Estudi de mètodes de generació de dades sintètiques

Isaac Cano Franco
ETSE-UAB

22 de juny de 2009

Sumari

Índex de taules

Índex de figures

Agraïments

Al Dr. Vicenç Torra, per haver-me donat la possibilitat de realitzar el projecte final de carrera sota la seva direcció i per haver estat en tot moment atent al treball realitzat.

A l'Institut d'Investigació en Intel·ligència Artificial (IIIA) per proporcionar-me tots els mitjans necessaris per a la realització d'aquest estudi. Especial agraïment a en Dani per proporcionar-me accés al cafè i pels seus acudits amb dubtosa gràcia.

Agrair també als companys de feina de la remodelada sala *polivalent* per estar sempre disponibles a les meves preguntes i per respectar sempre l'hora del te.

Finalment agrair a tota la meva família i amics, ja que sense ells no hauria arribat aquest dia, el qual representa el final d'una etapa molt important de la meva vida. El meu agraïment més profund per a la dona que m'ha fet costat durant aquests anys, Mireia.

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

Resum

Aquest projecte presenta un estudi científic dels mètodes de generació de dades sintètiques dins de l'àrea de la privadesa de dades. Aquests mètodes permeten controlar la transferència de dades sensibles a terceres parts i la utilitat estadística de les dades que es generen sintèticament. S'han introduït tots els conceptes bàsics necessaris per a situar al lector i s'ha analitzat un dels mètodes existents més amplament utilitzat (IPSO). Seguidament, s'ha proposat un nou mètode per a la generació de dades sintètiques (FCRM) que es basa en *Fuzzy c-Regression* i permet controlar l'equilibri entre pèrdua d'informació i risc de revelació mitjançant un paràmetre *c*.

1 Introducció

1.1 Motivacions

La gran majoria de les organitzacions, tant privades com públiques, obtenen i posteriorment emmagatzemen dades sobre individus i d'altres entitats en les seves bases de dades. En alguns casos, aquestes dades contenen informació sensible i per tant cal preservar la seva privadesa tant per motius ètics com legals, especialment si es vol donar accés a les dades a terceres parts per a possibles estudis estadístics.

Principalment, sense menysprear els motius morals o ètics, la legalitat actual és especialment estricta a Espanya, encara que també arreu del món però amb diferents nivells de compliment. Dins de la legislació Espanyola trobem la *Llei Orgànica 15/1999 del 13 de Desembre de Protecció de Dades de Caràcter Personal (LOPD)*, que té per objectiu garantir i protegir, en el que respecta al tractament de dades personals, les llibertats públiques i els drets fonamentals de les persones físiques i especialment el seu honor, intimitat i privacitat personal i familiar. També existeixen Declaracions Universals, com l'*Article 12 de la Declaració Universal dels Drets Humans*, adoptat per l'*Assemblea General de les Nacions Unides* i que estableix que el dret a la privadesa és un dret universal de la humanitat.

A causa d'aquestes obligacions legals, les organitzacions utilitzen diferents mètodes per prevenir la revelació de la informació, especialment sensible a atacs de privadesa. Un dels diferents controls que avui dia s'apliquen és la restricció d'accés a les dades personals d'individus concrets, només fent públics alguns estadístics considerats suficients com són la covariància o la mitjana. També trobem tècniques de control d'accés a les bases de dades i mètodes que emmascaren les dades originals abans d'alliberar-les a terceres parts.

Entre tots els mètodes d'emmascarament de dades, els pertorbatius s'han guanyat un racó especialment important dins de la literatura actual. La pertorbació de dades implica modificar les dades originals especialment sensibles mitjançant soroll, que pot ésser simplement aleatori. Tant si les dades són categòriques, en el cas d'atributs com *estudiant(Si/No)* o *lloc de naixement*, o si aquestes són numèriques, per exemple *edat* o *salari*, la pertorbació de dades és aplicable. En aquest projecte final de carrera (PFC) tractaré amb dades numèriques i contínues, com en la majoria d'aplicacions de pertorbació de dades.

Un cop afegit soroll a les dades originals, el procés és irreversible, és a dir, no podem obtenir les dades originals només coneixent les pertorbades. Ens faria falta conèixer el soroll i com aquest pot haver estat generat de forma aleatòria, el problema es trasllada a conèixer aquest terme aleatori. En canvi, si disposem d'una part de les dades originals, podem realitzar atacs que permeten enllaçar (*Record Linkage*) cada dada protegida amb la seva corresponent dada original. En-

tre aquests atacs trobem: *Probabilistic Record Linkage*, *Distance Based Record Linkage*, *Rank Swapping Record Linkage*, *Microaggregation Reidentification* i *Interval Disclosure* [11, 21].

S'ha realitzat molta recerca en l'àrea de la pertorbació de dades, algunes de les contribucions més importants pertanyen a Fienberg [13], Fuller [14] o Rubin [23]. L'àrea d'estudi que es vol tractar en aquest PFC és la pertorbació de dades a través de la generació de dades sintètiques. En aquest àmbit, i tractant-se de dades numèriques, un dels principals estudis que emprarem en aquest PFC, és el publicat per J. Burridge [3]. En el seu estudi, J. Burridge proposa un mètode de pertorbació de dades anomenat *Information Preserving Statistical Obfuscation*, que permet la protecció de dades, generant-ne de sintètiques, mantenint el vector de mitjanes i la matriu de covariàncies igual per a les dades originals i les sintètiques. Aquest mètode es diu que genera dades sintètiques perquè en certa manera, les dades pertorbades no són generades a través d'una funció directa de les originals. El vector de mitjanes i la matriu de covariàncies estan considerats com a estadístics suficients que s'han de mantenir per a que les dades pertorbades continuïn essent útils des del punt de vista estadístic.

Una de les principals idees subjacents al mètode de J.Burridge és la predicció de les dades pertorbades a través de rectes de regressió més alguns termes que s'encarreguen de corregir l'error de predicció, mantenint d'aquesta forma, els estadístics considerats suficients. Aquesta idea també és present en l'estudi de Richard J. Hathway i James C. Bezdek anomenat *Switching Regression Models and Fuzzy Clustering* [18]. Aquest mètode també permet generar dades sintètiques, però en aquest cas, es pretén pertorbar les dades sensibles classificant-les en *clústers*. Un cop classificades es procedeix a predir els valor sintètics mitjançant la recta de regressió que millor aproxima les dades per a cada clúster. Aquest mètode ofereix la possibilitat de parametritzar el nivell de pertorbació i per tant controlar la qualitat dels models de predicció.

En tots aquests mètodes es pretén mantenir un equilibri entre la utilitat de les dades i el risc de revelació. Quan no s'alliberen les dades originals a terceres parts sinó dades aleatòries, clarament no existeix cap tipus de risc de revelació i la privadesa de les dades es manté assegurada al cent per cent, en canvi, les dades no tenen cap utilitat perquè difereixen completament de les dades reals. D'altra banda, els mètodes pertorbatius permeten a les organitzacions facilitar les dades pertorbades a governs, investigadors i el públic en general per motius estadístics, mantenint el risc de revelació i la pèrdua d'informació sota un equilibri acceptable. Per tant, un dels principals objectius dels mètodes pertorbatius és proveir accés a les dades maximitzant la seva utilitat i minimitzant el risc de revelació. Caldrà llavors, avaluar els mètodes pertorbatius segons l'equilibri que cadascun aconsegueix entre aquest dos factors, o el que és el mateix, segons un *score*.

Per mesurar la utilitat de les dades, una possibilitat és mesurar la pèrdua d'informació que es produeix entre els estadístics de les dades pertorbades i els de les originals. Un dels principals estudis que pretén proporcionar una mesura d'aquesta pèrdua d'informació és el realitzat per J. M. Mateo Sanz, J. Domingo Ferrer i F. Sebé, anomenat *Probabilistic Information Loss Measures in Confidentiality Protection of Continuous Microdata* [15] .

1.2 Aportacions

L'anàlisi i la investigació realitzada en aquest projecte final de carrera contribueix en dos aspectes diferents.

Primerament, aquest estudi contribueix a les disciplines de *Statistical Disclosure Control* (SDC) i *Privacy Preserving Data Mining* (PPDM) amb la introducció d'una tècnica de mineria de dades, *Fuzzy c-Regression Models* (FCRM), per a l'anonimització de dades estadístiques que contenen informació confidencial sobre individus, tals com poden ser persones o entitats. Aquest estudi també dona a conèixer els conceptes bàsics subjacents als mètodes de generació de dades sintètiques estudiats (IPSO-A, IPSO-B i IPSO-C) i proposats (FCRM).

La segona contribució d'aquest estudi s'ha centrat en l'àrea dels mètodes de protecció de dades. Mitjançant la creació d'una nova tècnica per a la generació de dades sintètiques, com és el cas de FCRM, es preserva la utilitat estadística de les dades així com la privacitat de forma parametritzada segons els requisits de les terceres parts interessades en les dades. La creació d'una nova tècnica de protecció de dades també comporta la seva avaluació en termes de pèrdua d'informació i risc de revelació.

1.3 Estructura del document

Aquest document conté sis capítols/seccions i està organitzat en tes parts: la primera part agrupa la introducció i els preliminars (capítols 1 i 2), la segona part conté les aportacions d'aquest estudi (capítols 3, 4 i 5) i finalment les conclusions i el possible treball futur (capítol 6).

2 Preliminars

En aquesta secció es pretén introduir els conceptes bàsics que sorgeixen quan es treballa en la protecció de dades estadístiques així com un sumari de les tècniques que existeixen per protegir dades confidencials. Aquestes tècniques agrupen o pertorben les dades amb l'objectiu de reduir el risc d'identificar individus. El repte que es planteja en el problema de la confidencialitat és protegir les dades individuals sense degradar-ne excessivament la validesa analítica.

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

A la secció 2.1 es distingeixen i es defineixen els tipus de dades, macrodades i microdades, en els que es poden classificar les dades recollides a individus concrets. A la secció 2.2 es proporciona una descripció i classificació dels mètodes de protecció de microdades. En la següent secció 2.3 es descriuen les dades utilitzades per a l'avaluació dels mètodes de generació de dades sintètiques analitzats i proposat. En la secció 2.4 es defineixen i especifiquen els mètodes que s'utilitzaran per a mesurar la pèrdua d'informació i del risc de revelació de les dades generades sintéticament. Finalment s'introdueixen conceptes necessaris com són els mètodes de regressió, secció 2.5, la inversa generalitzada, secció 2.6, i el mètode de classificació difusa *Fuzzy c-means*, secció 2.7.

2.1 Classificació de les Dades

Els conjunts de dades que tractarem estan formats per variables i els valors d'aquestes corresponents a cada individus. Les variables poden ser qualitatives o quantitatives. Les variables *qualitatives* es poden classificar en:

- Qualitatives Nominals (o Categòriques), com per exemple la variable *sexe*.
- Qualitatives Ordinals, com per exemple una classificació segons les expressions *"baix, mitja, alt"*.

Les variables *quantitatives* poden ser mesurades en una escala mètrica (o numèrica), com per exemple les variables *edat* o *ingressos*.

Per agrupar diferents variables en estadístics i ser tractats pels diversos mètodes de control de la revelació existeix una tipologia proposada per Dalenius al 1988 [4]. Els estadístics es classifiquen segons:

- El seu format: en macrodades i microdades.
- La seva forma: en freqüències i magnituds.
- el mitjà emprat per publicar: en impremta i base de dades o altres mitjans.

En aquest estudi ens basarem en la primera classificació que s'ha fet dels estadístics que volem publicar, és a dir segons el seu format:

- Les **microdades** es defineixen segons Willenborg i De Waal [31] com un conjunt de registres sobre dades d'individus, els quals poden ser persones, empreses, companyies, etc...És a dir, les microdades consisteixen en la informació al nivell dels subjectes que responen. Tota la informació d'aquests subjectes ha de ser tractada com a confidencial i ha de ser protegida contra la revelació. Per a cada subjecte j tenim un vector individual de dades

Edat	Sexe	Estat civil	Nombre de fills	Ingressos mensuals (en euros)	...
19	femení	solter	0	1200	...
26	masculí	casat	2	1350	...
63	femení	vidu	5	900	...
...	...	...	...	...	...

Taula 1: Exemple de microdades amb variables qualitatives i quantitatives.

V_j, també anomenat registre de dades, el qual pot tenir variables qualitatives i/o quantitatives. De les variables que formen part d'un conjunt de microdades en podem distingir tres tipus bàsics segons el seu grau de compromís respecte la privacitat dels individus:

- **Identificadors directes:** Són variables el coneixement de les quals provoca la identificació de manera unívoca de l'individu al qual pertanyen aquests identificadors. Exemples d'aquest tipus de variables, quan es treballa amb persones, poden ser el nom complet o el número de DNI.
- **Identificadors indirectes:** Són variables que poden servir per identificar l'individu al qual pertanyen, però no de manera unívoca. El que si pot succeir és que hi hagi una combinació inusual d'identificadors indirectes que puguin provocar la identificació de l'individu en qüestió. Identificadors indirectes poden ser l'edat, el sexe o l'estat civil, quan es treballa amb persones. Un exemple d'aquest tipus de variables pot ser professió, que no és un identificador directe, però si els individus consultats formen part d'una comunitat concreta i molt reduïda, pot identificar unívocament a un individu, per exemple el metge d'un poble.
- **Variables confidencials:** Aquestes variables pertanyen al domini privat dels individus i s'ha d'evitar que es puguin relacionar amb l'individu al qual pertanyen. El criteri per decidir si una variable és sensible o no pot variar segons els països: el que en un país és una variable sensible en altres no ho és i a l'inrevés. Exemples de variables sensibles poden ser el passat criminal o les malalties que pateixen o han patit les persones.

En la taula 1 es mostren exemples de microdades amb variables qualitatives categòriques, com *Sexe* o *Estat civil*, i variables quantitatives, com *Edat*, *Nombre de fills* o *Ingressos mensuals*.

