

Estudi de mètodes de generació de dades sintètiques

Isaac Cano Franco

Escola Tècnica Superior d'Enginyeria
Universitat Autònoma de Barcelona
ETSE-UAB

01/07/09
Bellaterra

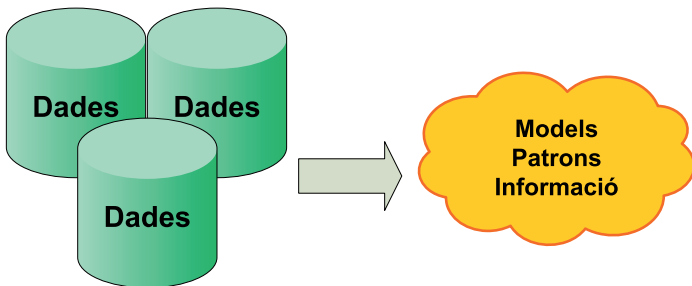
- 1 Introducció
- 2 Conceptes Previs
- 3 Mètodes de generació de dades sintètiques
 - Fuzzy c-Regression Models
- 4 Conclusions i millores

Sumari

- 1 Introducció
- 2 Conceptes Previs
- 3 Mètodes de generació de dades sintètiques
- 4 Conclusions i millores

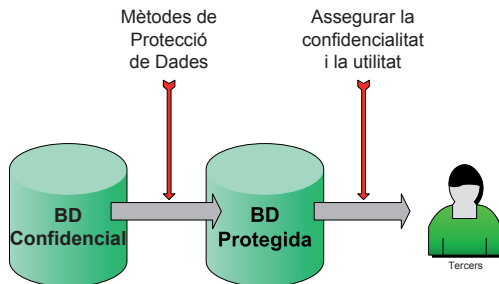
Privacy Preserving Data Mining (PPDM)

- **Data Mining**



Privacy Preserving Data Mining (PPDM)

- Privacy Preserving Data Mining



Sumari

- 1 Introducció
- 2 Conceptes Previs
- 3 Mètodes de generació de dades sintètiques
- 4 Conclusions i millores

Taula d'una BD (Microdades)

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: Exemple de microdades amb variables qualitatives i quantitatives.

Taula d'una BD (Microdades)

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: Exemple de microdades amb variables qualitatives i quantitatives.

Procés de protecció de microdades

DNI	Nom	Any de Naixement	C.P.	Malaltia
34556289T	David González	1978	08304	Grip A
38865344N	Pere Aguilar	1965	08301	SIDA
45878601A	Nuria Castella	1952	08802	Alzheimer
...	...	...	...	...

Tipus d'atributs:

- **Identificadors Directes:** DNI, número de la Seguretat Social, ...
- **Identificadors Indirectes:** edat, sexe, estat civil, ...
- **Atributs Confidencials:** malalties, passat criminal, ...

Procés de protecció de microdades

Identificadors DNI	directes Nom	Any de Naixement	C.P.	Malaltia
34556289T	David González	1978	08304	Grip A
38865344N	Pere Aguilar	1965	08301	SIDA
45878601A	Nuria Castella	1952	08802	Alzheimer
...	...	...	...	...

Tipus d'atributs:

- **Identificadors Directes:** DNI, número de la Seguretat Social, ...
- **Identificadors Indirectes:** edat, sexe, estat civil, ...
- **Atributs Confidencials:** malalties, passat criminal, ...

Procés de protecció de microdades

DNI	Nom	Identificadors	indirectes	Malaltia
		Any de Naixement	C.P.	
34556289T	David González	1978	08304	Grip A
38865344N	Pere Aguilar	1965	08301	SIDA
45878601A	Nuria Castella	1952	08802	Alzheimer
...	...	...	...	...

Tipus d'atributs:

- **Identificadors Directes:** DNI, número de la Seguretat Social, ...
- **Identificadors Indirectes:** edat, sexe, estat civil, ...
- **Atributs Confidencials:** malalties, passat criminal, ...

Procés de protecció de microdades

DNI	Nom	Any de Naixement	C.P.	Atribut Confidencial
				Malaltia
34556289T	David González	1978	08304	Grip A
38865344N	Pere Aguilar	1965	08301	SIDA
45878601A	Nuria Castella	1952	08802	Alzheimer
...	...	...	...	...

Tipus d'atributs:

- **Identificadors Directes:** DNI, número de la Seguretat Social, ...
- **Identificadors Indirectes:** edat, sexe, estat civil, ...
- **Atributs Confidencials:** malalties, passat criminal, ...

