Protecció de Dades Estadístiques

Daniel Baena Mirabete
http://www-eio.upc.es/~dbaena

Introducció

Vivim a l'era de la informació. Les noves tecnologies, com ara els dispositius mòbils o portàtils, o els darrers desenvolupaments software, com Facebook, Twitter o LinkedIn, han fet possible que la comunicació avui dia sigui gairebé immediata. Són moltes les persones i/o empreses que han trobat en la xarxa el seu principal canal de comunicació. La majoria veuen en ella el millor i més eficient recurs per arribar al major nombre de gent, en el menor temps possible. D'altra banda ens trobem diàriament amb milions i milions de dades movent-se a velocitats salvatges, amb empreses molt interessades en capturar-les i explotar-les, aplicant mètodes estadístics avançats i amb l'objectiu d'aconseguir el coneixement necessari per guanyar posicions en el difícil i cada cop més competitiu mercat . I és que, la informació és un recurs molt valuós i decisiu. Ara bé, que passa amb el dret que qualsevol persona té a la intimitat?. Segons algunes persones, el fi justifica els mitjans. Però, és justificable perdre "involuntàriament" l'anonimat envers una millor productivitat o servei al ciutadà? La declaració universal dels Drets Humans i la Llei Orgànica de Protecció de Dades (LOPD) consideren sense cap mena de dubte que no i així ho manifesten en els seus documents legals. Ens trobem doncs amb un problema de difícil solució. D'una banda, la necessitat o interès social de disposar de la major quantitat possible d'informació. En aquest sentit, els continus avenços tecnològics ho fan viable. Ara bé, per l'altre costat tenim el clar risc de revelació de dades de caràcter personal que això suposa. I per llei, aquests riscos s'han de limitar.

Són molts els sectors que es troben amb aquest problema. Diversos sectors com la banca, la indústria farmacèutica, el comerç electrònic o els motors de cerca a la web, entre molts altres, exploten diàriament les dades personals de clients per a benefici propi. Per exemple, els motors de cerca a la web, guarden primer i utilitzen després tota la informació introduïda pels clients en les seves cerques per fer-ne publicitat personalitzada (Google va tenir uns ingressos propers als 21128.5 milions de dòlars al 2008 en anuncis!). Amb aquestes dades ningú no dubta de la necessitat i/o interès de Google en disposar d'una gran quantitat de dades personals. Però, i el risc de revelació? A l'any 2006, AOL Research, amb tota la bona fe i amb l'objectiu de proporcionar dades als investigadors, va publicar una gran quantitat de consultes realitzades pels clients durant un periode de 3 mesos. Les dades, tot i ser prèviament anonimitzades, no van evitar que el diari New York Times identifiqués a un client. Aquest fet va obrir un important debat en defensa de la privacitat dels usuaris a Internet. I evidentment va danyar molt la imatge de la companyia AOL Research (a més d'haver de fer front a vàries denúncies i sancions econòmiques).

Veiem doncs que tenim davant un problema real i amb un important impacte a la societat. Tot i els diversos camps d'aplicació existents, en aquest article ens referirem únicament al sector de l'Estadística Oficial, que tenen en la revelació de dades estadístiques una de les seves majors preocupacions. L'objectiu principal dels Instituts d'Estadística (INE's) és publicar la major quantitat de dades fiables i amb el màxim de detall possible. Objectiu que entra clarament en conflicte amb l'obligació legal de preservar l'anonimat de les

dades obtingudes com a resultat de qualsevol activitat estadística portada a terme per les administracions públiques (secret estadístic). En aquest sector, existeixen fins i tot, lleis creades especialment per aquest àmbit en particular. Des de 1989, la llei de la funció Estadística Pública (BOE 11-5-1989, llei 12/1989) i en Catalunya, des de 1998, la llei 23/1998 (article 26) que defensen el correcte compliment del secret estadístic. Però més enllà de les raons legals, hi ha una raó molt més important. La confiança dels enquestats. Ells són el principal motor de l'Estadística Oficial. Si ells, els enquestats, senten que les seves dades personals no són segures, penseu que algú contestarà, per exemple, l'enquesta d'hàbits sexuals? Segurament s'ho pensaran dues vegades...