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

	dretà	esquerrà	TOTAL
Homes	43	9	52
Dones	44	4	48
TOTAL	87	13	100

Taula 2: Exemple de macrodades, taula de contingència.

Producció	Ind. X	Ind. Y	Ind. Z	Total
País 1	450	700	1050	2200
País 2	120	330	300	750
País 3	230	270	550	1050
Total	800	1300	1900	4000

Taula 3: Exemple de macrodades, taula de magnituds o agregats.

- Si per a cada cel·la es compta o s'estima el nombre d'elements que hi pertanyen, aleshores l'estadístic s'anomena **taula de contingència**. En la taula 2 es mostra un exemple de taula de contingència expressant la relació entre les variables *Sexe* (home o dona) i *despotisme* (dretà o esquerrà).
- Si s'agrega una variable quantitativa com *ingressos* o *producció* de tots els elements que pertanyen a una cel·la, aleshores anomenarem la taula resultant **taula de magnituds** o agregats. Un exemple de taula de magnituds es mostra a la taula 3.

L'ús d'ordinadors i paquets estadístics permet als usuaris fer avaluacions estadístiques i crear taules fetes a mida en lloc de rebre les anàlisis des d'un institut d'estadística. Aquesta tendència seguirà i, per tant, les oficines d'estadística experimentaran una demanda creixent, per part dels usuaris, de la difusió de fitxers de microdades en lloc de les anàlisis predefinides per les mateixes oficines. A causa de la seva mida, normalment les microdades es publiquen a través de bases de dades o altres fitxers. Per a les dues classes de macrodades (taules de contingència i de magnituds) hi ha dos camins de publicació: la publicació estadística impresa i les bases de dades o altres informàtics.

En aquest estudi es tractaran només **microdades** amb variables **quantitatives**, però els mètodes que s'estudiaran poden ser adaptats també a microdades amb variables categòriques.

2.2 Mètodes de Protecció de Microdades

Tal i com hem vist a la secció 2.1, una base de dades amb microdades X es pot veure com una matriu amb n files (registres/individus) i k columnes (variables/atributs). Cada fila conté, llavors, els valors dels atributs de cada individu.

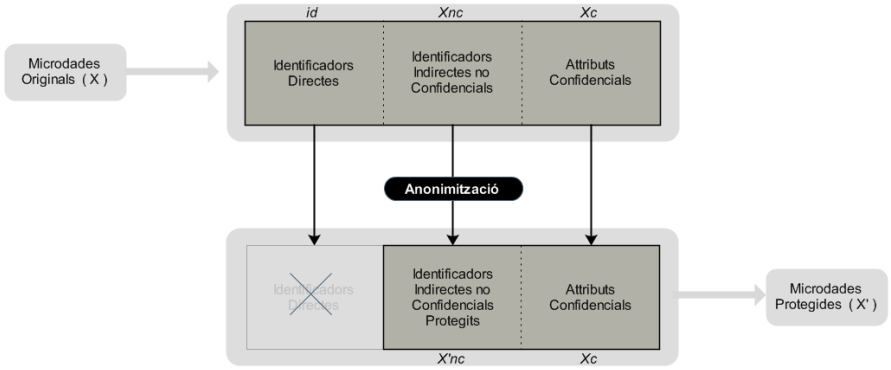

Figura 1: Procés de protecció de microdades

Considerant la classificació proposada en la secció 2.1 per a les variables, una base de dades X queda definida com $X = id||X_{nc}||X_c$, on id són les variables considerades identificadors directes, X_{nc} són els atributs no confidencials considerats identificadors indirectes i X_c són les variables confidencials. Normalment, abans d'alliberar X a terceres parts amb atributs confidencials, s'aplica primer un mètode de protecció ρ, passant a ésser microdades protegides X'. De fet, és usual assumir el següent escenari típic:

1. Els atributs considerats identificadors directes en X han estat eliminats o bé encriptats, per tant X passa a estar definida com $X = X_{nc}||X_c$.

2. Els atributs confidencials X_c no són modificats, per tant tenim que $X'_c = X_c$.

3. El mètode de protecció ρ és aplicat als atributs no confidencials considerats identificadors indirectes, per a preservar la privacitat dels individus dels quals s'està alliberant dades confidencials. Per tant, tenim que $X'_{nc} = \rho(X_{nc})$.

Aquest escenari permet a terceres parts tenir informació precisa de la informació confidencial sense revelar a qui pertany aquesta informació confidencial. La figura 1 mostra el procés de protecció i alliberació de microdades segons l'escenari típic considerat.
En aquest escenari, tal i com es mostra en la figura 2, un intrús podria intentar re-identificar individus obtenint identificadors indirectes no confidencials (X_{nc}) així com identificadors (Id) d'una altre font de dades. Llavors, aplicant tècniques de vinculació de registres *(Record Linkage)* entre els atributs protegits (X'_{nc}) i els mateixos atributs obtinguts a partir d'una altra font de dades (X_{nc}), l'intrús podria ser capaç de re-identificar un cert percentatge dels registres protegits juntament amb els corresponents atributs confidencials (X_c). Això és el que els mètodes de protecció intenten evitar. Aquest escenari de re-identificació serà considerat quan analitzem el risc de revelació dels diferents mètodes de protecció de dades i és si-

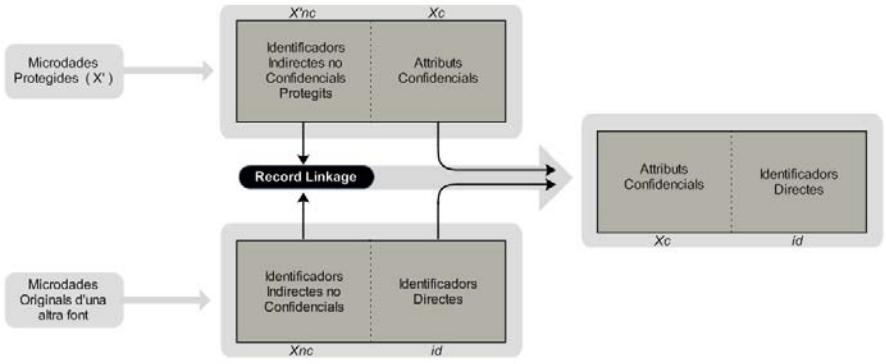

Figura 2: Escenari típic de Risc de Revelació.

milar als considerats en [26, 30].

Els mètodes de protecció poden ser classificats segons els seus efectes sobre les dades originals en tres categories diferents:

- **Pertorbatius.** El conjunt de dades es veu pertorbat afegint-li algun tipus de soroll, per exemple afegint soroll a les variables seguint una $N(0, a)$ per a una a determinada. D'aquesta forma, en el conjunt de dades originals, les combinacions de valors que sense ambigüitat identifiquen un individu (o registre) desapareixen i llavors, noves combinacions apareixen en el conjunt de dades protegides. Aquesta ofuscació fa difícil per a un atacant obtenir els valors del conjunt de dades originals. Un mètode pertorbatiu ha d'assegurar que la informació estadística del conjunt de dades originals es preserva en el conjunt de dades protegides. Exemple de mètodes que estan classificats com pertorbatius són *Rank Swapping* [16] i la *Microaggregació* [10].

- **No Pertorbatius.** Els mètodes classificats com a no pertorbatius no distorsionen el conjunt de dades originals sinó que modifiquen els valors originals per altres menys específics, per exemple reemplaçant un nombre real per un interval. També poden fer supressions parcials. Per tant, el procés de re-identificació és més difícil. En general, els mètodes no pertorbatius redueixen el nivell de detall de les dades, el que comporta una major pèrdua d'informació però en canvi un risc de revelació menor.

- **Generadors de Dades Sintètiques.** En aquest cas, es generen noves dades artificials que s'utilitzen per substituir les dades originals. Formalment, els generadors de dades sintètiques creen un model a partir de les dades originals que s'utilitza per a generar dades protegides de forma aleatòria però sotmeses al model creat. En aquest estudi analitzarem dos mètodes de protecció de dades considerats dins d'aquesta categoria, *Information Preserving*

Statistical Obfuscation (IPSO) [3] i un mètode que hem definit basat en la *Fuzzy c-Regression* [18].

Els mètodes de protecció de dades també poden ser classificats segons el tipus de dades que són capaços de suportar:

- **Numèrics.** Un atribut es considera numèric si es poden realitzar operacions aritmètics amb ell, per exemple $edat$ o $ingressos$. Cal dir que per a ser considerat numèric, un atribut no necessàriament ha de tenir un rang infinit, com és el cas de $númerodefills$. Quan es dissenyen mètodes per a la protecció de dades numèriques, es compta amb l'avantatge de que es poden realitzar operacions aritmètiques i amb el desavantatge de que cada combinació de valors numèrics en les dades originals és única, la qual cosa significa risc de revelació si no es realitza cap acció. Aquest estudi es centra en l'estudi de mètodes per a dades numèriques.

- **Categòrics.** Un atribut es considera categòric quan pren valors sobre un conjunt finit i quan les operacions aritmètiques no tenen sentit. Dins dels atributs categòrics es poden distingir entre *ordinals* i *nominals*. Quan es treballa amb l'escala ordinal, l'ordre entre valors és rellevant, per exemple $nivelld'estudis$, mentre que per a escales nominals l'ordre entre valors no ho és, per exemple $colord'ulls$. En l'escala ordinal, operacions com mínim o màxim tenen sentit, mentre que en l'escala nominal només tenen sentit les comparacions entre parelles. Quan es dissenyen mètodes per a la protecció de dades categòriques, la impossibilitat de realitzar operacions aritmètiques és un inconvenient, però en canvi, el fet de que les dades siguin finites és una propietat interessant que cal tenir en compte.

2.3 Dades Census

En aquesta secció, es descriu en detall el conjunt de dades utilitzades en els experiments, secció 5. El conjunt de dades considerat és un dels dos conjunts de dades emprats en el projecte Europeu CASC (Computational Aspects of Statistical Confidentiality) [2] com a conjunts de dades de referència per a la prova i comparació de mètodes per a la protecció de microdades numèriques. Aquest conjunt de dades de referència ha estat utilitzat en treballs com [5, 6, 8, 9, 7, 19, 34, 11, 26, 33].
Estem parlant del conjunt de dades anomenat $Census$, el qual conté un total de 1080 registres i 13 atributs numèrics per cada registre. Aquest conjunt de dades es va recopilar mitjançant *The Data Extraction System of the U.S. Census Bureau* [28].

El conjunt de dades utilitzat per a crear $Census$ ha sigut extret del grup de fitxers *March Questionnaire Supplement - Person Data Files* corresponents a la font de dades *Current Population Survey of the year 1995*. No tots els registres d'aquesta enquesta han estat considerats. Els registres sense valors o bé valors nuls per a com a mínim algun dels 13 atributs han estat descartats per tal d'obtenir els

Atribut	Nom	Descripció
a1	AFNLWGT	Pes final (es consideren 2 decimals)
a2	AGI	Ingrés brut ajustat
a3	EMCONTRB	Contribució a l'assegurança de salut de l'empleat
a4	ERNVAL	Guanys nets per negocis o producció agrària en 1995
a5	FEDTAX	Impost federal sobre la renda
a6	FICA	Deducció de nòmina per a la S.S. en concepte de pensió
a7	INTVAL	Quantitat d'ingressos per concepte d'interessos
a8	PEARNVAL	Total de guanys de la persona
a9	POTHVAL	Total d'ingressos d'altres persones
a10	PTOTVAL	Total d'ingressos de la persona
a11	STATETAX	Impostos estatals
a12	TAXINC	Impost estatal per els ingressos
a13	WSALVAL	Quantitat: salari i sou total

Taula 4: Atributs del conjunt de dades $Census$. En la primera columna trobem l'identificador de cada atribut, en la segona columna trobem el nom que rep cada atribut en el conjunt de dades originals de on es van extreure aquests atributs i l'última columna mostra una petita descripció del significat de cada atribut.

1080 registres finals. El motiu per haver escollit 1080 com a nombre de registres és que 1080 és el primer enter menor que 1200 i múltiple de 2,5,8 i 9, ja que aquests valors normalment són escollits com a valors de k en mètodes de protecció de dades com la Micro-agregació.
La taula 4 mostra la descripció dels 13 atributs escollits.

2.4 Pèrdua d'Informació i Risc de revelació

Aquesta secció pretén descriure les mesures emprades per a analitzar la pèrdua d'informació i el risc de revelació del mètode proposat en aquest PFC per a la generació de dades sintètiques. Aquestes mesures ens són d'utilitat en els experiments (secció 5). Mitjançant les mesures presentades a continuació, podem analitzar el comportament del mètode proposat segons els valors dels seus paràmetres així com comparar-ho de forma analítica amb altres mètodes, com és el cas de l'IPSO, agregant la pèrdua d'informació i el risc de revelació en una mesura anomenada $SCORE$. La pèrdua d'informació serà mesurada mitjançant la seva pèrdua d'informació probabilística, Probabilistic Information Loss (PIL), mètode aplicable de forma genèrica a qualsevol tècnica de control de la revelació estadística. En canvi, el risc de revelació (DR) és més difícil d'analitzar de forma general, per tant es basa en experiments de re-identificació mitjançant algoritmes que intenten enllaçar els registres protegits amb els seus corresponents registres originals ($Record - Linkage$).

El principal objectiu de qualsevol mètode de protecció de dades és minimitzar

tant la pèrdua d'informació (IL) com el risc de revelació (DR) de les dades protegides. Tanmateix, quan un d'aquests paràmetres disminueix, l'altre augmenta. Trobar la combinació òptima d'aquests dos paràmetres és una tasca realment difícil. A més, en algunes situacions, l'organització interessada en les dades protegides pot desitjar un determinat nivell d'alguns dels dos paràmetres. Per aquestes dues raons, és necessari calcular ambdues mesures d'una forma acurada abans d'alliberar les dades protegides, assegurant suficient protecció així com utilitat estadística.

