL |
|---|---|---|---|---|
| 1 | 404 | 0,427 | 48,77 | 627,63 |
| 2 | 414 | 0,437 | 44,97 | 606,21 |
| 3 | 312 | 0,329 | 37,53 | 530,57 |
| 4 | 306 | 0,323 | 40,71 | 567,89 |
| 5 | 376 | 0,397 | 42,97 | 577,95 |
| 6 | 358 | 0,378 | 41,99 | 569,02 |

**P=6**

| | CTBIL | ACTBIL | DIST | EBIL |
|---|---|---|---|---|
| 1 | 452 | 0,478 | 53,03 | 706,2 |
| 2 | 430 | 0,454 | 51,23 | 640,82 |
| 3 | 430 | 0,454 | 55,77 | 714,81 |
| 4 | 440 | 0,465 | 55,47 | 697,12 |
| 5 | 356 | 0,376 | 49,73 | 650,34 |
| 6 | 402 | 0,424 | 54,91 | 712,68 |

**P=7**

| | CTBIL | ACTBIL | DIST | EBIL |
|---|---|---|---|---|
| 1 | 432 | 0,456 | 59,04 | 708,24 |
| 2 | 470 | 0,496 | 63,63 | 751,14 |
| 3 | 468 | 0,494 | 59,47 | 709,32 |
| 4 | 504 | 0,532 | 61,96 | 773,23 |
| 5 | 484 | 0,511 | 65,41 | 776,7 |
| 6 | 514 | 0,543 | 63,68 | 758,37 |

**P=8**

| | CTBIL | ACTBIL | DIST | EBIL |
|---|---|---|---|---|
| 1 | 502 | 0,53 | 65,1 | 837,53 |
| 2 | 458 | 0,484 | 68,93 | 787,65 |
| 3 | 442 | 0,467 | 62,22 | 692,62 |
| 4 | 526 | 0,555 | 70,74 | 847,91 |
| 5 | 536 | 0,566 | 68,63 | 824,92 |
| 6 | 508 | 0,536 | 72,43 | 860,78 |

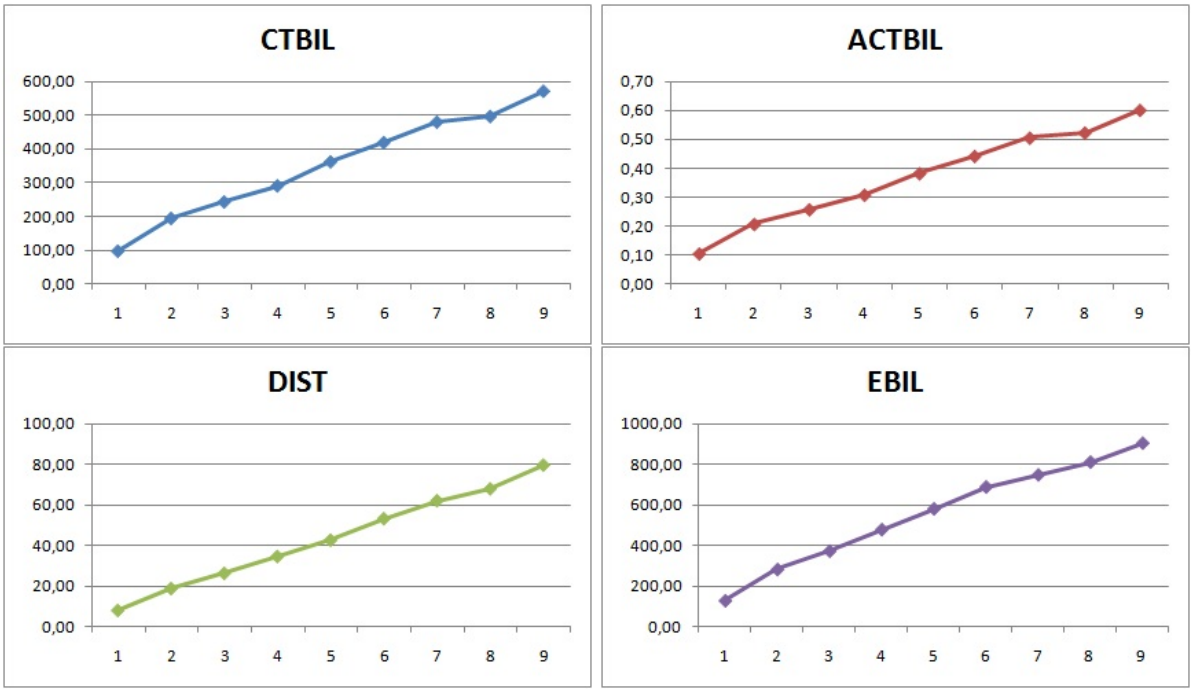
**P=9**

| | CTBIL | ACTBIL | DIST | EBIL |
|---|---|---|---|---|
| 1 | 604 | 0,638 | 81,99 | 898,92 |
| 2 | 640 | 0,676 | 83,76 | 915,8 |
| 3 | 554 | 0,586 | 77,37 | 891,43 |
| 4 | 538 | 0,568 | 73,59 | 848,03 |
| 5 | 530 | 0,559 | 77,39 | 896,78 |
| 6 | 552 | 0,583 | 83,72 | 957,65 |

  - *MITJANES*

| P | CTBIL | ACTBIL | DIST | EBIL | AIL |
|---|---|---|---|---|---|
| 1 | 98,33 | 0,10 | 8,34 | 130,07 | 1,411 |
| 2 | 195,00 | 0,21 | 19,09 | 283,05 | 3,110 |
| 3 | 243,33 | 0,26 | 26,56 | 373,07 | 4,163 |
| 4 | 291,00 | 0,31 | 34,83 | 478,26 | 5,368 |
| 5 | 361,67 | 0,38 | 42,82 | 579,88 | 6,541 |
| 6 | 418,33 | 0,44 | 53,36 | 687,00 | 7,875 |
| 7 | 478,67 | 0,51 | 62,20 | 746,17 | 8,771 |
| 8 | 495,33 | 0,52 | 68,01 | 808,57 | 9,526 |
| 9 | 569,67 | 0,60 | 79,64 | 901,44 | 10,815 |

- *GRÀFICS*




- *RISC DE REVELACIÓ*

- *EXPERIMENTS*

**P=1**

| | ID | DBRL | PRL | RSRL |
|---|---|---|---|---|
| 1 | 99,03 | 46,00 | 46,00 | 10,30 |
| 2 | 99,20 | 46,10 | 46,10 | 10,40 |
| 3 | 99,10 | 45,40 | 45,40 | 10,70 |
| 4 | 99,23 | 46,40 | 46,40 | 10,50 |
| 5 | 99,17 | 46,40 | 46,40 | 11,30 |
| 6 | 99,27 | 46,40 | 46,50 | 10,50 |