Procés de protecció de microdades

DNI	Nom	Any de Naixement	C.P.	Malaltia
34556289T	David González	1978	08304	Grip A
38865344N	Pere Aguilar	1965	08301	SIDA
45878601A	Nuria Castella	1952	08802	Alzheimer
...	...	...	...	...

Microdades Originals



Anonimització:

Mètodes Pertorbatius, Mètodes no Pertorbatius o Generadors de Dades Sintètiques



Any de Naixement	C.P.	Malaltia
1976	083XX	Grip A
1968	083XX	SIDA
1949	088XX	Alzheimer
...	...	...

Microdades Protegides

Risc de revelació

Microdades **Protegides**

Any de Naixement	C.P.	Malaltia
1976	083XX	Grip A
1968	083XX	SIDA
1949	088XX	Alzheimer
...	...	...

Microdades **Enllaçades**

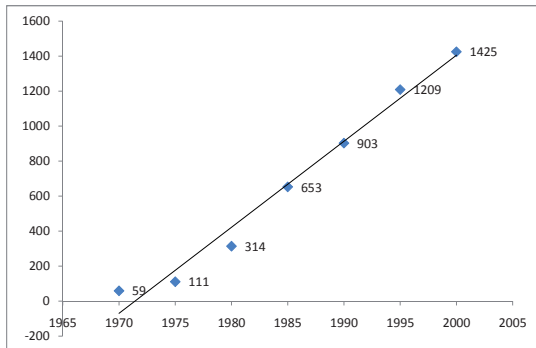
Nom	Malaltia
Nuria Castella	Alzheimer

Nom	Any de Naixement	C.P.
...	...	...
Nuria Castella	1952	08802
...	...	...

Microdades **Originals** d'una altre font

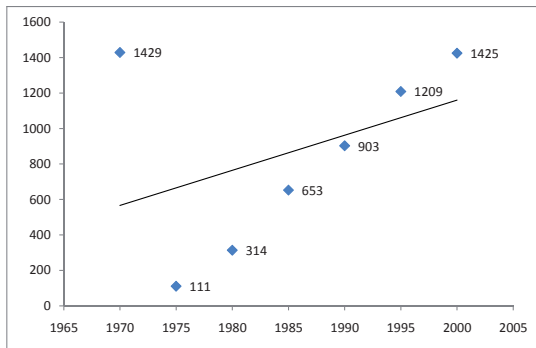
Pèrdua d'informació

Any	Ingressos per exportacions
1970	59
1975	111
1980	314
1985	653
1990	903
1995	1209
2000	1425



Pèrdua d'informació

Any	Ingressos per exportacions
1970	1429
1975	111
1980	314
1985	653
1990	903
1995	1209
2000	1425



SCORE

- **En general:**

↑ Pèrdua d'informació $\Rightarrow$ ↓ Risc de revelació

↓ Pèrdua d'informació $\Rightarrow$ ↑ Risc de revelació

- **SCORE:**

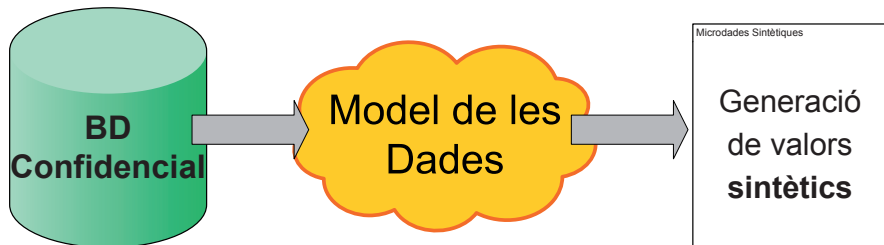
Agregar la pèrdua d'informació (IL) i el risc de revelació (DR)

$$SCORE = \frac{IL+DR}{2}$$

Sumari

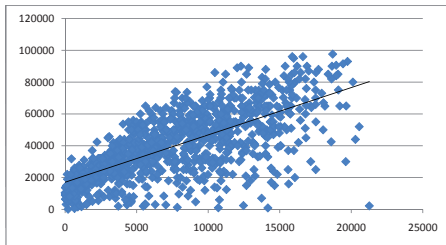
- 1 Introducció
- 2 Conceptes Previs
- 3 Mètodes de generació de dades sintètiques**
- 4 Conclusions i millores