2.4.1 Pèrdua d'informació (IL)

Estrictament parlant, l'avaluació de la pèrdua d'informació ha d'estar bassada en l'ús que es farà de les dades protegides. Com més gran sigui la diferència entre els resultats que es podrien obtenir amb les dades originals i les protegides, més gran serà la pèrdua d'informació. No obstant, molt sovint la protecció de microdades no es pot realitzar tenint en compte el seu ús posterior per les següents raons:

- La gran diversitat d'usos que es poden donar a les dades protegides, fins i tot resulta difícil identificar tots els possibles usos en el moment d'alliberar les dades protegides.
- Encara que tots els usos es poguessin identificar, oferir diferent versions de les dades, de forma que la versió número i tingués una pèrdua d'informació optimitzada per a l'ús i, resultaria en un risc de revelació inesperat.

Partint llavors, de que les dades s'haurien de poder protegir sense tenir en compte cap ús específic, és desitjable disposar de mesures genèriques per a la pèrdua d'informació. Aquestes mesures genèriques, haurien de ser capaces de capturar la quantitat de pèrdua d'informació per a un rang suficientment gran de dades. Per tant, direm que hi ha una pèrdua d'informació petita si les dades protegides són analíticament vàlides així com interessants segons les definicions que va proposar Winkler l'any 1999 [32]:

- Un conjunt de dades protegides és analíticament vàlid si preserva, de forma aproximada, els següents estadístics respecte les dades originals (algunes condicions són només aplicables a atributs continus):
  1. Les mitjanes i les covariàncies en un conjunt petit de subdominis (subconjunts de registres i/o atributs).
  2. Valors marginals per algunes tabulacions de les dades.
  3. Com a mínim una distribució característica.
- Un conjunt de microdades és analíticament interessant si 6 atributs d'importants subdominis són proveïts i es podem validar analíticament.

El mètode que s'utilitzarà en aquest estudi per avaluar la pèrdua d'informació serà el conegut com *Probabilistic Information Loss (PIL)*[15]. Aquest mètode calcula la diferència mitja entre alguns estadístics calculats tant en les dades originals

com en les protegides. Per simplicitat i per no abordar temes fora dels objectius d'aquest estudi, només s'ofereix una descripció molt reduïda dels estadístics considerats en aquest mètode. Per aprofundir en les seves definicions formals consultar la referència bibliogràfica. Els 5 estadístics tinguts en compte són els següents.

- Mitjana ($PIL(m_1^0)$)
- Variància ($PIL(m_2)$)
- Covariància ($PIL(m_1^1)$)
- Correlació de Pearson ($PIL(r)$)
- Quantils ($PIL(Q)$)

Finalment la pèrdua d'informació probabilística mitja es calcula com:

$$PIL = 100 * \frac{PIL(Q)+PIL(m_1^0)+PIL(m_2)+PIL(m_1^1)+PIL(r)}{5}$$

2.4.2 Risc de revelació (DR)

Segons les intencions de l'atacant, es consideren dos tipus de mesures del risc de revelació. Primerament, es suposa que l'intrús posseeix les dades protegides així com també alguns atributs originals obtinguts a partir de fonts de dades de terceres parts, aquest escenari es descrit a la figura 2. Aquí, l'intrús està interessat en enllaçar les dades protegides amb les corresponents originals. Aquest risc pot ésser mesurat mitjançant *record linkage*:

- **Distance Based Record Linkage (DBRL):** Aquesta mesura es calcula considerant només el número d'atributs que es suposen coneguts per l'atacant, per exemple els tres, cinc o vuit primers atributs. El valor final d'aquesta mesura és el número promig de registres que l'atacant ha pogut enllaçar amb els originals utilitzant alguna mesura de distancia entre registres. A continuació es procedeix amb una descripció més detallada d'aquesta mesura: *Distance Based Record Linkage* consisteix en calcular distancies entre totes les parelles de registres originals/protegits. La parella que resti amb mínima distància és considerada com una *Linked Pair (LP)* o el que és el mateix, parella enllaçada. Per una altre banda la resta de parelles de registre no enllaçats es consideren *Not Linked Pairs (NP)*. Dins de l'àmbit de privadesa de dades, el primer cop que es va utilitzar aquesta mesura va ser a [22], aplicada en aquell cas a un altre mètode de protecció de dades com és la Micro-agregació. Si definim $d(a, b)$ com la distancia entre un registre del conjunt de dades originals X i un registre del conjunt de dades protegides X', llavors és pot implementar el *Distance Based Record Linkage* com s'expressa en l'Algorisme 1. Cal remarcar que aquest algorisme es podrà aplicar sempre i quan la distancia entre els atributs pugui ser definida. Normalment

aquesta distancia es defineix en termes de distancia d_{attr_i} per a cada atribut $attr_i$ de la següent forma:

$$d(a,b) = \sum_{i=1}^{n} d_{attr_i}(attr_i^A(a), attr_i^B(b))$$

La distancia d_{attr_i} permet millorar el número de parelles enllaçades donant més importància a certs atributs com també permet definir distàncies entre atributs qualitatius o quantitatius.

Algorithm 1: Distance Based Record Linkage

Data: X: conjunt de dades originals, X': conjunt de dades protegides
Result: LP: linked pairs
begin
  foreach $a \in X$ **do**
    $b' = arg_m in_{b \in X'} d(a,b)$ $LP = LP \cup (a,b')$ **foreach** $a \in X$ **do**
      $NP = NP \cup (a,b)$
end

- **Probabilistic Record Linkage (PRL):** Aquesta mesura del risc de revelació és idèntica en essència a l'anterior però considerant en comptes de distàncies, les probabilitats condicionals d'enllaçar els registres. Aquest mètode es va descriure per primera vegada en [12]. A continuació es defineix formalment aquest mètode: Per a cada parella de registres (a,b), on a representa un registre del conjunt de dades originals X i b representa un registre del conjunt de dades protegides X', es defineix un vector de coincidència $\gamma(a,b) = (\gamma_1(a,b) \dots \gamma_n(a,b))$, on $\gamma_i(a,b)$ pren el valor 1 si $attr_i(a) = attr_i(b)$ i 0 si $attr_i(a) \neq attr_i(b)$. Cal remarcar que els valors $attr_i$ estan normalitzats. El següent pas consisteix en calcular un índex del vector de coincidències. Finalment, s'utilitza aquest índex per classificar (utilitzant un llindar) les parelles de registres originals/protegits segons LP o NP depenent de si han estat enllaçades o no. La versió del *Probabilistic Record Linkage* emprada en aquest estudi utilitza probabilitats condicionals per calcular els índexs. A més, aquestes probabilitats són estimades mitjançant l'algorisme *Expectation Maximization (EM)*. Els llindars considerats es calculen a partir de: (1) $P(LP|\mathbf{U})$, la probabilitat d'enllaçar una parella que en realitat és una parella no enllaçada (*fals positiu* o *enllaçat fals*) i (2) $P(NP|\mathbf{M})$, la probabilitat de no enllaçar una parella de registres que en realitat és una parella enllaçada (*fals negatiu* o *no enllaçat fals*). Encara que *PRL* és computacionalment més complex que *DBRL*, aquest mètode probabilístic resulta interessant ja que l'usuari tant sols ha d'especificar dues probabilitats: un límit superior de la probabilitat d'un fals positiu i un límit superior de la probabilitat d'un fals negatiu. El que representa un avantatge del *PRL* en front del *DBRL*.

- **Interval Disclosure (ID):** Aquest mètode modela un escenari en el qual un intrús intenta obtenir una aproximació dels valors originals. En aquest cas l'atacant no estaria interessat en obtenir els valors originals exactes o bé no ha pogut obtenir-los. El risc que mesura *ID* es calcula com 100 vegades el percentatge promig de que un valor original caigui dins d'un interval definit al voltant del corresponent valor protegit. Aquest interval es defineix normalment entre un 1% i un 10% dels valors.

Finalment, el risc de revelació (DR) es calcula com un promig ponderat dels resultats obtinguts per els mètodes anteriors:

$$DR = 0.5 \cdot \frac{DBRL+PRL}{2} + 0.5 \cdot ID$$

2.4.3 Agregant IL i DR (SCORE)

La pèrdua d'informació (IL) i el risc de revelació (DR) poden ser combinats per tal d'obtenir un únic valor final per a un mètode de protecció específic. Aquest valor final, anomenat *SCORE*, representa l'equilibri entre *IL* i *DR*. El millor mètode de protecció serà aquell que faci òptim aquest equilibri. Es considera que un mètode de protecció comença a ser interessant si aconsegueix obtenir valors del *SCORE* iguals o inferiors a 30. En aquest estudi agregarem la pèrdua d'informació i el risc de revelació com una mitja aritmètica de la forma següent:

$$SCORE = 0.5 \cdot IL + 0.5 \cdot DR$$

2.5 Mètodes de Regressió

En aquesta secció es dona una descripció general dels mètodes de regressió i més en concret es descriu el funcionament bàsic dels models de regressió anomenats *Switching Regression Models*.

Considerem un conjunt de dades $S = (x_1, y_1), \cdots, (x_n, y_n)$ on cada observació independent $x_k \in \Re^s$ te la seva corresponent observació depenent $y_k \in \Re^t$. En el cas més senzill, podem assumir que tant sols existeix una única relació funcional entre x i y per a tot el conjunt de dades S. En la majoria dels casos és necessari tenir en compte l'error comés per tal de trobar una solució òptima. Normalment, trobar la solució òptima passa per haver de realitzar la cerca de la "millor"funció f que representa la relació entre observacions:

$$y = f(x; \beta) + \epsilon,$$

on $\beta \in \Omega \subset \Re^k$ és un vector de paràmetres per determinar, i ϵ és un vector aleatori amb el seu corresponent vector de mitjanes $\mu = 0 \in \Re^t$ i la seva corresponent matriu de covariàncies Σ. Trobar una estimació òptima per β depèn en les assumpcions fetes sobre la distribució de ϵ, i el conjunt de valors factibles que pot prendre β. Aquest és un tipus de model multivariant ben conegut i àmpliament utilitzat.

El tipus de model que considerarem en aquest estudi i que està estretament relacionat amb el mètode proposat, *Fuzzy c-Regression*, es coneix com a "switching regression model" [29]. En comptes d'assumir que un únic model representa totes les n parelles d'observacions (x_k, y_y) en S, ara s'assumeix que les dades segueixen un total de c models:

$$y = f_i(x; \beta_i) + \epsilon_i \text{ , per a tota } i \text{ tal que } 1 \leq i \leq c,$$

on cada $\beta_i \in \Omega_i \subset \Re^k$, i on cada ϵ_i és un vector aleatori amb vector de mitjanes $\mu_i = 0 \in \Re_t$ i matriu de covariàncies Σ_i. Igual que en el cas del model únic més simple, en aquest cas també es desitgen bones estimacions per als paràmetres $\beta_1, \cdots, \beta c$. Tanmateix, tenim la dificultat afegida que S no està etiquetat, és a dir, donat un conjunt d'observacions (X_k, y_k) no sabem quin model s'ha d'aplicar.
En aquest estudi es proposa el mètode *Fuzzy c-Regression* (secció 4.1), basat en classificació difusa, que produeix bones estimacions de $\beta_1, \cdots, \beta c$ al mateix temps que assigna un vector d'etiquetes difuses a cada observació en S.

2.6 Inversa Generalitzada

La necessitat de calcular la inversa generalitzada sorgeix en certs problemes d'estadística, matemàtiques o enginyeria en general. Aquests problemes poden ser des de l'estimació de funcions lineals de classificació o de regressió fins a estimacions de circuits elèctrics o càlcul d'estructures. La metodologia que aquest estudi ha seguit alhora de calcular la inversa generalitzada es basa en la descomposició ortogonal basada en el càlcul d'una submatriu no singular donada una certa matriu singular.

En el cas d'una matriu rectangular o una matriu de rang no complet, el concepte *matriu inversa* es pot generalitzar. Sota certes condicions, és possible definir una matriu que es comporti com la inversa d'una matriu donada $A \in M_{m\times n}$, anomenada inversa generalitzada de A. Per a més informació consultar [25]. Aquesta inversa generalitzada A^- o 1-inversa s'anomena també inversa generalitzada dèbil ja que tant sols ha de complir la següent condició per tal de ser considerada com inversa generalitzada d'una matriu A.

Definició. Donada una matriu $A \in M_{m\times n}$, la matriu $A^- \in M_{m\times n}$ és una 1-inversa o inversa generalitzada dèbil de A si i només si $AA\_A = A$.

No existeix una única matriu 1-inversa però com a mínim n'existeix una. Existeixen diferents formes de calcular aquesta matriu inversa generalitzada però en aquest estudi s'utilitza el següent algorisme:

- **Entrada:** Una matriu A de dimensions $m \times n$. Una matriu inicial $W = I_n$ (matriu identitat de dimensió n) i un vector inicial $U = O_n$ (vector nul de dimensió n).

- **Sortida:** El vector U que indica quines files i columnes de A formen la submatriu B_{11} i la matriu W que conté la inversa B_{11}^{-1}.
- **Pas 1: Inicialització.** Inicialitzar $W = I_n$, $U = O_n$ i $i = 1$.
- **Pas 2: Producte escalar.** Calcular el producte escalar $t_j = a_i^T w_j$ per a $j = 1, \cdots, n$. És a dir, el producte escalar de la fila número i de A amb cada columna de W.
- **Pas 3: Seleccionar una columna per pivotar.** Trobar el primer t_p diferent de zero corresponent a un element nul en la posició p de U, el qual indica la columna p de W que es farà servir com a pivot. Reemplaçar la component p de U per i. Si no existeix tal columna, anar al Pas 6, en cas contrari continuar al Pas 4.
- **Pas 4. Modificar la columna pivot.** Dividir la columna p de W per t_p.
- **Pas 5. Pivotatge.** Per a $j = 1$ fins a n, $j \neq p$ i $t_j \neq 0$ fer $w_{kj} = w_{kj} - t_j w_{kp}$ per a $k = 1, \cdots, n$.
- **Pas 6.** Si $i = m$ continuar amb el Pas 7, en altre cas, incrementar i en una unitat i tornar al Pas 2.
- **Pas 7. Sortida.** La matriu W conté la inversa d'una submatriu de rang màxim de A. Aquesta submatriu està composta per la intersecció de les files, els índexs de les quals es troben al vector U, i les columnes indicades per la posició d'aquests índexs en la matriu U. La inversa d'aquesta submatriu està composada per les columnes de W corresponents als elements no nuls de U i les mateixes files. El rang de A és igual al nombre de elements no nuls de U.