**P=2**

| | ID | DBRL | PRL | RSRL |
|---|---|---|---|---|
| 1 | 98,43 | 44,60 | 44,70 | 10,10 |
| 2 | 98,30 | 45,50 | 45,50 | 10,40 |
| 3 | 98,03 | 44,60 | 44,60 | 9,90 |
| 4 | 97,90 | 43,80 | 43,90 | 10,20 |
| 5 | 98,00 | 44,00 | 44,00 | 9,90 |
| 6 | 97,90 | 43,60 | 43,70 | 10,00 |

**P=3**

| | ID | DBRL | PRL | RSRL |
|---|---|---|---|---|
| 1 | 97,73 | 44,30 | 44,20 | 10,00 |
| 2 | 97,90 | 44,40 | 44,40 | 9,50 |
| 3 | 96,80 | 43,10 | 43,20 | 9,60 |
| 4 | 97,23 | 43,10 | 43,10 | 9,40 |
| 5 | 97,27 | 43,50 | 43,60 | 9,20 |
| 6 | 91,13 | 42,80 | 43,00 | 9,90 |

**P=4**

| | ID | DBRL | PRL | RSRL |
|---|---|---|---|---|
| 1 | 96,37 | 42,20 | 42,30 | 9,60 |
| 2 | 96,57 | 42,90 | 43,30 | 9,60 |
| 3 | 96,10 | 41,70 | 41,60 | 9,90 |
| 4 | 96,83 | 43,00 | 42,80 | 9,50 |
| 5 | 96,47 | 42,10 | 42,10 | 10,00 |
| 6 | 96,73 | 43,00 | 43,20 | 9,70 |

**P=5**

| | ID | DBRL | PRL | RSRL |
|---|---|---|---|---|
| 1 | 95,23 | 41,20 | 41,30 | 8,60 |
| 2 | 95,27 | 41,60 | 41,80 | 9,30 |
| 3 | 96,07 | 42,40 | 42,50 | 8,90 |
| 4 | 95,73 | 41,10 | 41,20 | 9,30 |
| 5 | 95,53 | 42,00 | 42,00 | 8,70 |
| 6 | 95,77 | 41,50 | 41,60 | 8,60 |

**P=6**

| | ID | DBRL | PRL | RSRL |
|---|---|---|---|---|
| 1 | 94,67 | 40,00 | 41,50 | 10,10 |
| 2 | 94,97 | 40,90 | 41,00 | 9,40 |
| 3 | 94,37 | 40,00 | 40,20 | 8,90 |
| 4 | 94,37 | 38,90 | 39,60 | 8,80 |
| 5 | 94,93 | 40,30 | 40,70 | 8,90 |
| 6 | 94,43 | 39,30 | 39,40 | 9,50 |

**P=7**

| | ID | DBRL | PRL | RSRL |
|---|---|---|---|---|
| 1 | 94,20 | 40,20 | 40,70 | 9,20 |
| 2 | 93,70 | 40,00 | 40,40 | 8,50 |
| 3 | 94,13 | 40,30 | 41,10 | 9,20 |
| 4 | 93,80 | 38,10 | 38,40 | 8,90 |
| 5 | 93,57 | 39,00 | 39,70 | 8,80 |
| 6 | 93,67 | 38,80 | 39,70 | 8,50 |

**P=8**

| | ID | DBRL | PRL | RSRL |
|---|---|---|---|---|
| 1 | 93,17 | 37,50 | 38,50 | 8,70 |
| 2 | 93,43 | 39,80 | 40,30 | 8,30 |
| 3 | 94,27 | 39,00 | 39,70 | 9,00 |
| 4 | 92,93 | 38,10 | 38,50 | 8,50 |
| 5 | 93,07 | 37,40 | 38,20 | 9,00 |
| 6 | 92,97 | 37,50 | 38,00 | 8,20 |

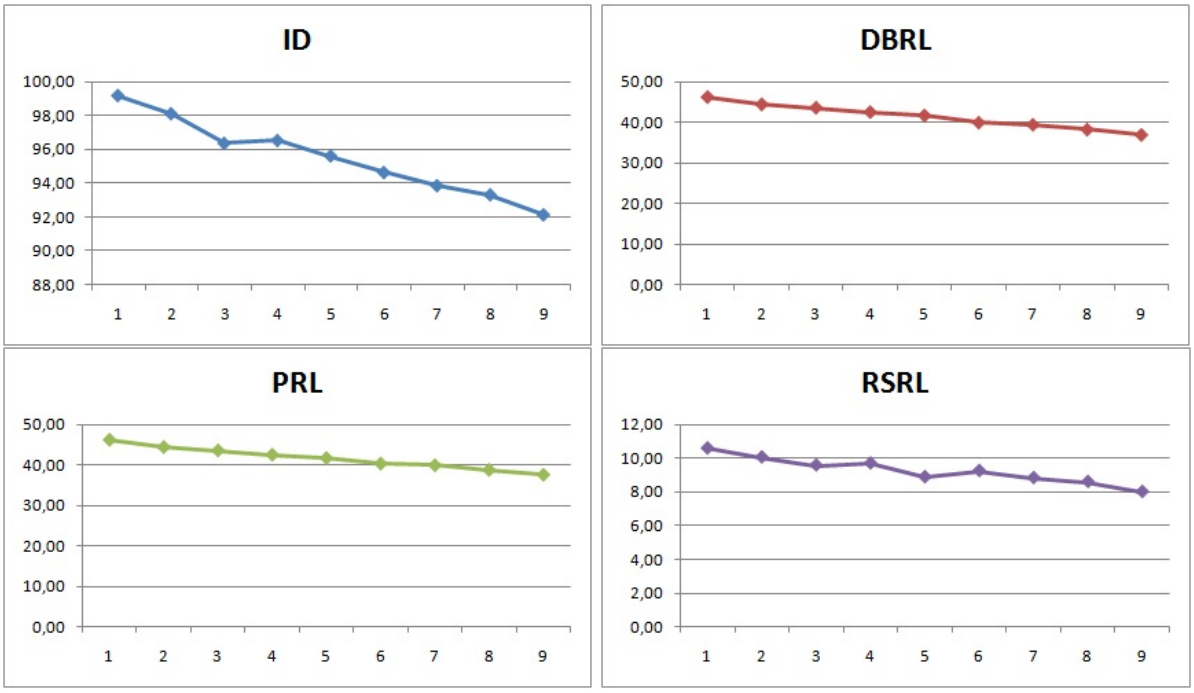
**P=9**

| | ID | DBRL | PRL | RSRL |
|---|---|---|---|---|
| 1 | 91,83 | 36,80 | 37,70 | 8,40 |
| 2 | 91,97 | 36,80 | 37,60 | 8,10 |
| 3 | 92,30 | 36,30 | 37,00 | 8,30 |
| 4 | 92,73 | 37,40 | 38,30 | 7,80 |
| 5 | 92,27 | 37,30 | 38,00 | 7,60 |
| 6 | 91,63 | 37,00 | 37,60 | 7,90 |