Esquema de funcionament

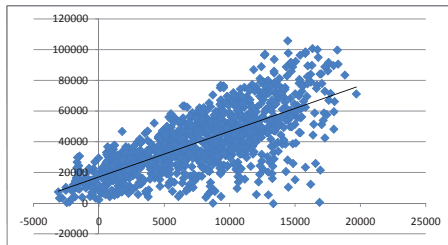


Information Preserving Statistical Obfuscation (IPSO)

Fitxer original



Fitxer protegit amb IPSO-C

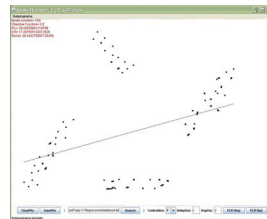
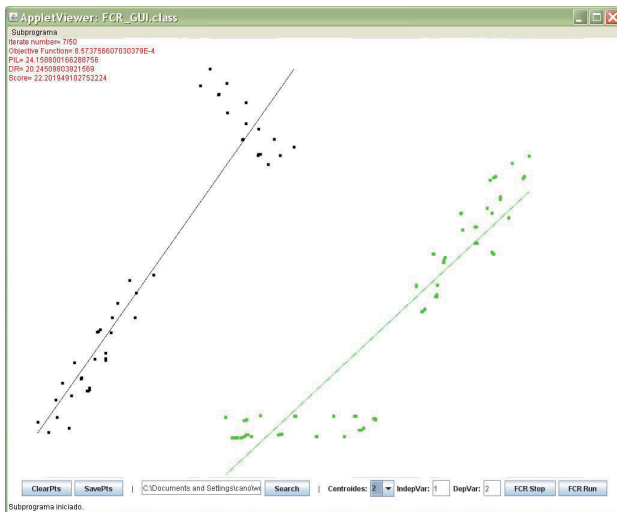


- **Pèrdua d'informació:** 17.04%
- **Risc de revelació:** 11.022%
- **SCORE:** 14.03%

Sumari

- 1 Introducció
- 2 Conceptes Previs
- 3 Mètodes de generació de dades sintètiques**
 - Fuzzy c-Regression Models
- 4 Conclusions i millores

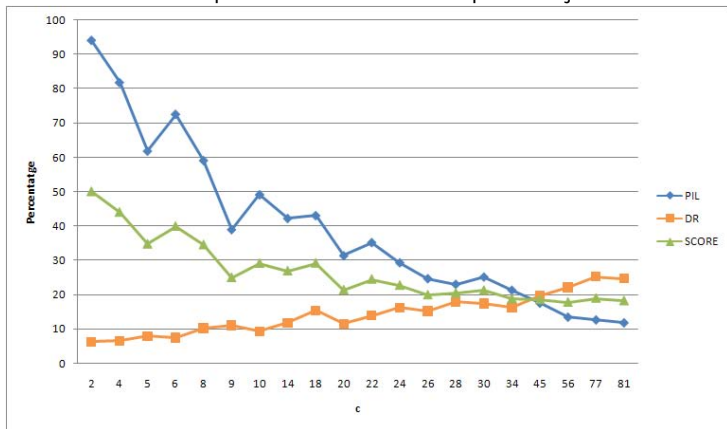
Fuzzy c-Regression Models (FCRM)



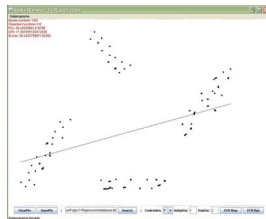
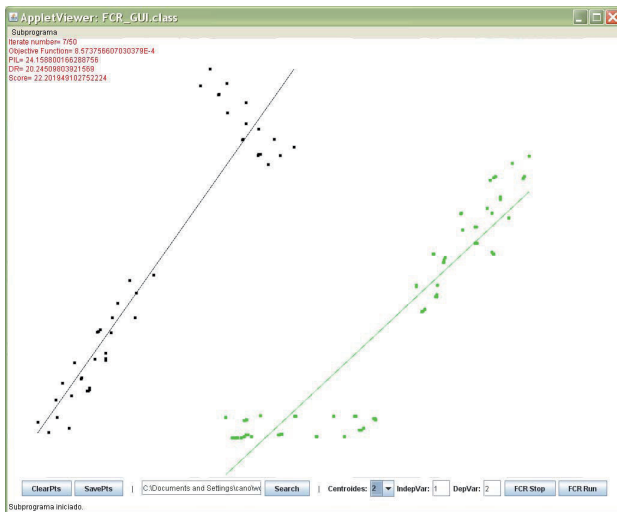
- $c = 2$
- Pèrdua d'informació:
24.16%
- Risc de revelació:
20.24%
- SCORE: 22.20%

Fuzzy c-Regression Models (FCRM)

Relació entre PIL i DR respecte el nombre de classes per al conjunt de dades *Census*.