En aquest estudi, l'aplicació de la inversa generalitzada sorgeix quan es produeix singularitat en l'equació 2, dins del càlcul de *Fuzzy c-Regression* (secció 4.1)

2.7 Fuzzy c-Means

L'algorisme *fuzzy c-means* [1] és un dels mètodes més utilitzats per a realitzar classificació difusa (*fuzzy clustering*). Es basa en el concepte *fuzzy c-partition*, introduït per Ruspini l'any 1969 [24]. Donat un conjunt d'elements, *Fuzzy c-means* fa una partició difusa d'aquests. Des de un punt de vista conceptual, les categories subjacents de les dades es consideren difuses, és a dir, la pertinença de cada element a cada categoria és un valor continu dins de l'interval [0,1]. Per tant, donat un conjunt d'elements $X = \{x_1, x_2, \ldots, x_N\}$ avaluats en termes d'atributs $A = \{A_1, A_2, \ldots, A_M\}$, *Fuzzy c-means* fa una partició difusa dels objectes X. A més, es consideren c categories ($C = \{C_1, \ldots, C_c\}$) i per tant el problema consisteix en determinar c funcions de pertinença $\mu_1, \mu_2, \ldots, \mu_c$, on μ_i és la funció de

pertinença corresponent a C_i. Per a ser considerada com a funció de pertinença, μ_i ha de ser tal que per a cada objecte x la pertinença a les diferents categories C suma 1. Finalment, es requereix que com a mínim hi hagi un element amb una pertinença diferent de 0 per a cada categoria. Es poden formalitzar les anteriors restriccions sobre les funcions de pertinença com:

$$\sum_{i=1}^{c} \mu_i(x) = 1 \qquad \text{per a tota } x \in X$$

$$0 < \sum_{x \in X} \mu_i(x) < N \qquad \text{per a tota } C_i \in C$$

Un cop hem definides les restriccions per a les pertinences, el problema es pot formular de la forma següent:

$$\text{minimitzar} \qquad FO(\mu, P) = \sum_{k=1}^{c} \sum_{x} (\mu_k(x))^m \left\| A(x) - p_k \right\|^2$$

restringida a

$$\mu \in M_f = \left\{ (\mu_k(x)) \mid \mu_k(x) \in [0,1], \sum_{k=1}^{c} \mu_k(x) = 1, \forall x \in X \right\}$$

On c és un valor constant que representa el nombre de categories difuses permeses. L'altre valor constant m, el qual ha de ser major que 1, representa el grau de borrositat de les categories. Com més gran és el valor de m, més difuses són les categories. Quan $m \to \infty$, totes les categories cobreixen tots els punts. Per contra, com més petit és el valor de m, menys difuses són les categories. Quan $m = 1$ el problema és equivalent al de l'algorisme *c-means*.

L'algorisme *fuzzy c-means* construeix una solució factible pel problema anterior de la forma següent:

1. Definir una partició inicial μ i calcular els centroides P. Aquesta inicialització por ésser aleatòria.

2. El següent pas consisteix en, per a cada element x_i actualitzar la pertinença de x_i a cada categoria C_k de la forma següent:

   - Si $\left\| x_i - p_k \right\|^2 > 0$ llavors per a cada categoria C_k:

     $$\mu_k(x_i) = \left[\sum_{j=1..c} \left(\frac{\left\| x_i - p_k \right\|^2}{\left\| x_i - p_j \right\|^2} \right)^{\frac{1}{(m-1)}} \right]^{-1}$$

   - Si hi ha alguna categoria C_k per a la qual $\left\| x_i - p_k \right\|^2 = 0$ significa que x_i té el mateix valor que algun centroide p_k. Per tant, en aquest cas, la pertinença de x_i s'ha de compartir de forma aleatòria amb tots els centroides que tinguin igual valor que x_i.

3. L'objectiu del pas següent és actualitzar el valor dels centroides. En conseqüència, per a cada categoria C_k el seu centroide es defineix com:

$$p_k = \frac{\sum_{i=1}^{N}(\mu_k(x_i))^m A(x_i)}{\sum_{i=1}^{N}(\mu_k(x_i))^m}$$

i per la component *j*, es defineix com:

$$A_j(p_k) = \frac{\sum_{i=1}^{N}(\mu_k(x_i))^m A_j(x_i)}{\sum_{i=1}^{N}(\mu_k(x_i))^m}$$

4. Parar d'iterar quan superem el llindar de convergència establert, en altre cas anar al pas número 2. El llindar de convergència es pot definir com la comparació de les funcions de pertinença de dues iteracions consecutives. Considerant un llindar λ i també μ' i μ com les funcions de pertinença de dues iteracions consecutives, definim la condició per a finalitzar les iteracions de la forma següent:

$$\lambda > \max_{k=1,\ldots,c} \max_{x \in X} |\mu'_k(x) - \mu_k(x)|$$

3 IPSO

S'han proposat diversos mètodes per a la generació de dades sintètiques. Un d'ells és la família de mètodes anomenada *Information Preserving Statistical Obfuscation* (IPSO) [3]. L'objectiu de IPSO és enfosquir la identitat de les dades però també preservar alguns estadístics. Aquesta família està formada per tres mètodes IPSO-A, IPSO-B i IPSO-C, on IPSO-C és el mètode amb menor pèrdua d'informació i l'IPSO-A és el més simple de tots tres, és a dir, produeix més pèrdua d'informació. Els mètodes IPSO es basen en dividir les dades originals en conjunts d'atributs X i Y. Els atributs dins del conjunt X es consideren independents i els atributs dins del conjunt Y són considerats dependents. Llavors, és construeix un model de les dades, com pot ser un model de regressió múltiple multivariant, capaç de representar la informació continguda en Y i es genera un conjunt d'atributs protegits Y', seguint la distribució condicional $Y|(T, x)$. Com que el model es construeix a partir de les dades dependents Y, la protecció que s'aconsegueix depèn de les dades. Per tant, el grau de protecció és una funció directa de les dades i no es pot parametritzar per a obtenir un determinat equilibri entre la pèrdua d'informació i el risc de revelació. A continuació detallem les diferències entre IPSO-A, IPSO-B i IPSO-C:

- **IPSO-A.** Aquest mètode realitza una regressió múltiple de Y sobre X i els valors generats segons la regressió múltiple Y'_A són utilitzats per a substituir els valors originals de Y. Per tant, les dades protegides estan formades per els atributs de X i Y'_A en comptes del conjunt d'atributs originals X i Y.

Suposem que els atributs a Y segueixen una distribució normal multivariant amb matriu de covariàncies $\sum = \sigma_{jk}$ i vector de mitjanes $x_i B$, on B és la matriu dels coeficients de regressió. El desavantatge de l'IPSO-A és que, si una regressió múltiple és aplicada a (y'_a, x) obtindrem estimacions per a $\hat{B}_A$ i $\hat{\Sigma}_A$ que, en general, són diferents de les estimacions B i Σ que podem obtenir si apliquem el model a les dades originals (y, x).

- **IPSO-B.** El següent mètode IPSO, IPSO-B, millora y'_A de forma que el vector de mitjanes estimat $\hat{B}_B$ obtingut a partir de (y'_B, x) és igual a $\hat{B}$. Per tant, el nou valor y'_B es pot utilitzar com a valor pertorbat per a ésser alliberat juntament amb les dades públiques, preservant l'estadístic considerat suficient $\hat{B}$ i per tant la pèrdua d'informació disminueix.

- **IPSO-C.** L'últim dels tres mètodes IPSO, IPSO-C, novament millora y'_A i y'_B per tal de mantenir $\hat{B}_C = \hat{B}$ i també $\hat{\Sigma}_C = \hat{\Sigma}$. Això s'aconsegueix aplicant un model de regressió múltiple multivariant a (y'_c, x).

Els mètodes IPSO són similars a la classe de mètodes anomenats *General Additive Data Perturbation* (GADP) i que van ser descrits per primera vegada per Muralidhar and Sarathy [17]. Tots dos mètodes intenten mantenir certes característiques estadístiques de les dades originals com les mitjanes, variàncies o les covariàncies. La principal diferència entre GADP i IPSO és que aquest darrer manté els estadístics esmentats anteriorment fins i tot per al cas de conjunts de mostres de mida petita o mitjana.

4 FCRM

4.1 Fuzzy c-Regression

Els models *Fuzzy c-Regression* són una família de funcions objectiu que es poden utilitzar per a aplicar models de regressió *switching* tant a dades numèriques com contínues. Donat c (el número d'agrupacions o *clústers*, $1 < c < n$), l'algorisme *Fuzzy c-Regression* és capaç d'estimar els paràmetres de c diferents models de regressió, al mateix temps que realitza una c-partició difusa de les dades. En aquest cas considerem un conjunt de dades de mida n, $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, on cada vector de característiques (x_i, y_i) té una observació depenent $y_i \in \mathbb{R}^t$ corresponent a una certa observació independent $x_i \in \mathbb{R}^s$. La principal diferència entre els models de regressió *Fuzzy c-Regression* i els problemes més simples de regressió, és que aquests últims consideren una única relació funcional entre x i y per a totes les observacions, mentre que *Fuzzy c-Regression* considera que les dades segueixen c models diferents.

$$y = f_i(x; \beta_i) + \epsilon, \qquad 1 \leq i \leq c \tag{1}$$

on cada $\beta_i \in \Omega_i \subset \mathbb{R}^{k_i}$, i cada ϵ_i és un vector aleatori amb el corresponent vector de mitjanes $\mu_i = 0 \in \mathbb{R}^t$ i matriu de covariàncies Σ_i. Cal dir que S no està

etiquetat, és a dir, donat un vector de característiques (x_i, y_i), no se sap quin model lineal és el que cal aplicar-li. Hathaway i Bezdek van publicar en [18] una solució factible per aquest problema. La seva solució es basa en tècniques de classificació difusa i és capaç de produir bones estimacions de $\{\beta_1, \ldots, \beta_c\}$ al mateix temps que etiqueta de forma difusa les dades en S. El problema d'etiquetar S es resol llavors, mitjançant classificació difusa assignant vectors d'etiquetes amb restriccions que representen la pertinença de cada objecte (x_i, y_i) a cada classe c.

L'algorisme per a obtenir els *Fuzzy c-Regression Models* (FCRM) consta de passos similars als de *Fuzzy c-Means*:

1. **Pas 1.** Donat un conjunt de dades $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$. Establir $m > 1$ (un valor adequat, segons les nostres proves, és $m = 1.5$), especificar els models de regressió segons l'equació 1, i escollir una mesura de l'error $E = \{E_{ik}\}$ de forma que $E_{ik}(\beta_i) \geq 0$ per a i i k que satisfaci la propietat de minimització [18]. Establir un llindar on finalitzar les iteracions $\epsilon > 0$ (nosaltres hem escollit ϵ dins del rang $[0.0001, 0.00001]$ perquè normalment proporciona bones estimacions) i una partició inicial $U^{(0)} \in M_f$. En aquest estudi, s'ha utilitzat l'algorisme *fuzzy c-means* per obtenir aquesta partició inicial. Llavors establir un llindar r_{max}, número màxim d'iteracions, de forma que $r = 1, \ldots, r_{max}$ en cas que FCRM no convergeixi (en aquest estudi s'ha utilitzat $r_{max} = 30$).

2. **Pas 2.** Actualitzar els valors del paràmetres dels c models de regressió $\beta_i = \beta_i^{(r)}$ així com la mesura de l'error $E_{ik}(\beta_i)$ per a $f_i(x_k; \beta_i)$ que minimitza globalment (sobre $\Omega_1 \times \Omega_2 \times \cdots \times \Omega_c$) la funció restringida:

$$\psi(\beta_1, \ldots, \beta_c) \equiv E_m(U^{(r)}, \beta_1, \ldots, \beta c)$$

   Un exemple de mesura de l'error $E_{ik}(\beta_i)$ és la norma:

$$E_{ik}(\beta_i) = \|f_i(x_k; \beta_i) - y_k\|^2$$

   En aquest cas, aquest segon pas es pot especificar establint $\Omega_i = \mathbb{R}^s$, $f_i(x_k; \beta_i) = ((x_k)^T \beta i)$ i $1 \leq i \leq c$, per tant, la funció objectiu $E_m(U^{(r)}, \beta_1, \ldots, \beta c)$ passa a ser una extensió multi-model difusa del criteri habitualment utilitzat quan s'apliquen models de regressió, estimador per mínims quadrats:

$$E_{ik}(\beta_i) = (y_k - (x_k)^T \beta i)^2.$$

   Cal afegir que els nous valors per als paràmetres del model de regressió $\beta_i^{(r)}$, $1 \leq i \leq c$ es poden calcular mitjançant la següent formula, sempre i quan les columnes de X siguin linealment independents i $U_{ik}^{(r)} > 0$ per a $1 \leq k \leq n$, en altre cas caldrà aplicar la inversa generalitzada (secció 2.6):

$$\beta_i^{(r)} = \left[X^T D_i X\right]^{-1} X^T D_i Y \tag{2}$$

on X denota la matriu de $\mathbb{R}^{n\times s}$ que té x_k com a la seva columna k. Y representa el vector de $\mathbb{R}^n$ que té y_k com a la seva component k i D_i és la matriu diagonal de $\mathbb{R}^{n\times n}$ que té $(U_{ik}^{(r)})^m$ com al seu element diagonal k.