- *MITJANES*

| P | ID | DBRL | PRL | RSRL | ADR |
|---|---|---|---|---|---|
| 1 | 99,17 | 46,12 | 46,13 | 10,62 | 72,650 |
| 2 | 98,09 | 44,35 | 44,40 | 10,08 | 71,247 |
| 3 | 96,34 | 43,53 | 43,58 | 9,60 | 70,472 |
| 4 | 96,51 | 42,48 | 42,55 | 9,72 | 69,556 |
| 5 | 95,60 | 41,63 | 41,73 | 8,90 | 68,666 |
| 6 | 94,62 | 39,90 | 40,40 | 9,27 | 67,427 |
| 7 | 93,84 | 39,40 | 40,00 | 8,85 | 66,922 |
| 8 | 93,31 | 38,22 | 38,87 | 8,62 | 66,086 |
| 9 | 92,12 | 36,93 | 37,70 | 8,02 | 64,911 |

- *GRÀFICS*



# ANNEX III

## RESULTATS PARCIALS DELS EXPERIMENTS CORRESPONENTS AL RANK SWAPPING

- *PÈRDUA D'INFORMACIÓ*
  - *EXPERIMENTS*

**P=1**

| | CTBIL | ACTBIL | DIST | EBIL |
|---|---|---|---|---|
| 1 | 392 | 0,414 | 23,76 | 773,19 |
| 2 | 398 | 0,42 | 26,24 | 793,7 |
| 3 | 446 | 0,471 | 21,89 | 828,6 |
| 4 | 386 | 0,408 | 22,78 | 805,58 |
| 5 | 378 | 0,399 | 26,83 | 775,81 |

**P=2**

| | CTBIL | ACTBIL | DIST | EBIL |
|---|---|---|---|---|
| 1 | 538 | 0,568 | 38,62 | 1295,45 |
| 2 | 504 | 0,532 | 33,026 | 1192,17 |
| 3 | 542 | 0,572 | 40,13 | 1152,46 |
| 4 | 564 | 0,596 | 43,79 | 1201,58 |
| 5 | 602 | 0,636 | 39,89 | 1230,95 |

**P=4**

| | CTBIL | ACTBIL | DIST | EBIL |
|---|---|---|---|---|
| 1 | 752 | 0,794 | 65,53 | 1817,89 |
| 2 | 802 | 0,847 | 73,47 | 1859,25 |
| 3 | 780 | 0,824 | 71,62 | 1809,41 |
| 4 | 752 | 0,794 | 64,75 | 1832,47 |
| 5 | 758 | 0,8 | 80,44 | 1920,05 |

**P=6**

| | CTBIL | ACTBIL | DIST | EBIL |
|---|---|---|---|---|
| 1 | 1008 | 1,064 | 107,59 | 2366,42 |
| 2 | 930 | 0,982 | 116,54 | 2397,43 |
| 3 | 1010 | 1,067 | 116,22 | 2376,34 |
| 4 | 936 | 0,988 | 108,21 | 2312,13 |
| 5 | 968 | 1,022 | 111,03 | 2392,64 |

**P=8**

| | CTBIL | ACTBIL | DIST | EBIL |
|---|---|---|---|---|
| 1 | 1002 | 1,058 | 135,74 | 2754,13 |
| 2 | 1054 | 1,113 | 140,29 | 2773,4 |
| 3 | 1018 | 1,075 | 133,89 | 2650,44 |
| 4 | 1028 | 1,086 | 137,24 | 2749,01 |
| 5 | 1032 | 1,089 | 139,24 | 2731,96 |

**P=10**

| | CTBIL | ACTBIL | DIST | EBIL |
|---|---|---|---|---|
| 1 | 1100 | 1,162 | 170,53 | 3069,64 |
| 2 | 1138 | 1,202 | 161,37 | 3081,36 |
| 3 | 1158 | 1,223 | 171,19 | 3073,92 |
| 4 | 1160 | 1,225 | 161,25 | 3065,61 |
| 5 | 1112 | 1,174 | 165,6 | 3076,96 |

**P=12**

| | CTBIL | ACTBIL | DIST | EBIL |
|---|---|---|---|---|
| 1 | 1218 | 1,286 | 191,65 | 3387,5 |
| 2 | 1046 | 1,105 | 191,32 | 3312,7 |
| 3 | 1200 | 1,267 | 183,72 | 3370,12 |
| 4 | 1198 | 1,265 | 188,93 | 3339,58 |
| 5 | 1130 | 1,193 | 185,34 | 3311,69 |

**P=14**

| | CTBIL | ACTBIL | DIST | EBIL |
|---|---|---|---|---|
| 1 | 1148 | 1,212 | 214,22 | 3575,04 |
| 2 | 1202 | 1,269 | 213,35 | 3594,12 |
| 3 | 1186 | 1,253 | 214,28 | 3602,17 |
| 4 | 1184 | 1,25 | 210,82 | 3554,14 |
| 5 | 1286 | 1,358 | 219,94 | 3634,27 |

**P=16**

| | CTBIL | ACTBIL | DIST | EBIL |
|---|---|---|---|---|
| 1 | 1218 | 1,286 | 240,15 | 3764,89 |
| 2 | 1280 | 1,352 | 242,67 | 3771,59 |
| 3 | 1190 | 1,257 | 238,77 | 3809,32 |
| 4 | 1266 | 1,337 | 233,87 | 3765,26 |
| 5 | 1246 | 1,316 | 250,39 | 3780,77 |

**P=18**

| | CTBIL | ACTBIL | DIST | EBIL |
|---|---|---|---|---|
| 1 | 1270 | 1,341 | 268,59 | 3952,31 |
| 2 | 1236 | 1,605 | 258,97 | 3927,84 |
| 3 | 1334 | 1,409 | 273,79 | 4016,79 |
| 4 | 1284 | 1,356 | 263,49 | 3931,23 |
| 5 | 1258 | 1,329 | 261,11 | 3877,42 |