Són molts els esforços dedicats en els darrers anys a la recerca de mètodes de protecció de dades estadístiques, que permetin publicar la major quantitat de dades possible tot minimitzant al màxim el risc de revelació individual. A Catalunya, l'Idescat (Institut d'Estadística de Catalunya) ha prestat una atenció especial a aquest tema [3, 4, 18, 19].

Mètodes de protecció de dades

Els INE's treballen bàsicament amb dos tipus de fitxers de dades. Fitxers de microdades (censals o mostrals) compost per N registres (individus/empreses) amb M variables cadascú (numèriques, p.e: edat, salari,... o categòriques, p.e: estat civil, sexe, professió,...) i els fitxers de macrodades (dades tabulars), format pel creuament d'una o més variables categòriques d'un fitxer de microdades.

Protecció de Microdades

Parlarem primer del món de les microdades. I començarem amb un petit exemple. Imaginem que els INE's disposen d'un fitxer original (Taula 1) molt sensible, amb informació personal de salut. Suposem ara que davant l'elevada demanda d'informació, decideixen publicar-ho. Abans però, li fan un tractament previ. Eliminen algunes variables d'identificació directe (Nom,Adreça,DNI) i eliminen i/o agreguen alguns valors d'algunes variables. El fitxer resultant d'aquest tractament és el de la taula 2.

Nom	Adreça	DNI	Sexe	Edat	Ciutat	E.Civil	Malaltia
Antoni	c/...	001	H	33	Banyoles	Casat	Cor
Alba	c/...	002	D	40	Manresa	Divorciat	Asma
Imma	c/...	003	D	36	Banyoles	Casat	Asma
Pep	c/...	004	H	36	Banyoles	Solter	Apendicitis
Albert	c/...	005	H	22	Badalona	Solter	SIDA
Eduard	c/...	006	H	81	Banyoles	Vidu	Fractura

Taula 1: Exemple fitxer original de microdades

Sexe	Edat	Ciutat	E.Civil	Malaltia
H	33	Banyoles	Casat	Cor
D	40	–	Vidu-o-Divorciat	Asma
D	36	–	Casat	Asma
H	36	Banyoles	Solter	Apendicitis
H	22	Badalona	Solter	SIDA
H	81	Banyoles	Vidu-o-Divorciat	Fractura

Taula 2: Exemple fitxer públic de microdades

Algú pot pensar que havent fet aquests tractaments previs a la publicació, ja no hi ha cap risc de revelació i que per tant el fitxer és segur a qualsevol atac extern. Però malauradament, no és així. Per exemple, imaginem que una persona disposa de la informació de la taula 3. A aquesta persona li resultarà molt fàcil (a partir del creuament de dades entre les taules 2 i 3) identificar la malaltia del senyor Antoni.

Nom	Sexe	Ciutat	E.Civil	Edat
Antoni	H	Banyoles	Casat	33

Taula 3: Exemple informació externa

És molt important recordar que davant la publicació de dades sensibles, sempre existirà risc de revelació. L'única manera d'evitar-ho totalment és no publicar absolutament res. Els mètodes de protecció de dades només poden limitar-ho.

Pel que fa a l'àmbit de les microdades tenim, entre d'altres, els següents mètodes:

- Mètodes pertorbatius: És publiquen totes les dades però modificades (el mínim necessari).
  - Microagregació: Orientat a variables contínues. Té com a objectiu fer grups de com a mínim k individus "semblants" i substituir els valors de les variables que poden ajudar a identificar un registre per un valor comú al grup (p.ex: mitjana del grup).
  - Intercanvi de dades: Aplicat a variables numèriques i/o categòriques-ordinals. S'ordenen els valors de les variables i cada valor s'intercanvia amb un altre valor (escollit aleatòriament) entre un rang de valors restringit dintre un percentatge p% del nombre total de registres. Si p petit, utilitat de les dades i risc augmenten. Si p alt, utilitat i risc disminueixen.
- Mètodes no pertorbatius: Només es publiquen una part de les dades originals (màxim possible).
  - Mostreig: Publicació únicament d'una mostra de la població. "Unicitat" a la mostra no implica "unicitat" a la població.
  - Recodificació global: Per a variables categòriques. Consisteix en fer una recodificació de les variables categòriques (agrupant categories) evitant "unicitats" en les dades publicades.
  - Supressió local: Consisteix en eliminar el registre o els valors d'aquelles variables amb alt risc d'identificació.