3. **Pas 3.** L'objectiu d'aquest pas és actualitzar $U^{(r)} \rightarrow U^{(r+1)} \in M_f$, interpretant U_{ik} com el pes o *l'importància* en quant el valor del model $f_i(x_k;\beta_i)$ és igual a y_k (pertinença difusa a cadascun dels c diferents models). Aquesta actualització es realitza seguint la fórmula següent:

$$U_{ik} = \left[\sum_{j=1}^{c}\left(\frac{E_{ik}}{E_{jk}}\right)^{\frac{1}{m-1}}\right]^{-1}, \text{ si } E_{ik} > 0 \text{ per a } 1 \leq i \leq c$$

   En el cas que trobem algun $E_{ik} = 0$, el seu valor es pot reemplaçar per algun número positiu molt petit (en aquest estudi hem utilitzat $E_{ik} = 10^{-100}$), per tant aquest tercer pas es pot realitzar de totes formes.

4. **Pas 4.** Aquest darrer pas és l'encarregat de comprovar si l'algorisme ha de finalitzar la seva execució. Si la diferència entre U^r i U^{r+1}, corresponents a dos iteracions consecutives, és major que el llindar de final de l'execució o r és més petit o igual que r_{max} (número màxim d'iteracions) llavors $r := r+1$ i anar al pas 2. En altre cas finalitzar l'execució de l'algorisme.

4.2 Utilitzant *Fuzzy c-Regression* per a generar dades sintètiques

Un cop introduïts tots els conceptes necessaris i relatius a aquest estudi, el pas següent és combinar la classificació difusa amb els models de regressió *switching* per tal de generar dades sintètiques a partir d'unes dades originals. En la secció anterior hem assenyalat les fórmules necessàries per a implementar els algorismes *Fuzzy c-Means* i *Fuzzy c-Regression model*, ara presentem els passos bàsics que s'han de seguir per tal de generar les dades sintètiques conservant al mateix temps la privacitat de les dades originals:

- El primer pas consisteix en diferenciar del conjunt de dades originals quins són els atributs independents X i quins els dependents Y.

- En el següent pas s'ha d'aplicar la *Fuzzy c-Regression (FCRM)* (secció 4.1), inicialitzant la partició difusa $U^{(0)} \in M_f$ mitjançant *Fuzzy c-Means* (secció 2.7).

- Un cop finalitza el FCRM, el proper pas és generar les dades sintètiques. Per a cada vector de característiques $(x_s, y_s) \in S$, $1 \leq s \leq n$, hem de seleccionar el *clúster* i amb màxima pertinença. És a dir, hem de seleccionar el $arg\ max_{i=1}^{c} U_{is}$. Un cop hem arribat a aquest punt sabem quin model de regressió β_i, corresponent al centroide p_i, utilitzarem per generar les dades sintètiques. És a dir, podem utilitzar el model de regressió β_i corresponent

de cada centroide p_i per tal de predir el valor sintètic $y_s^{'}$ que reemplaçarà el valor original y_s.

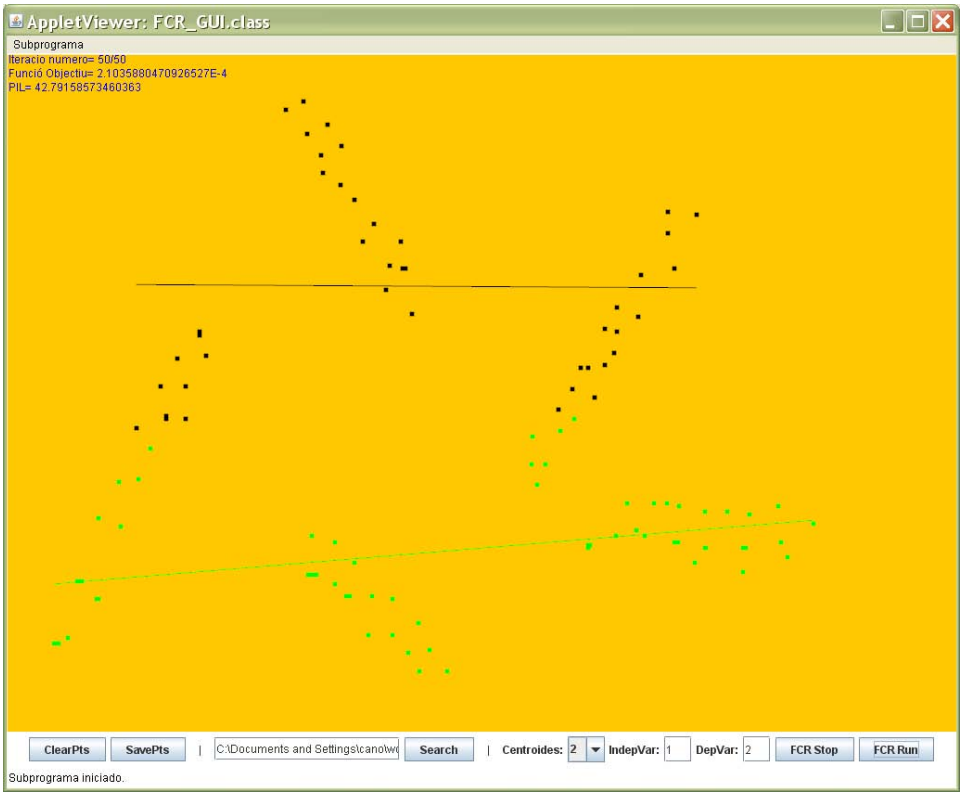

Figura 3: Estats finals dels models d'acord amb l'equació 1 amb $c = 2$ per a les dades de l'exemple.

4.3 Exemple

Arribats a aquest punt, es vol mostrar com s'utilitza FCRM amb un exemple petit. Considerarem dos atributs, el primer d'ells farà el paper d'atribut independent i el segon de depenent. D'aquesta forma podrem mostrar gràficament (només ens calen dues dimensions) cadascun dels vectors de característiques etiquetats (els diferents colors faran el paper d'etiquetes) amb la classe o *clúster* a la qual té una pertinença major. També es mostren les rectes de regressió corresponents a cada centroide (vegeu les Figures 3 i 4). Com és habitual, l'eix de les X és l'horitzontal i l'eix de les Y és el vertical.

Hem considerat per aquest exemple valors de c entre 2 i 10 tot i que només hem representat gràficament els casos de c que satisfan $2 \leq c \leq 7$. El conjunt

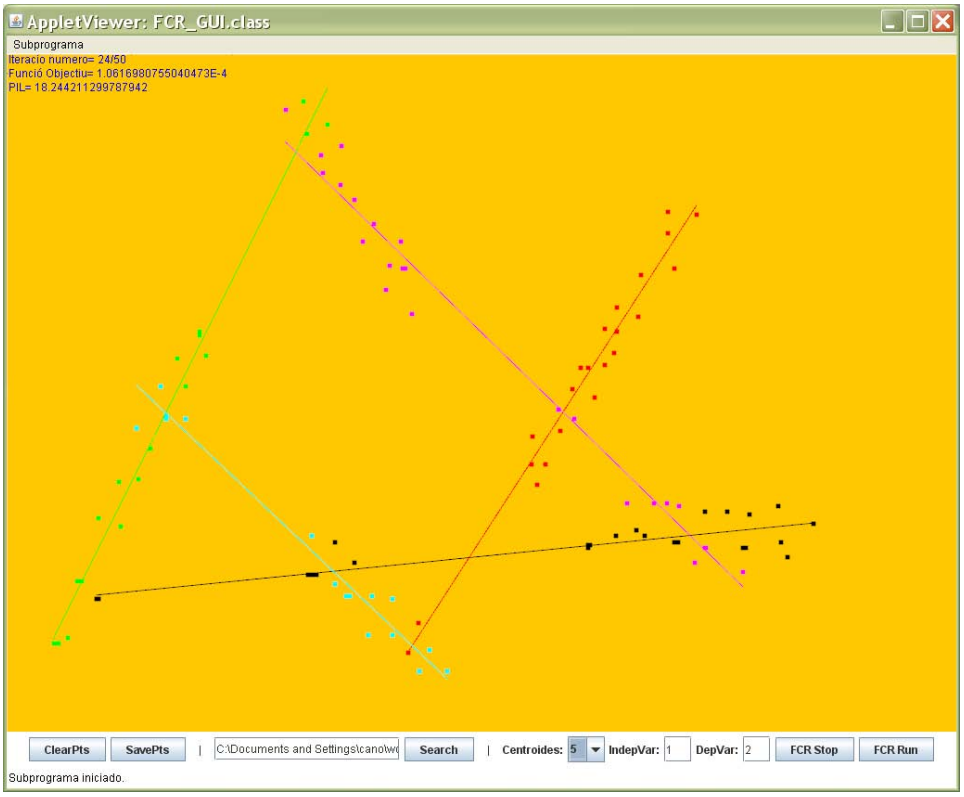

Figura 4: Estats finals dels models d'acord amb l'equació 1 amb $c = 5$ per a les dades de l'exemple.

de dades/objectes $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ està format per cinc conjunts independents de vectors de característiques. Per a cada valor de c s'ha calculat la pèrdua d'informació *PIL* i el risc de revelació *DR* obtinguts, així com el promig de totes dues mesures, *score* [35, 11] (calculat com $score = 0.5 * PIL + 0.5 * DR$). En la taula 5 i en la figura 5 es mostra la relació entre c (número de centroides) i PIL/DR: a mesura que c augmenta, la pèrdua d'informació disminueix i de forma contraria el risc de revelació augmenta. Per exemple, el menor risc de revelació / màxima pèrdua d'informació s'aconsegueix amb $c = 2$ on $PILmax = 40.41\%$ i $DRmin = 5.88\%$. En el cas de $c = 5$ obtenim $PIL = 20.11\%, DR = 9.27\%$. Finalment, la mínima pèrdua d'informació/màxim risc de revelació s'aconsegueix amb $c = 10$ obtenint valors de $PILmin = 12.79\%$ i $DRmax = 22.48\%$. Aquest exemple mostra com utilitzant FCRM podem utilitzar *c* per tal d'obtenir l'equilibri desitjat entre la utilitat de les dades sintètiques i la privadesa de les dades originals. Aquest és un clar avantatge entre el mètode proposat en aquest estudi, FCRM, i la família de mètodes IPSO, la qual genera dades sintètiques amb una pèrdua d'informació i un risc de revelació fixes.

C	F.O.	PIL	DR	SCORE
2	1.09E-04	40.41	5.88	23.15
3	1.62E-04	30.3	6.09	18.2
4	4.63E-03	23.97	8.03	16
5	1.46E-04	20.11	9.27	14.69
6	1.67E-04	18.89	9.98	14.43
7	1.07E-04	15.96	15.12	15.54
8	3.66E-03	15.24	16.17	15.7
9	1.12E-03	14.22	21.92	18.07
10	1.09E-01	12.79	22.48	17.64

Taula 5: Resultats obtinguts per al conjunt de dades d'exemple. C representa el número de classes o *clústers*, F.O. funció objectiu, PIL significa *Probabilistic Information Loss*, DR *Disclosure Risk* i *score* és el promig entre PIL i DR.

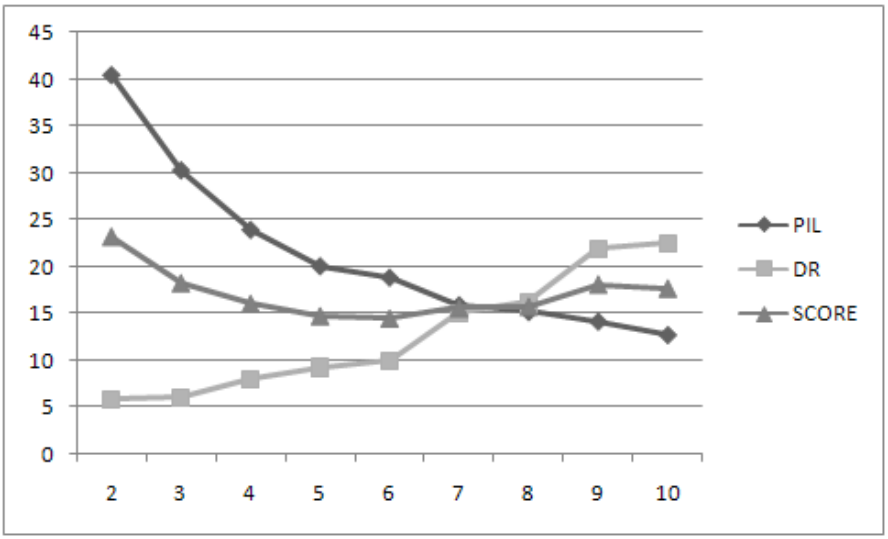

Figura 5: Relació entre PIL/DR i el nombre de classes o *clústers* C pel conjunt de dades d'exemple.

5 Experiments

Tal i com es menciona a la secció 2.3, en els experiments s'ha utilitzat el conjunt de dades $Census$. S'han tingut en compte dos conjunts de proves diferents. El primer conjunt de proves $S1$ està format per 9 variables dependents $v1, v3, v4, v6, v7, v9, v11, v12, v13$ i 4 variables independents, $v2, v5, v8, v10$. Per poder incloure la variable $v5$ s'ha necessitat la inversa generalitzada atès que aquesta variable causa singularitat de $[X^T D_i X]$ en l'equació 2. El segon conjunt de proves $S2$ està format per 4 variables dependents $v4, v7, v12, v13$ i 9 variables independents, $v1, v2, v3, v5, v6, v8, v9, v10, v11$.
Per a cada conjunt de proves hem generat les dades sintètiques segons diferents valors de c i a continuació s'ha calculat el valor de la funció objectiu (F.O.), *Probabilistic Information Loss* (PIL), *Disclosure Risk* (DR) i l'*standard score*. De la mateixa forma que en l'exemple (secció 4.3), per a valor petits de c la pèrdua d'informació és màxima mentre que el risc de revelació és mínim. De forma contraria, per a valors grans de c, la pèrdua d'informació és mínima mentre que el risc és màxim. Això ens aporta informació sobre la relació directa que existeix entre el nombre de centroides i la pèrdua d'informació així com sobre la relació inversa entre el nombre de centroides i el risc de revelació. Aquesta propietat es compleix en tots dos conjunts de proves. Per exemple, en $S1$ per a $c = 2$ obtenim $PIL_{max} = 44.677\%, DR_{min} = 9.583\%$ mentre que per a $c = 15$ obtenim $PIL_{min} = 7.164\%, DR_{max} = 26.97\%$. També en $S2$, per a $c = 2$ obtenim $PIL_{max} = 94.053\%, DR_{min} = 6.21\%$ mentre que per a $c = 81$ obtenim $PIL_{min} = 11.895\%, DR_{max} = 24.662\%$.