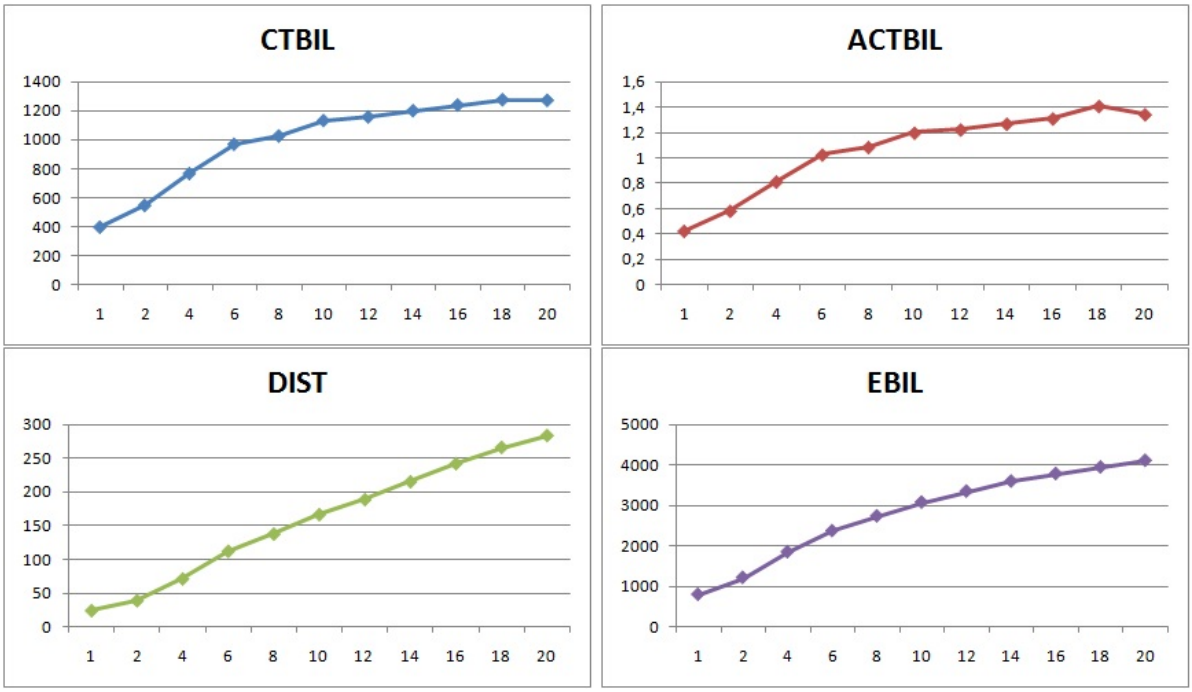
**P=20**

| | CTBIL | ACTBIL | DIST | EBIL |
|---|---|---|---|---|
| 1 | 1284 | 1,356 | 284,74 | 4099,98 |
| 2 | 1216 | 1,284 | 277,75 | 4098,31 |
| 3 | 1312 | 1,385 | 274,09 | 4066,27 |
| 4 | 1258 | 1,328 | 288,95 | 4114,23 |
| 5 | 1292 | 1,364 | 287,54 | 4143,41 |

- *MITJANES*

| P | CTBIL | ACTBIL | DIST | EBIL | AIL |
|---|---|---|---|---|---|
| 1 | 400 | 0,4224 | 24,3 | 795,376 | 7,241 |
| 2 | 550 | 0,5808 | 39,0912 | 1214,522 | 11,133 |
| 4 | 768,8 | 0,8118 | 71,162 | 1847,814 | 17,490 |
| 6 | 970,4 | 1,0246 | 111,918 | 2368,992 | 23,436 |
| 8 | 1027 | 1,0842 | 137,28 | 2731,788 | 27,397 |
| 10 | 1134 | 1,1972 | 165,988 | 3073,498 | 31,385 |
| 12 | 1158 | 1,2232 | 188,192 | 3344,318 | 34,497 |
| 14 | 1201 | 1,2684 | 214,522 | 3591,948 | 37,646 |
| 16 | 1240 | 1,3096 | 241,17 | 3778,366 | 40,356 |
| 18 | 1276 | 1,408 | 265,19 | 3941,118 | 42,780 |
| 20 | 1272 | 1,3434 | 282,614 | 4104,44 | 44,828 |

- *GRÀFICS*



- *RISC DE REVELACIÓ*

  - *EXPERIMENTS*

**P=1**

| | ID | DBRL | PRL | RSRL |
|---|---|---|---|---|
| 1 | 92,93 | 38,5 | 38,4 | 9 |
| 2 | 92,67 | 37 | 37,1 | 8,4 |
| 3 | 92,4 | 36,9 | 37 | 7,9 |
| 4 | 92,33 | 37,3 | 37,3 | 7,1 |
| 5 | 92,87 | 38,4 | 38,3 | 9,1 |

**P=2**

| | ID | DBRL | PRL | RSRL |
|---|---|---|---|---|
| 1 | 86,33 | 31 | 30,4 | 7,3 |
| 2 | 87,53 | 31,1 | 31,4 | 7,6 |
| 3 | 88,2 | 32,7 | 31,7 | 6,7 |
| 4 | 86,67 | 31,3 | 31 | 6,1 |
| 5 | 87,2 | 31,1 | 30,6 | 7,5 |

**P=4**

| | ID | DBRL | PRL | RSRL |
|---|---|---|---|---|
| 1 | 77,4 | 21,8 | 21,7 | 5,6 |
| 2 | 76,8 | 21,6 | 22 | 6 |
| 3 | 78,13 | 22,2 | 22,7 | 4,9 |
| 4 | 77,53 | 21 | 21,3 | 4,9 |
| 5 | 76,27 | 19,8 | 19 | 5,5 |

**P=6**

| | ID | DBRL | PRL | RSRL |
|---|---|---|---|---|
| 1 | 67 | 13,2 | 13,5 | 4,1 |
| 2 | 65,8 | 12 | 11,2 | 3,5 |
| 3 | 66,2 | 13,1 | 11,6 | 4,5 |
| 4 | 68 | 16,2 | 14,8 | 4,3 |
| 5 | 67,33 | 14 | 12,1 | 4,3 |

**P=8**

| | ID | DBRL | PRL | RSRL |
|---|---|---|---|---|
| 1 | 58,2 | 7,2 | 6,2 | 3,3 |
| 2 | 58,67 | 7,3 | 6,7 | 3,6 |
| 3 | 60,53 | 9,3 | 8 | 3,5 |
| 4 | 60,67 | 8,8 | 7,5 | 3,1 |
| 5 | 59,33 | 8,7 | 8,4 | 3,7 |

**P=10**

| | ID | DBRL | PRL | RSRL |
|---|---|---|---|---|
| 1 | 53,93 | 7,1 | 6,1 | 2,4 |
| 2 | 51,93 | 6,2 | 4,2 | 2,3 |
| 3 | 51,73 | 5,4 | 4,8 | 2,2 |
| 4 | 53,53 | 6,2 | 5 | 2,6 |
| 5 | 53,2 | 6,8 | 5,4 | 2,6 |

**P=12**

| | ID | DBRL | PRL | RSRL |
|---|---|---|---|---|
| 1 | 46,33 | 4,5 | 4 | 2 |
| 2 | 48,87 | 3,4 | 2,7 | 1,7 |
| 3 | 47,87 | 4,5 | 4,4 | 2,4 |
| 4 | 48,67 | 4,9 | 4,1 | 2,1 |
| 5 | 48,87 | 4,4 | 3,3 | 1,9 |

**P=14**

| | ID | DBRL | PRL | RSRL |
|---|---|---|---|---|
| 1 | 43,8 | 3,3 | 2,6 | 1,7 |
| 2 | 43,2 | 3,3 | 3,1 | 1,6 |
| 3 | 43,2 | 3,6 | 3 | 1,5 |
| 4 | 45,27 | 4,1 | 2,7 | 0,9 |
| 5 | 41,33 | 2,7 | 2,3 | 1 |