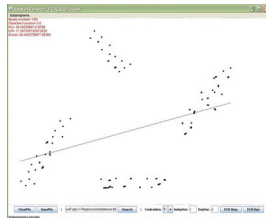
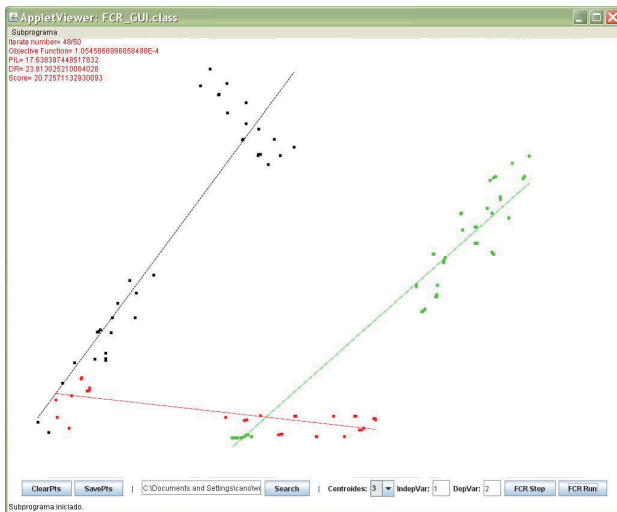


Fuzzy c-Regression Models (FCRM)



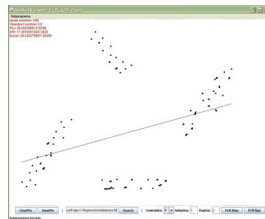
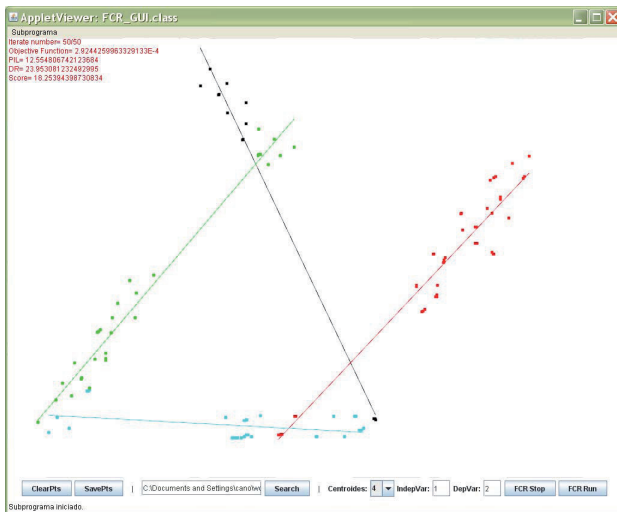
- $c = 2$
- Pèrdua d'informació:
24.16%
- Risc de revelació:
20.24%
- SCORE: 22.20%

Fuzzy c-Regression Models (FCRM)



- $c = 3$
- Pèrdua d'informació:
17.64%
- Risc de revelació:
23.81%
- SCORE: 20.73%

Fuzzy c-Regression Models (FCRM)



- $c = 4$
- **Pèrdua d'informació:**
12.55%
- **Risc de revelació:**
23.95%
- **SCORE:** 18.25%

Sumari

- 1 Introducció
- 2 Conceptes Previs
- 3 Mètodes de generació de dades sintètiques
- 4 Conclusions i millores

Conclusions

- FCRM obté en general millors resultats que IPSO-A i IPSO-B.
- FCRM permet que l'usuari pugui obtenir un equilibri entre Pèrdua d'informació i Risc de revelació determinat.

Conclusions

- S'han estudiat els conceptes bàsics de *Privacy Preserving Data Mining*:
 - Mesures de pèrdua d'informació.
 - Mesures de risc de revelació.
 - Fuzzy c-means.
 - Fuzzy c-regression.
- Cano, I., Torra, V., Generation of Synthetic Data by means of Fuzzy c-Regression. *2009 IEEE International Conference on Fuzzy Systems*, Jeju Island, Corea, 2009.

Millores

- Millorar FCRM per tal d'obtenir millor SCORE.
 - Per exemple utilitzant xarxes RBF.
- Incorporar informació sobre restriccions entre atributs.

Gràcies!