En aquest experiment amb les dades *Census* es vol estudiar en detall l'evolució de la pèrdua d'informació i el risc de revelació, per aquest motiu hem incrementat c fins a que els valors de PIL han esdevinguts menors que els de DR. D'aquesta forma podem mostrar la relació existent entre aquestes dues mesures. Una bona elecció normalment correspon al valor de c mínim, per exemple en $S1$ $SCORE_{min} = 16.912\%$ corresponent a $c = 13$. Cal dir que per a escenaris particulars (per exemple, amb dades molt sensibles) altres valors de c amb menys risc de revelació (menor DR) poden ser més adequats. Aquesta és la raó per la qual és especialment significatiu el disposar d'un paràmetre per a controlar el mètode de generació de dades sintètiques.

Per tal de comparar FCRM i IPSO, hem mesurat per a tots dos conjunts de proves la pèrdua d'informació, el risc de revelació així com també l'*score* en cas de fer servir IPSO-A, IPSO-B o IPSO-C per a generar les dades sintètiques. En el conjunt de proves $S1$ s'obté, en el cas de IPSO-A i IPSO-B, una pèrdua d'informació màxima de 49.16% juntament amb un risc de revelació de 10.04% mentre que amb FCRM s'obté amb $c = 2$ una pèrdua d'informació màxima de 44.67% juntament amb un risc de revelació de 9.58%. En canvi, s'hi utilitzem IPSO-C obtenim el millor *score* de 7.95% amb valors de $PIL = 9.52\%$ i $DR = 6.39\%$.
En relació a l'escenari $S2$, IPSO-A i IPSO-B obtenen un valor de pèrdua d'infor-

C	F.O.	PIL	DR	SCORE
2	0.184	44.677	9.583	27.13
3	0.005	32.614	12.294	22.454
4	0.078	26.668	14.791	20.73
5	0.195	21.797	16.9516	19.374
6	0.104	18.357	17.137	17.747
7	0.191	16.249	18.791	17.52
8	0.031	13.495	20.492	16.994
9	0.683	12.941	22.999	17.97
10	0.362	11.424	24.256	17.84
11	0.295	10.249	24.255	17.252
12	0.993	9.104	23.969	16.536
13	0.405	8.449	25.374	16.912
14	0.208	8.551	25.408	16.98
15	0.753	7.164	26.970	17.067
IPSO-A	-	49.163	10.044	29.603
IPSO-B	-	49.164	10.04	29.602
IPSO-C	-	9.522	6.392	7.957

Taula 6: Resultats obtinguts per al conjunt de proves S1: conjunt de dades *Census* amb 9 variables dependents.

mació de 44.44% i un risc de revelació del 14.493%, en canvi, utilitzant FCRM per a generar les dades sintètiques obtenim amb $c = 26$ un valor similar de risc de revelació però gairebé la meitat de pèrdua d'informació. Un altre cop, en el cas de l'IPSO-C, s'obté el millor *score* atès que FCRM per a valors similars de risc de revelació produeix gairebé el doble de pèrdua d'informació.

Finalment, fora dels dos conjunts de proves $S1$ i $S2$, s'ha estudiat el comportament del nombre de variables/atributs considerades independents, a partir d'ara anomenarem a aquest conjunt de proves $S3$. En aquest sentit s'ha procedit a, donat un valor de c fixat, anar agafant una a una totes les variables del fitxer Census (13 variables) i considerar-la com a única variable depenent. Per a cadascun dels 13 casos possibles de variables dependents és procedeix a considerar com a variables independents els següents n atributs. D'aquesta forma podem comprovar experimentalment quin és el comportament, respecte a la pèrdua d'informació i al risc de revelació, dels models de regressió a mesura que n augmenta. En altres paraules, l'experiment servirà per corroborar que a mesura que el model disposa de més variables independents aquest generarà dades sintètiques més semblants a les originals però que al mateix temps tindran més risc de revelació.

En l'experiment s'han considerat valors de c tals que $1 \leq c \leq 10$ i per a cada valor de c s'han considerat grups de n atributs independents, per a $1 \leq n \leq 19$.

C	F.O.	PIL	DR	SCORE
2	0,999945	94,053	6,210	50,132
4	0,999999966	81,792	6,428	44,110
5	0,999999715	61,834	7,782	34,808
6	0,999999995	72,550	7,424	39,987
8	0,999995425	59,113	10,152	34,633
9	0,9999968	38,990	11,042	25,016
10	9,917E-06	49,157	9,261	29,209
14	0,999895297	42,246	11,710	26,978
18	0,999992791	43,083	15,407	29,245
20	0,999993721	31,422	11,454	21,438
22	0,999995385	35,233	13,895	24,564
24	0,999910478	29,362	16,184	22,773
26	0,999999944	24,750	15,186	19,968
28	0,999999785	23,098	17,857	20,478
30	0,999713871	25,222	17,376	21,299
34	0,967993197	21,397	16,296	18,847
45	4,70034E-25	17,606	19,612	18,609
56	0,999999909	13,541	22,085	17,813
77	0,999832048	12,751	25,257	19,004
81	0,999997699	11,895	24,662	18,279
IPSO-A	-	44,44	14,493	29,467
IPSO-B	-	44,441	14,493	29,467
IPSO-C	-	17,037	11,022	14,029

Taula 7: Resultats obtinguts per al conjunt de proves S2: conjunt de dades *Census* amb 4 variables dependents.

C	N	PIL	DR	SCORE
2	1	63,96063083	4,273251671	34,11694125
2	2	60,63304442	5,141774576	32,8874095
2	3	56,2637375	5,278536899	30,7711372
2	4	54,39703789	5,860955186	30,12899654
2	5	52,56616912	6,726047306	29,64610821
2	6	49,92355693	6,670284562	28,29692074
2	7	46,4322841	7,271157573	26,85172084
2	8	46,18347629	7,421209861	26,80234308
2	9	45,13569801	7,931527846	26,53361293
2	10	39,94796621	9,594126784	24,7710465
2	11	36,94578884	10,74188761	23,84383822
2	12	35,48314376	10,60215459	23,04264918
2	13	34,02018227	11,11389769	22,56703998
2	14	33,10528376	11,54676863	22,32602619
2	15	31,90011836	11,54262144	21,7213699
2	16	30,28635472	11,65917584	20,97276528
2	17	29,72438823	12,4056938	21,06504101
2	18	29,21648118	12,24106227	20,72877173
2	19	28,97730173	12,66517107	20,8212364
...	...	...	...	...
4	1	49,03334158	8,844761165	28,93905137
4	2	45,23332296	9,482227075	27,35777502
4	3	39,30468361	10,3359029	24,82029325
4	4	36,24124824	11,57408418	23,90766621
4	5	32,95952474	12,42167251	22,69059862
4	6	31,70845549	13,38087219	22,54466384
4	7	30,31749243	14,28737466	22,30243354
4	8	28,96340469	14,4206434	21,69202405
4	9	28,23074474	13,79608811	21,01341642
4	10	23,76855495	16,00384986	19,88620241
4	11	22,38342589	18,20894721	20,29618655
4	12	20,78628072	18,59577233	19,69102652
4	13	20,28540616	18,89287801	19,58914209
4	14	19,88971323	18,93580575	19,41275949
4	15	19,52214067	18,97848615	19,25031341
4	16	18,45040876	19,44524073	18,94782474
4	17	18,29724024	19,80165785	19,04944904
4	18	18,15666619	20,42717176	19,29191897
4	19	17,4605634	20,3484739	18,90451865
...	...	...	...	...

Taula 8: Resultats obtinguts per a l'experiment $S3$ en el cas de $c = 2$ i $c = 4$. C representa el nombre de classes o *clústers*, N número de variables independents, PIL significa *Probabilistic Information Loss*, DR *Disclosure Risk* i *score* és el promig entre PIL i DR.

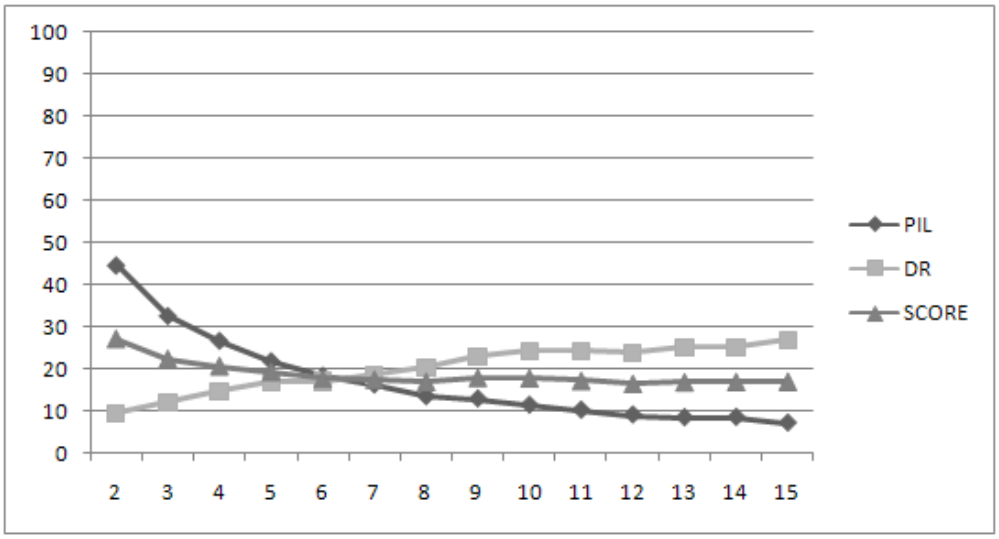

Figura 6: Relació entre PIL i DR respecte el nombre de clústers per al conjunt de proves S1.

En la taula 8 es mostren els resultats obtinguts per a valors de $c = 2$ i $c = 4$. Per raons d'espai i per tal de no proveir una taula de resultats excessivament gran i per tant impossible d'interpretar no es mostren la resta de resultats per a valors de c diferents de 2 i 4. En la figura 7 es proporciona una perspectiva 3D dels resultats obtinguts, en aquest cas si s'han considerat tots els resultats obtinguts. L'eix *Percent* representa el percentatge obtingut tant en risc de revelació (color verd), pèrdua d'informació (color vermell) com per al valor de *score* (color lila). En aquesta figura es pot observar tant el comportament de c com de n. Per a una c fixada obtenim que a mesura que augmentem n la pèrdua d'informació disminueix, atès que el model de regressió conté més paràmetres (variables independents) per ajustar-se al model real de les dades. Pel que fa al risc de revelació, aquest augmenta a mesura que augmenta n, també en el cas d'una c fixada.

6 Conclusions i possibles millores

En aquest estudi hem realitat un recorregut a través dels conceptes bàsics de *Statistical Disclosure Control*. Hem introduït els conceptes necessaris per a treballar amb microdades i algunes de les mesures més comuns de la pèrdua d'informació i del risc de revelació. De la mateixa forma s'han detallat els conceptes bàsics relacionats amb FCRM. Hem proposat aquest mètode per a la generació de dades sintètiques i s'ha avaluat en termes de privadesa i utilitat de les dades. Alguns d'aquest conceptes són els models de regressió *switching* i el mètode de classificació difusa *Fuzzy c-means*.

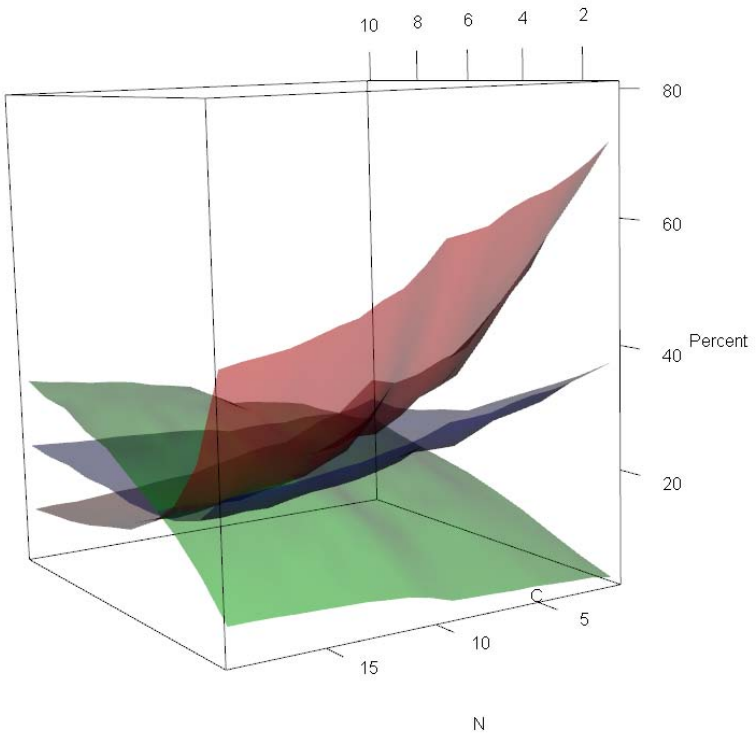

Figura 7: Pèrdua d'informació (vermell), risc de revelació (verd) i *score* (lila) per a diversos valors de centroides (C) i nombre de variables independents (N)

Pel que respecta a la viabilitat de FCRM per a la generació de dades sintètiques, hem analitzat la pèrdua d'informació i el risc de revelació segons diferents valors del nombre de centroides, principal paràmetre de FCRM. Aquesta proposta ha estat comparada amb una de les principals famílies de mètodes per a la generació de dades sintètiques, IPSO. Hem mostrat que els resultats de FCRM són, en general, millors que IPSO-A i IPSO-B però en canvi són, també en general, pitjors que IPSO-C. També hem destacat l'avantatge del paràmetre c (nombre de centroides) a l'utilitzar FCRM. Mentre que FCRM permet obtenir un equilibri determinat entre PIL/DR, a través dels diferents valors de c, IPSO sempre presenta el mateix nivell de privacitat i utilitat de les dades generades sintèticament.