**P=16**

| | ID | DBRL | PRL | RSRL |
|---|---|---|---|---|
| 1 | 40,13 | 3 | 2,2 | 1,7 |
| 2 | 39,07 | 2,3 | 2,2 | 1,2 |
| 3 | 40,53 | 2,6 | 2,4 | 1,1 |
| 4 | 41,4 | 2,4 | 2,1 | 1,2 |
| 5 | 37,87 | 2,1 | 1,9 | 1,4 |

**P=18**

| | ID | DBRL | PRL | RSRL |
|---|---|---|---|---|
| 1 | 36 | 1,7 | 1,8 | 1,7 |
| 2 | 36,4 | 2 | 2 | 1,2 |
| 3 | 35,73 | 2,1 | 2 | 1,4 |
| 4 | 34,87 | 1,7 | 0,6 | 1,3 |
| 5 | 37 | 2,2 | 1,9 | 0,8 |

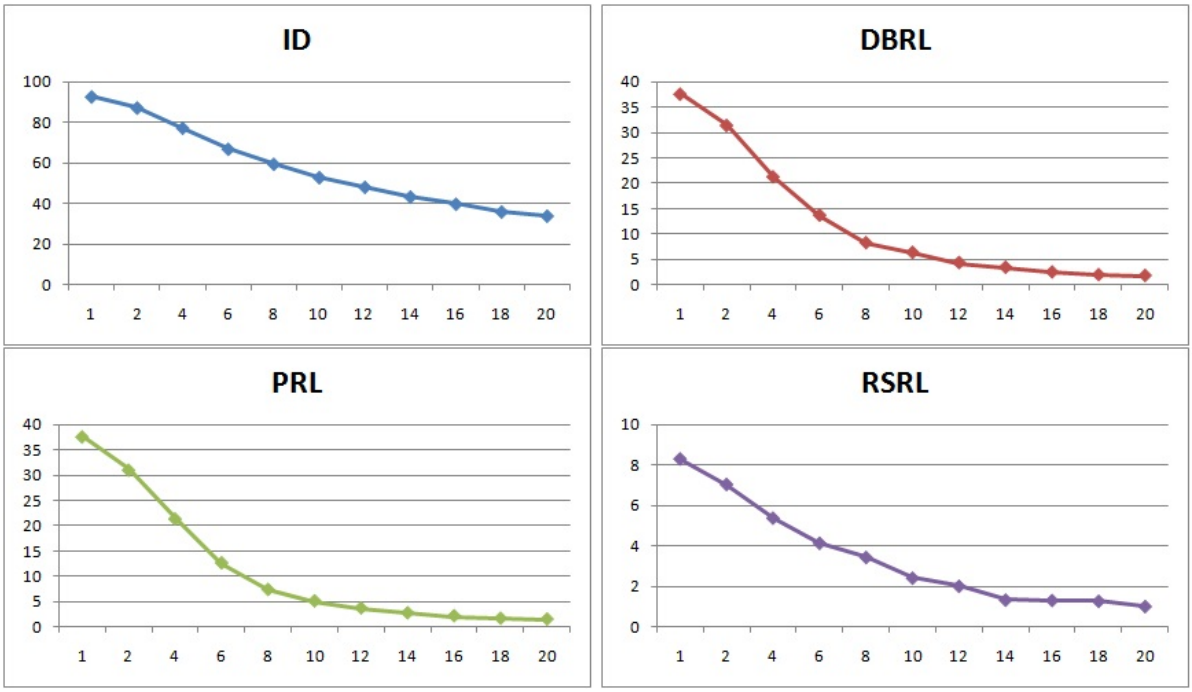
**P=20**

| | ID | DBRL | PRL | RSRL |
|---|---|---|---|---|
| 1 | 34,07 | 2,2 | 2 | 1,3 |
| 2 | 34,53 | 2,1 | 1,9 | 0,7 |
| 3 | 34,8 | 2 | 1,7 | 0,9 |
| 4 | 33,07 | 1 | 0,4 | 0,9 |
| 5 | 33,13 | 1,6 | 1,5 | 1,3 |

- *MITJANES*

| P | ID | DBRL | PRL | RSRL | ADR |
|---|---|---|---|---|---|
| 1 | 92,64 | 37,62 | 37,62 | 8,3 | 65,150 |
| 2 | 87,19 | 31,44 | 31,02 | 7,04 | 59,344 |
| 4 | 77,23 | 21,28 | 21,34 | 5,38 | 49,373 |
| 6 | 66,87 | 13,7 | 12,64 | 4,14 | 40,314 |
| 8 | 59,48 | 8,26 | 7,36 | 3,44 | 33,869 |
| 10 | 52,86 | 6,34 | 5,1 | 2,42 | 29,604 |
| 12 | 48,12 | 4,34 | 3,7 | 2,02 | 26,340 |
| 14 | 43,36 | 3,4 | 2,74 | 1,34 | 23,379 |
| 16 | 39,8 | 2,48 | 2,16 | 1,32 | 21,141 |
| 18 | 36 | 1,94 | 1,66 | 1,28 | 18,980 |
| 20 | 33,92 | 1,78 | 1,5 | 1,02 | 17,850 |