Existeixen altres mètodes de protecció de microdades que no explicarem en aquest article. El lector interessat por consultat més informació sobre el tema en [7, 8, 14, 20].

D'altra banda, tot i els esforços dedicats al desenvolupament de potents mètodes de protecció de microdades que facin possible la publicació de cada cop més quantitat d'informació, els investigadors necessiten encara més nivell de detall per dur a terme les seves investigacions. Un nivell d'informació que no es pot aconseguir amb els actuals mètodes protecció. Per aquest motiu, els INE's estan darrerament desenvolupant noves modalitats d'accés segur a microdades. El primer pas, va ser la creació de centres d'accés segur a les dades originals (anonimitzades, sense noms ni adreces personals). En aquests centres, els investigadors poden realitzar els seus anàlisis sense la possibilitat d'exportar cap informació a l'exterior. Qualsevol exportació haurà d'estar prèviament acceptada per l'INE. Aquests centres han d'estar equipats amb potents ordinadors i software especialitzat. A més, per tal d'evitar cap revelació de dades personals, no es disposa ni de connexió a Internet (que permeti l'enviament d'emails) ni tampoc possibilitat de gravar dades a dispositius externs (usb, disc o cd's). L'utilització d'impressores o telèfons també presenta un clar risc que pot ser solucionat amb la presència de personal de l'INE que vetlli per la seguretat de la informació. Les variables a les que l'investigador pot accedir pot ser limitat prèviament en funció de la recerca que l'investigador vol dur a terme. Un cop l'investigador finalitza la seva feina, els resultats obtinguts poden ser exportats sempre i quan el responsable de l'INE (o equip de persones de l'INE especialitzats en el tema) donin el seu vist-i-plau. En països com Canada, Holanda, Itàlia o Alemanya, ja s'han creat aquests centres amb resultats molt satisfactoris. La publicació de bons articles va ser possible gràcies només a la disponibilitat d'aquests centres d'accés segur. Tot i això, segur que el lector haurà fàcilment detectat un gran problema en aquesta iniciativa. Es tracta de la necessitat de que l'investigador es trobi físicament en el centre i fora del seu entorn habitual de treball. Una primera solució a aquest problema és la possibilitat de que els investigadors desenvolupin els seus programes en software estadístic especialitzat com SAS, R, SPSS, STATA o d'altres. Aquests programes són enviats als INE's que els executen i analitzen els resultats obtinguts (per tal de descartar qualsevol risc de revelació personal). Un cop l'INE aprova el contingut, retorna els resultats a l'investigador. El problema d'aquesta iniciativa és la pèrdua de temps en la correcció de possibles errors en els programes desenvolupats pels investigadors. En aquest sentit i per tal de minimitzar aquests errors, els INE's poden posar a disposició dels investigadors un petit exemple amb dades inventades però amb l'estructura real del fitxer original. Països com Austràlia o Luxemburg en són un exemple d'execució remota.

I per acabar, tenim l'accés remot segur a dades originals. L'investigador pot connectar-se remotament (per VPN, Virtual Private Network) al servidor dels INE's, on es troben les dades originals i amb una diversitat de programes instal·lats que permetin explotar estadísticament la informació (R, SPSS, SAS, STATA,...). Només els investigadors autoritzats poden connectar-se i treballar en aquest servidor. L'investigador treballaria com si estigués present en el centre de recerca de l'INE, fent més fàcil i còmoda la seva feina. D'altra banda, l'INE tindria control absolut dels resultats que obté l'investigador. Els resultats no sortirien del servidor fins que l'equip de l'INE l'analitzi i doni la seva conformitat.