Es consideren dues possibles millores del treball. En primer lloc es considera la possibilitat de que FCRM obtingui millors resultats que IPSO-C substituint el model lineal per un model basat en xarxa neuronal com el model basat en *Radial Basis Function (RBF) networks*. En segon lloc es considerarà si es possible incorporar en els models informació sobre restriccions entre les variables de la base de dades, seguint la línia de [2a].

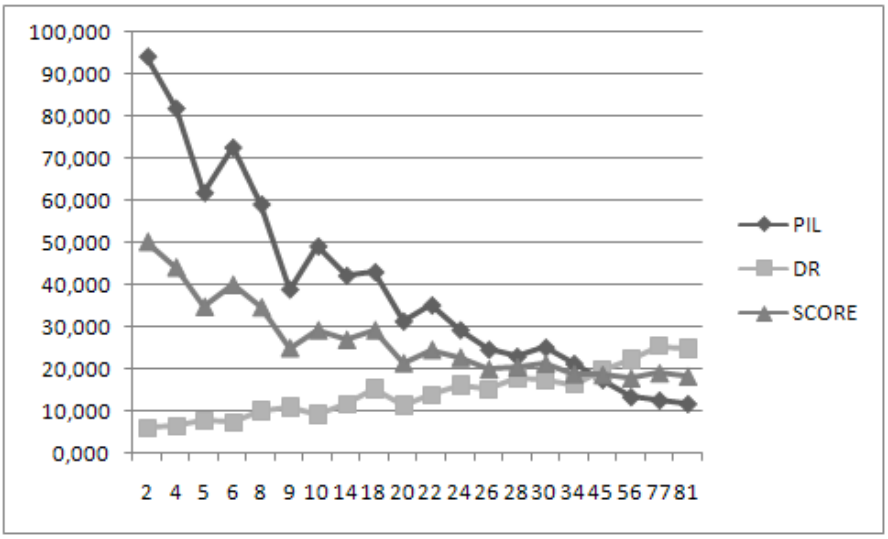

Figura 8: Relació entre PIL i DR respecte el nombre de clústers per al conjunt de proves S2.

A Apèndix

En aquest apèndix es descriu a molt alt nivell el codi generat per tal d'implementar la família de mètodes IPSO-A, IPSO-B i IPSO-C, així com també la implementació de FCRM. S'ha fet servir tant el llenguatge de programació *R* com *Java*. És important aclarir que l'objectiu principal de la implementació d'aquests mètodes és experimental, és a dir, poder obtenir resultats quantitatius del comportament dels mètodes segons els seus paràmetres.

Finalment s'adjunta el document format *paper* referent a aquest estudi. Aquest document formava part dels objectius secundaris d'aquest projecte final de carrera.

A.1 IPSO

Per implementar la família de mètodes IPSO, s'ha optat per utilitzar l'entorn de programació estadística *R* juntament amb la llibreria *sdcMicro* que conté mètodes per al control del risc de revelació estadística (*Statistical Disclosure Control*) per a la generació de fitxers de dades d'ús públic o científic.

R és un entorn estadístic sota la llicència *Open-Source (GPL)* que va sorgir després de *S* i *S-Plus* (http://www.insightful.com). Mentre que el llenguatge *S* es va desenvolupar al final dels anys 80 als laboratoris de AT&T, el projecte *R* va començar al 1995 sota la direcció de Robert Gentleman i Ross Ihaka, tots dos investigadors del departament d'estadística de la universitat de Auckland. Actualment *R* es manté gracies a la col·laboració de científics d'arreu del món. La pàgina web del projecte *R* és http://www.r-project.org. En aquesta pàgina es pot obtenir el software necessari per a utilitzar *R* així com les llibreries necessàries, com és el cas de la llibreria sdcMicro.

A continuació es proporciona el codi creat en aquest estudi per, a partir d'un fitxer de microdades (fitxer original), obtenir dades sintètiques (fitxer protegit) segons els mètodes de generació de dades sintètiques IPSO-A, IPSO-B o IPSO-C.

```
1  ###################################################
   # IMPLEMENTACIO IPSO-A, IPSO-B, IPSO-C
3  #      (basat en Muralidhar 2005)
   ###################################################
5  library(sdcMicro)

7  #Per a executar IPSO
   #1. Llegir el fitxer de microdades (separades per comes)
9  data <- read.table(file="rutaCompleta\nomFitxer", sep=",")
   #2. Establim atributs 1 i 2 com a dependents
11 Lxxx <- c(1, 2)
```

```
  #3. Establim atributs 3 i 4 com a independents
13 Lsss <- c(3, 4)

15 #IPSO-A
  IPSOA <- IPSOabc(data,Lxxx,Lsss,1)
17 IPSOB <- IPSOabc(data,Lxxx,Lsss,2)
  IPSOC <- IPSOabc(data,Lxxx,Lsss,3)
19


21
  IPSOabc <- function (data, Lxxx, Lsss, IPSO) {
23 # IPSO = 1: IPSO-a
  # IPSO = 2: IPSO-b
25 # IPSO = 3: IPSO-c
  # Defecte : IPSO-c
27
  nomTaula  <- "data"
29 dataIPSOa <- data
  dataIPSOb <- data
31 dataIPSOc <- data
  numRegistres <- nrow(data)
33


35
  # nomTaula: nom de les variables
37 # Lxxx, Lsss: llista de variables de xxx i sss
  varsx <- paste(nomTaula,"$","V", Lxxx, sep="")
39 varss <- paste(nomTaula,"$","V", Lsss, sep="")

41 sumaVx <- paste(varsx, collapse="+")
  sumaVs <- paste(varss, collapse="+")
43
  ################################################################
45 # IPSO-A: Model basat en la regressio
  ################################################################
47 for (i in Lxxx) {
     formula <- paste(nomTaula,"$V",i,"~", sumaVs, sep="")
49   rxresy <- predict(lm(as.formula(formula)))
     dataIPSOa[i] <- rxresy
51 }


53
  ################################################################
55 # IPSO-B: Afegim soroll
  #         quan la regressio fa servir xxx i sss: preserva B
57 #         quan la regressio nomes fa servir sss: NO preserva B
  ################################################################
59 for (i in Lxxx) {
         y <- rnorm(numRegistres)
```

```
61          y <- (y-mean(y))/sqrt(var(y))
            formula <- paste("y ~ ", sumaVs, " + ", sumaVx, sep="")
63          rxresy <- dataIPSOa[i] + residuals(lm(as.formula(formula)))
            dataIPSOb[i] <- rxresy
65  }

67
    ################################################################
69  # IPSO-C: Preserva B i Covariança
    ################################################################
71  for (i in Lsss) {
        if (i == Lsss[[1]]) { matSSS <- data[i] }
73          else { matSSS <- cbind(matSSS,data[i]) }
    }
75
    for (i in Lxxx) {
77          y <- rnorm(numRegistres)
            y <- (y-mean(y))/sqrt(var(y))
79
            formula <- paste("y ~ ", sumaVx, "+", sumaVs, sep="")
81
            yIPSOaResidu <- residuals(lm(as.formula(formula)))
83
            if (i == Lxxx[[1]]) {
85                  matYYYr <- yIPSOaResidu
                    matXXX <- data[i]
87                  }
            else {  matYYYr <- cbind(matYYYr,yIPSOaResidu)
89                  matXXX <- cbind(matXXX,data[i]) }
    }
91
    matYYYr <- as.matrix(matYYYr)
93  matXXX <- as.matrix(matXXX)
    matSSS <- as.matrix(matSSS)
95
    ## Sigma_{BB}^{-0.5} * b_i :
97  ##      aixo ha de tenir covarianca identitat
    ## ecs son els residus
99  part2 <- matYYYr %*% t(chol(solve(cov(matYYYr))))
    cov(part2)
101
    ## Sigma_{ee}^{0.5} * part2
103 eee <- cov(matXXX) - cov(matXXX,matSSS) %*% solve(cov(matSSS))
          %*% cov(matSSS,matXXX)
105
    part1 <- part2 %*% chol(eee)
107 yyyNou <- part1

109 ## Copiem els resultats a la matriu amb totes les dades
```

```
  for (i in 1:length(Lxxx)) {
111   for (j in 1:nrow(dataIPSOa)) {
        dataIPSOc[j,Lxxx[[i]]] <- dataIPSOa[j,Lxxx[[i]]] + yyyNou[j,i]
113 }}

115
  if (IPSO==1) {return(dataIPSOa)}
117 if (IPSO==2) {return(dataIPSOb)}
  if (IPSO==3) {return(dataIPSOc)}
119
  return(dataIPSOc)
121
  }
```

A.2 FCRM

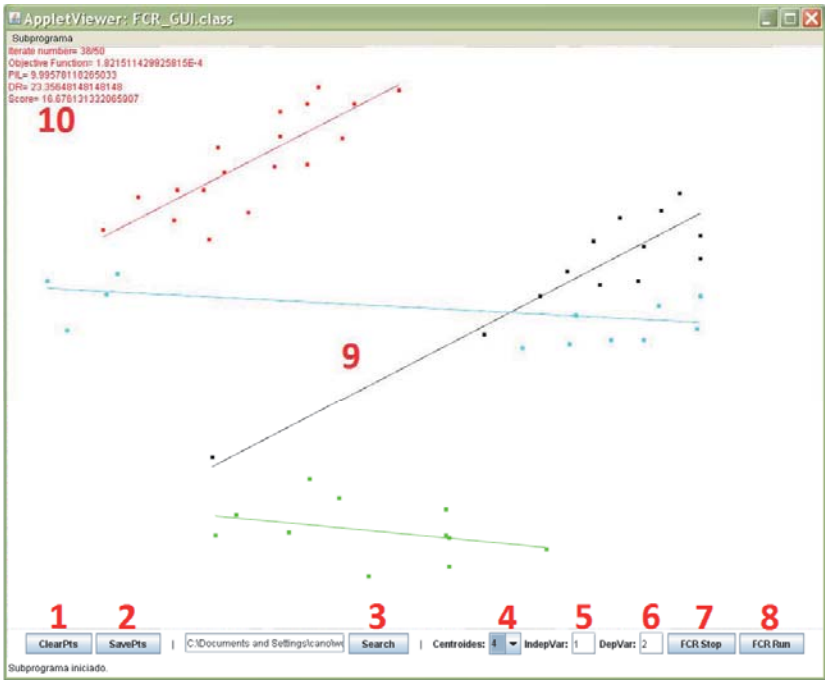

Figura 9: Captura de pantalla de la interfície gràfica d'usuari que implementa FCR_GUI.java

La implementació del mètode proposat en aquest estudi, utilitzar *Fuzzy c-Regression models (FCRM)* per a la generació de dades sintètiques, ha estat en *Java*. En la figura 10 es mostra el diagrama de classes simplificat (sense mètodes ni atributs) que implementa *FCRM*, no obstant, aquest diagrama de classes no inclou el codi necessari per a mesurar la pèrdua d'informació (*Probabilistic Information*

Loss) ni per a mesurar el risc de revelació (*Distance Based Record Linkage, Probabilistic Record Linkage i Interval Disclosure*), ja que es considera fora de l'àmbit de generació de dades sintètiques i aquestes mesures han estat implementades amb anterioritat a aquest estudi. Tot i això, al CD-ROM que s'entrega juntament amb aquesta memòria, està disponible el programari sencer, incloent també les mesures no considerades aquí. Cal remarcar que el software ha estat desenvolupat amb l'objectiu de poder realitzar els diferents experiments i obtenir dades rellevants per aquest estudi, en aquest sentit no s'ha seguit estrictament cap tipus de metodologia de desenvolupament de software encara que les classes creades són prou abstractes com per a poder oferir funcionalitat addicional si fos necessari. A continuació es mostra una petita descripció del propòsit de cada classe:

- **FCR_GUI**
  Aquesta classe implementa una petita interfície gràfica d'usuari (figura 9) útil per a representar gràficament els resultats de *FCRM* per al cas bidimensional (un atribut independent i un atribut dependent). En aquesta classe ens permet introduir, mitjançant simples *clics* de ratolí, dins de l'espai reservat per a la representació gràfica dels resultats (zona 9 en la figura), diferents punts que seran emmagatzemats dins de la classe *Database* un cop pitgem el botó *SavePts*(zona 2 en la figura). Si volem buidar la base de dades de punts introduïts manualment tant tan sols hem de pitjar el botó *ClearPts* (zona 1 en la figura). Una altre opció, en el cas de que disposem d'un fitxer de text en clar amb els punts a tractar (un punt (x_i, y_i) per fila i separant cadascuna de les coordenades per espai o comes), és especificar el fitxer mitjançant el botó *Search* (zona 3 en la figura). El menú desplegable *Centroides* (zona 4 en la figura) permet especificar el valor de c amb un valor màxim de 10. Si volem que la primera coordenada de cada punt es consideri com a independent caldrà especificar 1 en el quadre de text (zona 5 en la figura). De la mateixa forma, si volem que la segona coordenada de cada punt es consideri com a dependent caldrà especificar 2 en el quadre de text (zona 6 en la figura). Finalment els botons *FCR Stop* i *FCR Run* serveixen per poder realitzar una parada dins de l'execució de l'algorisme.
  Durant cada iteració del mètode, a la part superior esquerra (zona 10 en la figura) de la interfície gràfica d'usuari, es mostra informació útil com el número d'iteració, el valor de la funció objectiu, la pèrdua d'informació (PIL), el risc de revelació (DR) i l'*score*. Els resultats es mostren de forma gràfica (zona 9 en la figura) de la següent forma: els punts originals es pinten com quadrats, cada quadrat tindrà el color del clúster al que té més pertinença i les rectes de regressió també es pinten amb el color del clúster al qual pertanyen.