- *GRÀFICS*



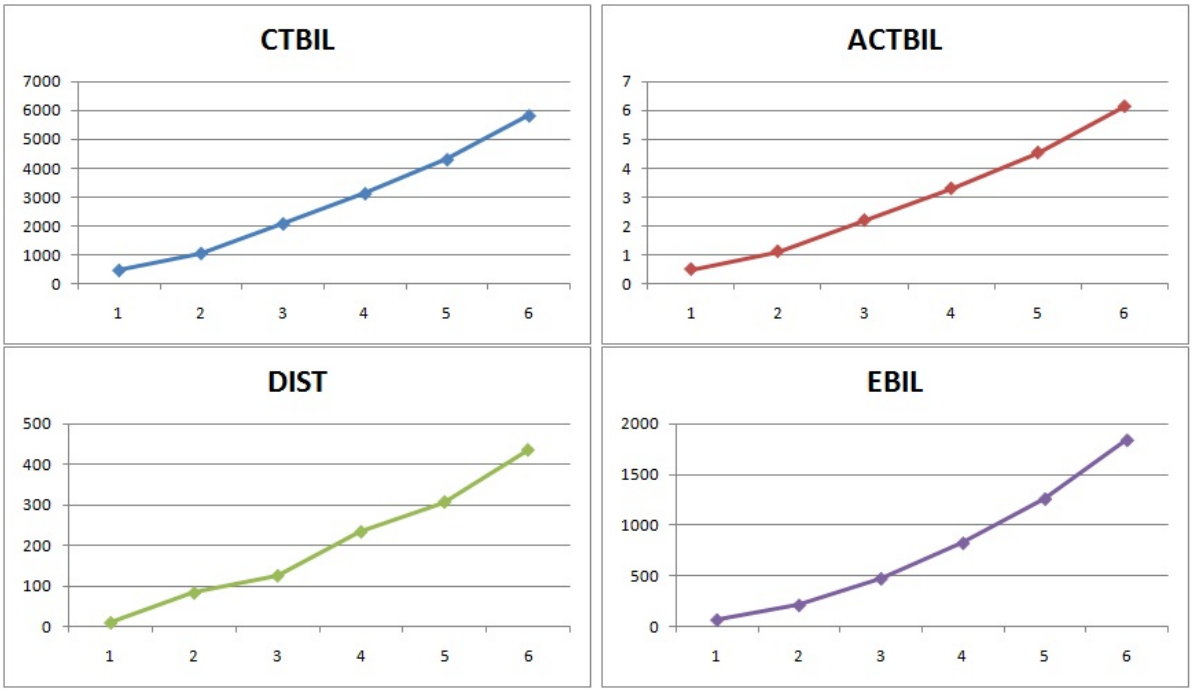
# ANNEX IV

## RESULTATS PARCIALS DELS EXPERIMENTS CORRESPONENTS AL GLOBAL RECODING

- *PÈRDUA D'INFORMACIÓ*
  - *EXPERIMENTS*

| P | CTBIL | ACTBIL | DIST | EBIL | AIL |
|---|---|---|---|---|---|
| 1 | 480 | 0,507 | 10 | 68,25 | 1,171 |
| 2 | 1068 | 1,128 | 83,5 | 213,97 | 6,083 |
| 3 | 2088 | 2,205 | 126 | 475,85 | 10,485 |
| 4 | 3132 | 3,307 | 234,36 | 826,23 | 18,796 |
| 5 | 4306 | 4,547 | 307,82 | 1262,13 | 26,070 |
| 6 | 5822 | 6,148 | 435,89 | 1842,49 | 37,227 |

  - *GRÀFICS*

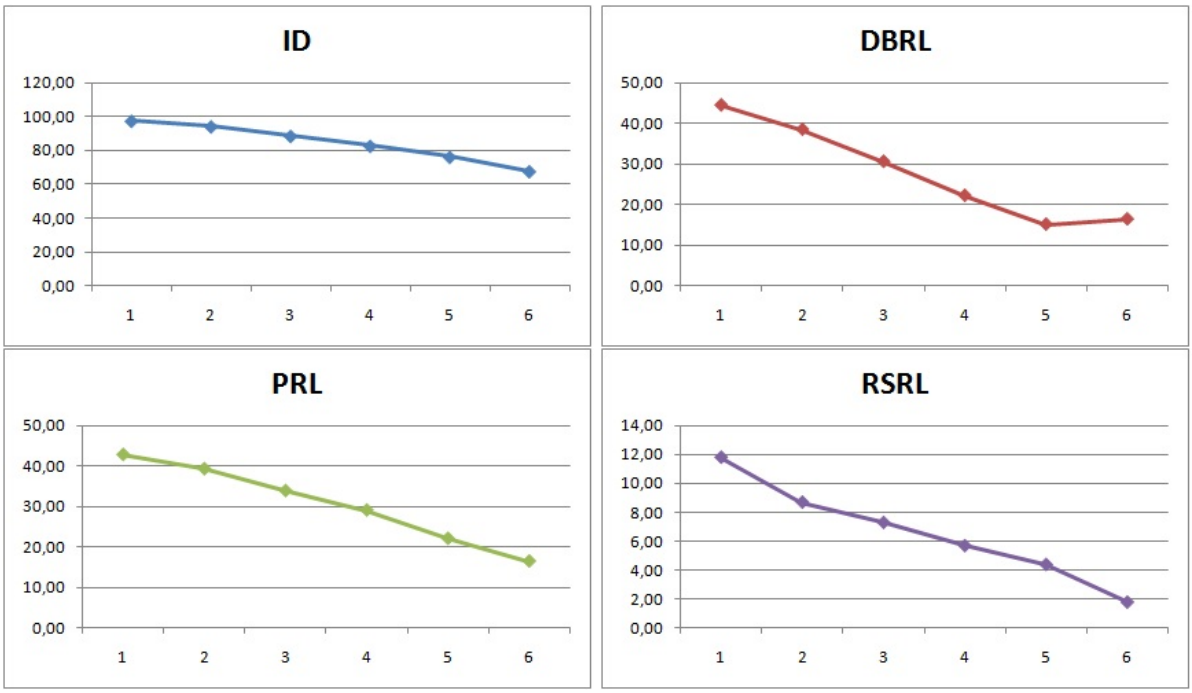


- *RISC DE REVELACIÓ*

  - *EXPERIMENTS*

| P | ID | DBRL | PRL | RSRL | ADR |
|---|---|---|---|---|---|
| 1 | 97,33 | 44,50 | 42,70 | 11,80 | 70,917 |
| 2 | 94,07 | 38,50 | 39,20 | 8,70 | 66,633 |
| 3 | 88,40 | 30,60 | 33,80 | 7,30 | 61,100 |
| 4 | 82,60 | 22,20 | 29,00 | 5,70 | 55,800 |
| 5 | 76,07 | 15,10 | 22,00 | 4,40 | 49,033 |
| 6 | 67,60 | 16,40 | 16,40 | 1,80 | 42,000 |