- **Database**
  Tal i com s'ha explicat anteriorment, aquesta classe s'encarrega, entre altres coses, d'emmagatzemar els punts introduïts a mà o bé llegits de fitxer.

Només s'utilitza aquesta classe quan s'està executant la interfície gràfica d'usuari atès que només es permet la introducció de punts bidimensionals.

- **FCR**
  Aquesta classe conforma el nucli de la implementació d'aquest mètode i per tant s'encarrega principalment de llegir les dades des de fitxer i emmagatzemar-les internament, classes *File* i *Data*. A continuació realitza tots els passos necessaris per a la generació de les dades sintètiques, secció 4.2. Per tant, utilitza tant les classes *Step1*, *Step2* i *Step3* com la classe *YRegressionModel*. Les tres primeres implementen els tres passos bàsics de que consta la *Fuzzy c-Regression* i la darrera classe emmagatzema els models de regressió que es faran servir per a generar les dades sintètiques.

- **Arguments**
  Aquesta classe simplement emmagatzema els arguments que rep *FCR*, en forma de cadena de caràcters, en diferents variables per a un accés més directe.

- **File**
  Aquesta classe s'encarrega de llegir un fitxer de microdades (fitxer en clar). Conté els mètodes necessaris per carregar les dades llegides a la classe *Data*.

- **Data**
  La classe *Data* emmagatzema les microdades llegides i proporciona mètodes per accedir i editar aquestes dades.

- **Step1**
  Com el nom d'aquesta classe indica, la funcionalitat que implementa està relacionada amb el primer pas de l'algorisme.

- **Step2**
  Anàlogament a l'anterior classe, *Step2* s'encarrega de realitzar els càlculs associats al segon pas de l'algorisme.

- **Step3**
  De la mateixa forma que les dues classes anteriors a aquesta, *Step3* realitza els càlculs associats al tercer pas de l'algorisme.

- **YRegressionModel** Aquesta classe és l'encarregada d'emmagatzemar i proporcionar accés a tots els paràmetres relacionats amb els c diferents models de regressió.

- **FcMeans**
  Aquesta classe s'utilitza per a inicialitzar els clústers abans de començar a iterar els passos que conformen *Fuzzy c-Regression*. Aquesta classe també servirà com a contenidor de les matrius de pertinences i dels c centroides.

- **Centroide**
  La classe *Centroide* conté tota la informació relacionada amb cada clúster, més concretament, sobre el centre de cada clúster. També conté els coeficients de les rectes de regressió relacionades amb cada clúster.

- **ObjecteFcMeans**
  Aquesta classe emmagatzema els atributs que formen cada objecte, és a dir, els valors dels atributs de cada línia de microdades (objecte). També conté les pertinences anterior i actuals de l'objecte a cadascun dels clústers. Això és útil per a calcular el valor de la funció objectiu i verificar si hem arribat al llindar de convergència que aturaria les iteracions de *Fuzzy c-Regression*.

- **Pertinenca**
  Simplement emmagatzema el valor d'una pertinença en concret i a quin centroide fa referència aquesta pertinença.

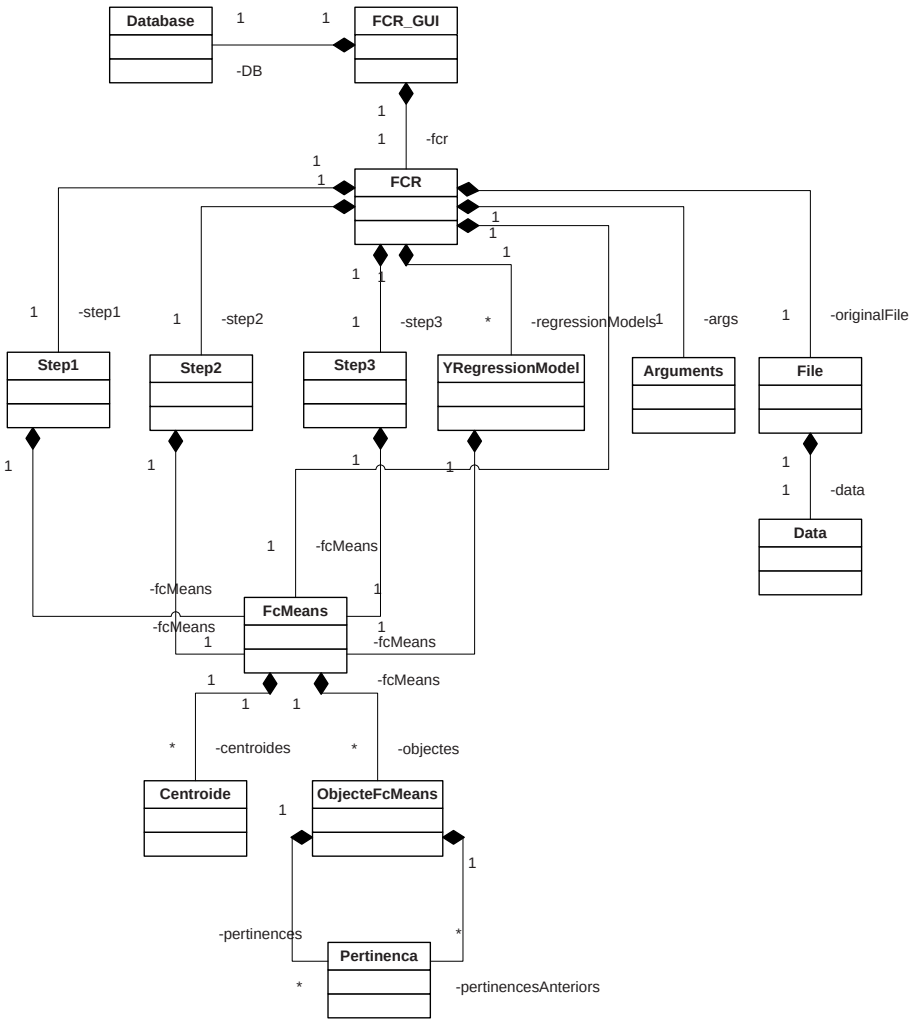

Figura 10: Diagrama de classes simplificat del software que implementa FCRM

A.3 Publicacions en format *paper* relacionades amb aquest estudi

Per a complementar el treball realitzat en aquest projecte final de carrera he redactat el *paper* que s'inclou a continuació. El títol del mateix és "Generation of Synthetic Data by Means of Fuzzy c-Regression" [1a]. Aquest *paper* ha estat acceptat per a la seva presentació en el FUZZ-IEEE2009, 2009 IEEE International Conference on Fuzzy Systems, ICC Jeju, Jeju Island, Korea, del 20 al 24 d'Agost del 2009 (http://www.fuzz-ieee2009.org/). Està previst que jo faci la presentació d'aquest treball a Corea.

Contribucions

[1a] Cano, I., Torra, V., (2009) Generation of Synthetic Data by Means of Fuzzy c-Regression. 2009 IEEE International Conference on Fuzzy Systems, FUZZ-IEEE2009. En premsa.

[2a] Cano, I., Navarro-Arribas, G., Torra, V., (2009) A new framework to automate constrained microaggregation. The 4th International Workshop on Security, IWSEC2009. Acceptat com a *short paper*.

Bibliografia

[1] Bezdek, J., (1981) Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York.

[2] Brand, R., Domingo-Ferrer, J., Mateo-Sanz, J., (2002) Reference data sets to test and compare SDC methods for protection of numerical microdata. European Project IST-2000-25069 CASC, http://neon.vb.cbs.nl/casc.

[3] Burridge, J., (2003) Inforation preserving statistical obfuscation. Statistics and Computing, vol. 13, pp. 321-327.

[4] Dalenius, T., (1988) Controlling Invasion of Privacy in Surveys. Statistics Sweden, Stockholm.

[5] Dandekar, R., Domingo-Ferrer, J., Sebé, F., (2002) LHS-based hybrid microdata vs rank swapping and microaggregation for numeric microdata protection. In J. Domingo-Ferrer, editor, Inference Control in Statistical Databases, volume 2316:153-162 of Lecture Notes in Computer Science, Berlin Heidelberg, Springer.

[6] Domingo-Ferrer, J., Mateo-Sanz, J.M., Torra, V., (2001) Comparing SDC methods methods for microdata on the basis of information loss and disclosure risk. In Pre-proceedings of ETK-NTTS'2001, Vol. 2:807-826, Luxemburg. Eurostat.

[7] Domingo-Ferrer, J., Torra, V., Mateo-Sanz, J.M., Sebé, F., (2006) Empirical Disclosure risk assessment of the ipso synthetic data generators. In Monographs in Official Statistics-Work Session On Statistical Data Confidenciality, pages 227-238, Luxemburg. Eurostat.

[8] Domingo-Ferrer, J., Sebé, F., Solanas, A., (2008) A polynomial-time approximation to optimal multivariate microaggregation. Computers and Mathematics with Applications, Vol. 55(4):714-732.

[9] Domingo-Ferrer, J., Torra, V., (2005) Ordinal, continuous and heterogeneous k-anonymity through microaggregation. Data Mining and Knowledge Discovery, Vol. 11(2):195-212.

[10] Domingo-Ferrer, J., Mateo-Sanz, J.M., (2002) Practical data-oriented microaggregation for statistical disclosure control. IEEE Transactions on Knowledge and Data Engineering, 14(1):189-201.

[11] Domingo-Ferrer, J., Torra, V., (2001) A quantitative comparison of disclosure control methods for microdata, Confidentiality, disclosure and data acces: Theory and practical applications for statistical agencies. Doyle, P., Lane, J.I., Theeuwes, J.J.M., Zayatz, L.V., eds., Elsevier, pp.111-133.

[12] Fellegi, I.P., Sunter, B., (1969) A theory for record linkage. Journal of the American Statistical Association, 64(328):1183-1210.

[13] Fienberg, S.E., (2000) Confidentiality and data protection through disclosure limitation: Evolving principles and technical advances. The Philippine Statistician, vol. 49, pp. 461-468.

[14] Fuller, W.A., (1993) Masking procedures for microdata disclosure limitation. Journal of Official Statistics, vol. 9, pp. 338-406.

[15] Mateo-Sanz J.M., Domingo-Ferrer J.D., Sebé F. (2005) Probabilistic Information Loss Measures in Confidentiality Protection of Continuous Microdata. Data Mining and Knowledge Discovery. Vol. 11, Issue 2, pp. 181-193.

[16] Moore, R.A., (1996) Controlled data swapping techniques for masking public use microdata sets. U.S. Bureau of the Census.

[17] Muralidhar, K., Sarathy, R., (2003) A theoretical basis for perturbation methods. Statistics and Computing 13:329-335.

[18] Hathaway R.J., Bezdek J.C. (1993) Switching regression models and fuzzy clustering. IEEE Transactions on Fuzzy Systems. Vol. 1(3):195-204.

[19] Laszlo, M., Mukherjee, S., (2005) Minimum spanning tree partitioning algorithm for microaggregation. IEEE Transactions on Knowledge and Data Engineering. Vol. 17(7):902-911.

[20] Nin, J., (2008) Contributions to record linkage for disclosure risk assessment. Tesi doctoral - Universitat Autònoma de Barcelona. Escola Tècnica Superior d'Enginyeria, Departament de Ciències de la Computació. ISBN: 978846916595.

[21] Nin, J., Herranz, J., Torra. V., (2008) Towards a More Realistic Disclosure Risk Assessment, Provacy in Statistical Databases (PSD), vol. 5262 of Lecture Notes in Computer Science, pp. 152-165. Springer.

[22] Pagliuca, D., Seri, G., (1999) Some results of individual ranking method on the system of enterprise accounts annual survey. Technical report, Esprit SDC Project, Deliverable MI-3/D2.

[23] Rubin, D.B., (1993) Discussion on 'Statistical Disclosure Limitation'. Journal of Official Statistics, vol. 9, pp. 461-468.

[24] Ruspini, E., (1969) A new approach to clustering, Information and Control, Vol. 15:22-32.

[25] Schott, J.R., (2005) Matrix Analysis for Statistics. Wiley-Interscience. ISBN: 0471669830.

[26] Torra, V., Abowd, J.M., Domingo-Ferrer, J., (2006) Using mahalanobis distance based record linkage for disclosure assessment. In Privacy in Statistical Databases, volume 4302 of Lecture Notes in Computer Science, pages 175-186. Springer.

[27] Torres, A., (2003) Contribucions a la microagregació per a la protecció de dades estadístiques.Tesis doctoral - Universitat Politècnica de Catalunya. ISBN: 8468838829.

[28] U.S. Census Bureau, Data Extraction System, http://www.census.gov.

[29] De Veaux, R. D., (1989) "Mixtures of linear regressions", Computational Statistics and Data Analysis, vol. 8, pp. 227-245.

[30] Wang, D.W., Liau, C.J., Hsu, T.S., (2007) An epistemic framework for privacy protection in database linking. Data and knowledge engineering, 61:176-205.

[31] Willenborg, L., De Waal, T., (1996) Statistical Disclosure Control in Practice, Springer LNS 111.

[32] Winkler, W.E., (1999) Re-identification methods for evaluating the confidentiality of analytically valid microdata. In Statistical Data Protection, J. Domingo-Ferrer (Ed.), Luxemburg: Office for Official Publications of the European Communities. (Journal version in Research in Official Statistics, vol. 1, no. 2, pp. 50-69, 1998).

[33] Winkler, W.E., (1995) Matching and record linkage. Business Survey Methods, pages 355-284, 1995.

[34] Yancey, W.E., Winkler, W.E., Creecy, R.H., (2002) Disclosure risk assessment in perturbative microdata protection. In J. Domingo-Ferrer, editor, Inference Control in Statistical Databases, Vol. 2316:195-152 of Lecture Notes in Computer Science, Berlin Heidelberg, Springer.

[35] http://ppdm.iiia.csic.es/