  - *GRÀFICS*



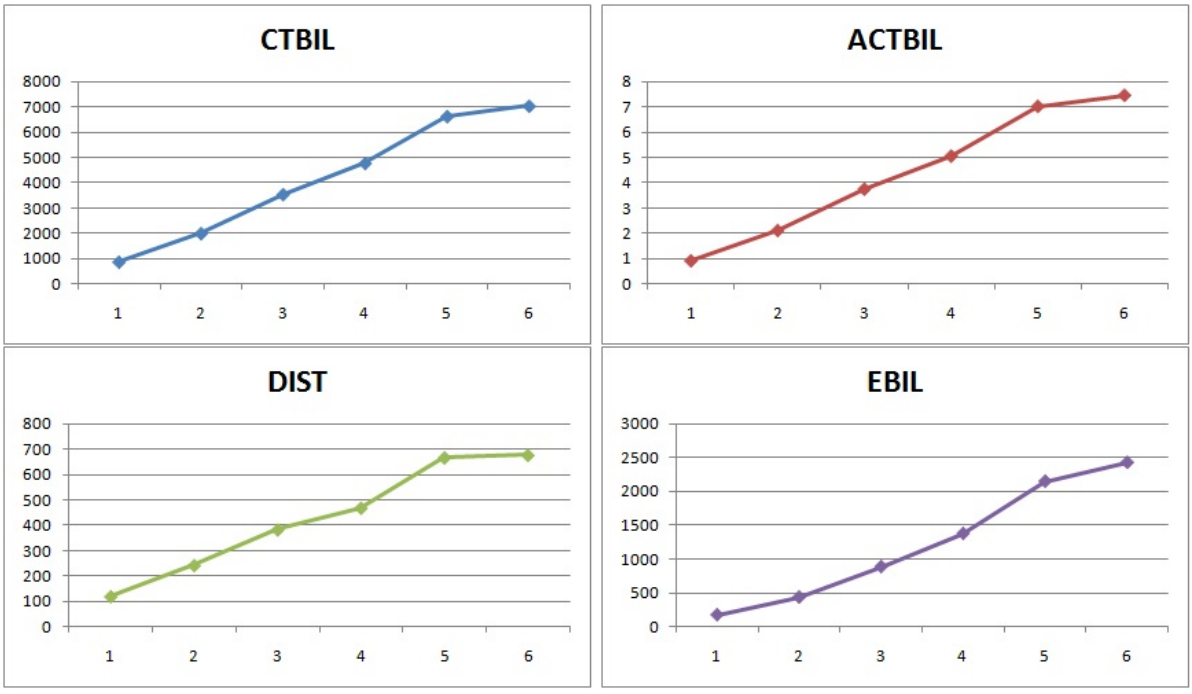
# ANNEX V

## RESULTATS PARCIALS DELS EXPERIMENTS CORRESPONENTS AL TOP CODING

- *PÈRDUA D'INFORMACIÓ*
  - *EXPERIMENTS*

| P | CTBIL | ACTBIL | DIST | EBIL | AIL |
|---|---|---|---|---|---|
| 1 | 876 | 0,925 | 117,63 | 171,96 | 7,384 |
| 2 | 2006 | 2,118 | 240,71 | 433,02 | 15,789 |
| 3 | 3550 | 3,749 | 381,63 | 885,22 | 26,638 |
| 4 | 4780 | 5,048 | 467,7 | 1377,68 | 34,972 |
| 5 | 6634 | 7,005 | 666,68 | 2152,25 | 51,185 |
| 6 | 7044 | 7,438 | 676,68 | 2431,02 | 53,892 |

  - *GRÀFICS*

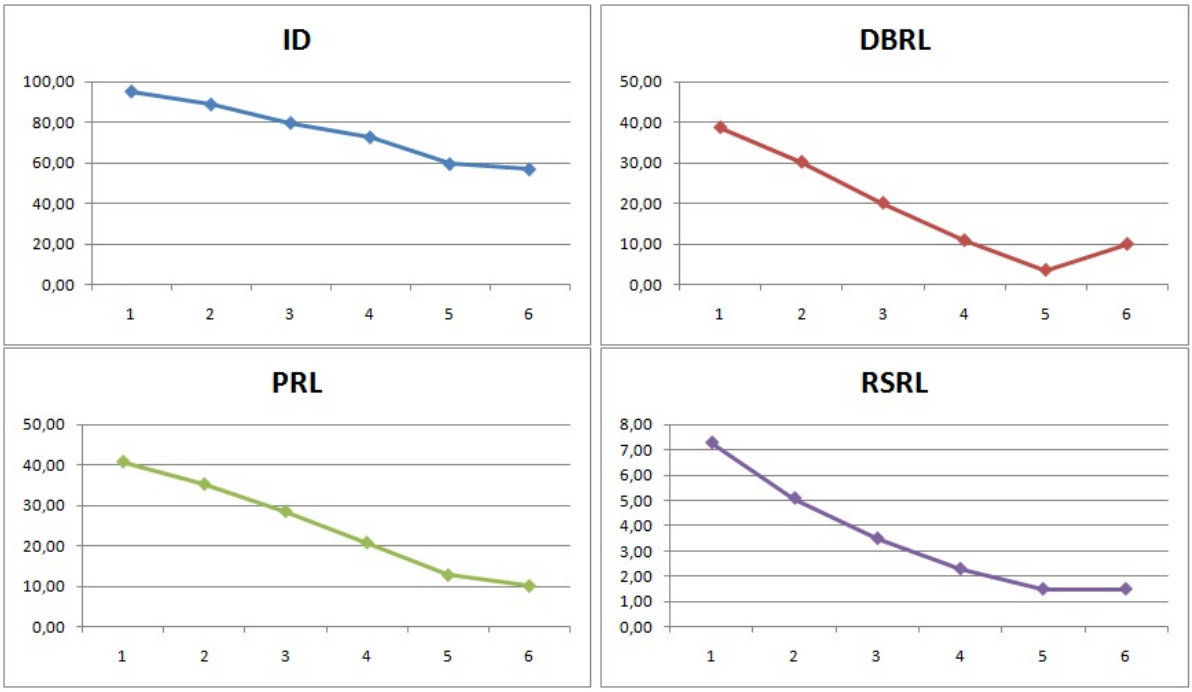


- *RISC DE REVELACIÓ*

  - *EXPERIMENTS*

| P | ID | DBRL | PRL | RSRL | ADR |
|---|---|---|---|---|---|
| 1 | 95,13 | 38,70 | 40,70 | 7,30 | 67,917 |
| 2 | 88,80 | 30,20 | 35,20 | 5,10 | 62,000 |
| 3 | 79,80 | 20,10 | 28,50 | 3,50 | 54,150 |
| 4 | 72,73 | 10,90 | 20,80 | 2,30 | 46,767 |
| 5 | 59,68 | 3,60 | 12,80 | 1,50 | 36,283 |
| 6 | 57,10 | 10,00 | 10,10 | 1,50 | 33,600 |

  - *GRÀFICS*



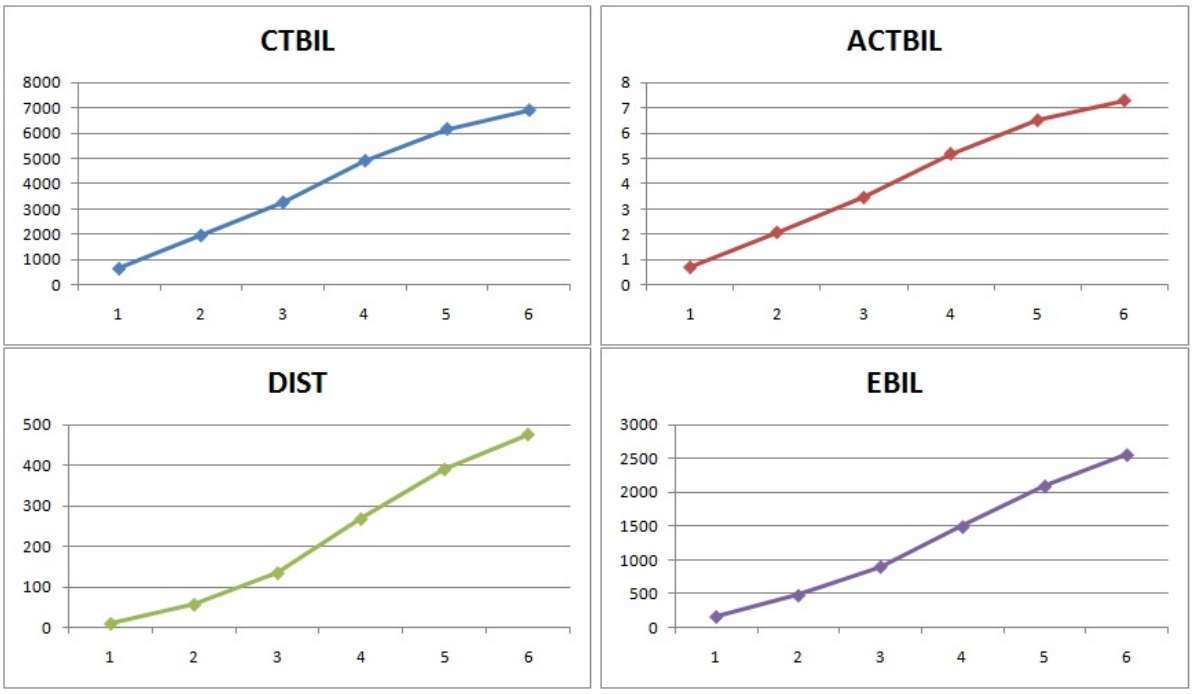
# ANNEX VI

## RESULTATS PARCIALS DELS EXPERIMENTS CORRESPONENTS AL BOTTOM CODING

- *PÈRDUA D'INFORMACIÓ*
  - *EXPERIMENTS*

| P | CTBIL | ACTBIL | DIST | EBIL | AIL |
|---|---|---|---|---|---|
| 1 | 654 | 0,691 | 11,25 | 162,2 | 1,992 |
| 2 | 1960 | 2,069 | 57,95 | 477,22 | 7,097 |
| 3 | 3258 | 3,44 | 135,25 | 901,74 | 14,519 |
| 4 | 4896 | 5,17 | 267,88 | 1493,14 | 26,027 |
| 5 | 6148 | 6,492 | 389,24 | 2093,59 | 36,909 |
| 6 | 6886 | 7,271 | 474,98 | 2557,32 | 44,837 |

  - *GRÀFICS*

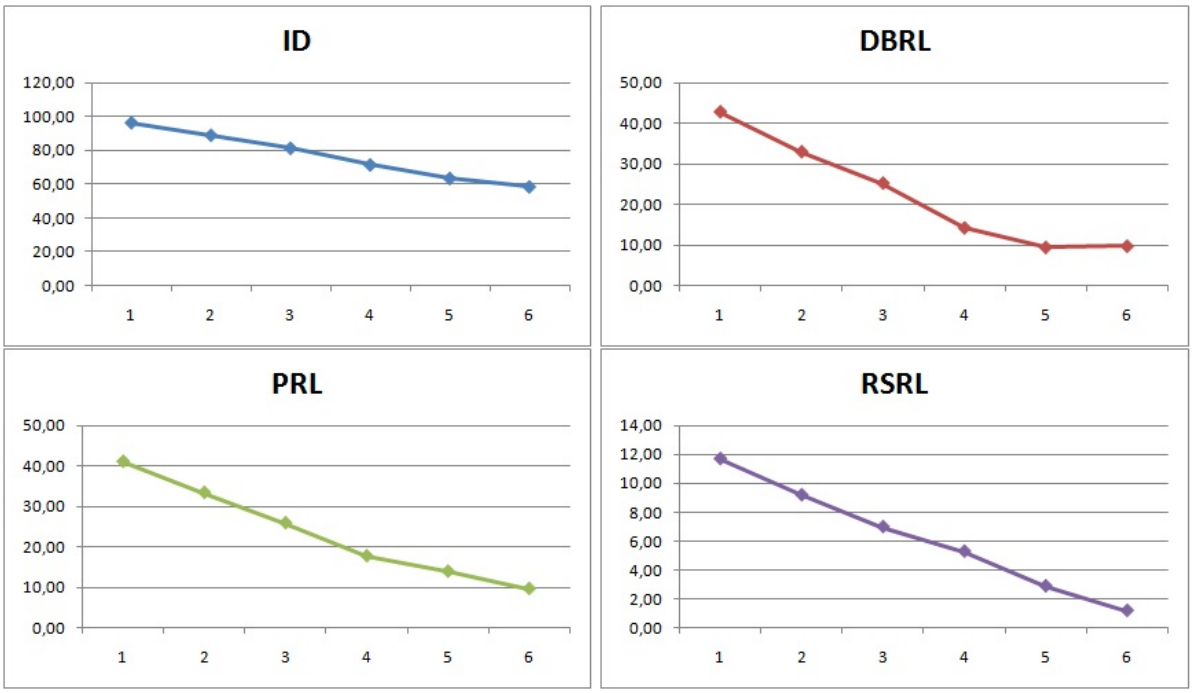


- *RISC DE REVELACIÓ*

  - *EXPERIMENTS*

| P | ID | DBRL | PRL | RSRL | ADR |
|---|---|---|---|---|---|
| 1 | 96,37 | 42,90 | 41,20 | 11,70 | 69,633 |
| 2 | 89,00 | 33,10 | 33,50 | 9,20 | 61,250 |
| 3 | 81,37 | 25,40 | 26,00 | 7,00 | 53,683 |
| 4 | 71,53 | 14,40 | 17,80 | 5,30 | 44,667 |
| 5 | 63,57 | 9,60 | 14,00 | 2,90 | 39,783 |
| 6 | 58,60 | 9,90 | 9,70 | 1,20 | 34,250 |

  - *GRÀFICS*



Signat: Jordi Marés Soler

Bellaterra, 17 de juny de 2010

**RESUM**

Degut a l'expansió de la nostra societat cada dia hi ha més fonts de dades públiques (mèdiques, financeres, ...) per a realitzar-hi estudis estadístics. Aquestes fonts de dades són perilloses per a la informació confidencial de les persones o institucions ja que són accessibles per a tothom, per tant necessiten ser protegides abans de ser publicades. En aquest projecte es presenten els diferents mètodes de protecció corresponents a dades categòriques així com un anàlisi de cadascun per a determinar-ne la pèrdua d'informació i el risc de revelació. Finalment també s'ha desenvolupat un mètode per optimitzar els resultats obtinguts pel mètode PRAM.

**RESUMEN**

Debido a la expansión de nuestra sociedad, cada vez hay mas fuentes de datos públicos (medicas, financieras, ...) para realizar estudios estadísticos. Éstas fuentes de datos son peligrosas para la información confidencial de las personas o instituciones ya que son accesibles por todo el mundo, por eso deben ser protegidas antes de ser publicadas. En este proyecto se presentan los diferentes métodos de protección correspondientes a datos categóricos así como un análisis de cada uno para determinar su pérdida de información y su riesgo de revelación. Finalmente también se ha desarrollado un método para optimizar los resultados del método PRAM.

**ABSTRACT**

Due to the expansion of our society, every day appear more and more public data sources (medical, financial, ...) for statistical studies. Those data sources are dangerous for confidential information about persons or institutions because they are accessible for everybody, so they need to be protected before publish. This project presents the different protection methods for categorical data as well as the analysis of everyone to determine their information loss and disclosure risk. Finally a new method for the optimization of PRAM results has developed.