Protecció de Macrodades

Tot i que cada cel·la d'una taula mostra informació agregada per a varis individus, existeix un risc de revelació de dades individuals, com es pot observar en la figura 1. La taula (a) dóna el salari mig pel creuament de les variables categòriques edat i codi postal. La taula (b) mostra la freqüència d'individus pel creuament de les mateixes variables categòriques. Observem com hi ha només un individu amb codi postal z_2 i interval d'edat $51 - 55$. Per tant, és fàcil per a qualsevol persona externa deduir el salari d'aquesta persona,

en aquest cas 40000€. Amb una freqüència de dos individus, qualsevol d'ells podria també determinar el salari de l'altre. En general, podem tenir problemes de revelació confidencial de dades individuals amb totes aquelles cel·les amb una freqüència baixa d'individus, anomenades sensibles. Existeixen altres regles per a determinar quines cel·les és consideren sensibles [9, 16] que el lector interessat pot consultar.

Les macrodades poden tenir informació del nombre d'individus (taules de freqüències) o informació addicional (suma, mitjana, etc) d'una nova variable numèrica (taules de magnituds) pel creuament de les variables categòriques de cada cel·la. A més a més, segons el signe de les cel·les, tenim taules positives (totes les cel·les amb valor ≥ 0, p.ex: totes les taules de freqüència) o generals (Valors $\in \mathbb{R}$. p.ex: "variació del IPC" entre anys segons "anys" x "autonomies"). Segons l'estructura (depèn de les variables categòriques de creuament i vàlid tant per a taules de freqüències/magnituds i positives/generals) podem tenir taules úniques k-dimensionals, $k \geq 1$, resultat de creuar k variables categòriques (les més populars són les bidimensionals (k=2),p.ex: Taula de freqüències "Sexe" x "Edat" (categoritzada)) o taules jeràrquiques, que són un conjunt de taules que tenen variables de creuament relacionades de forma jeràrquica (p.ex: Jerarquia a nivell de Comarca, Municipi i Codi Postal).

		z_1	z_2	
⋮	...	...	...	...
51–55	...	38000€	40000€	...
56–60	...	39000€	42000€	...
⋮	...	...	...	...

(a)

		z_1	z_2	
⋮	...	...	...	...
51–55	...	20	1 or 2	...
56–60	...	30	35	...
⋮	...	...	...	...

(b)

Figura 1: Exemple d'una taula de dades confidencials, extret de [1]. (a) salari mig per edat i codi postal. (b) Nombre d'individus per edat i codi postal. Si només hi ha una persona a z_2 i interval d'edat $51 - 55$, qualsevol persona externa podria conèixer que aquest individu cobra 40000€ de mitjana. En cas de dos individus, un d'ells podria arribar a deduir el salari de l'altre.

La figura 1 mostra un exemple de dos taules bi-dimensionals. Tot i que la protecció de grans taules bi-dimensionals amb moltes cel·les sensibles no és gens fàcil, podem considerar aquests tipus de taules com el cas més simple de tots. La major complexitat esdevé en el moment en que tenim estructures més complexes, com poden ser les taules multidimensionals o bé un conjunt de taules jeràrquiques (que s'han de protegir de manera conjunta).

Podem classificar els mètodes actuals de protecció de dades tabulars com pertorbatius –canvien el valor original de les cel·les– i els no pertorbatius –no canvien els valor originals–. El mètode no pertorbatiu més utilitzat és el CSP (Cell Supression Problem). Es basa en l'eliminació d'un conjunt (mínim) de cel·les per tal de garantir la protecció de les taules. Entre els mètodes pertorbatius hi ha una tècnica recent anomenada CTA (Controlled Tabular Adjustment) que donada una taula de dades busca la taula "segura" més propera (que no revela informació confidencial dels individus a partir dels quals s'ha calculat la taula).

Tots dos mètodes es basen en la formulació i resolució d'un difícil problema d'optimització. Qualsevol taula o llista de taules, de qualsevol dimensió, mida i estructura, la podem representar com un vector de cel·les $a_i, i = 1, \ldots, n$, que satisfan un conjunt m de restriccions lineals $Aa = b$ (on a representa el vector de cel·les a_i's, $b \in \mathbb{R}^m$ i $A \in \mathbb{R}^{m \times n}$) i un conjunt un conjunt $\mathcal{P} = \{i_1, i_2, \ldots, i_p\} \subseteq \{1, \ldots, n\}$ d'índexs de cel·les sensibles o confidencials (no es pot publicar el seu valor original). La majoria de taules acostumen a tenir valors de cel·les positius ($a \geq 0$).

Mètode CTA: El mètode CTA troba els valors segurs x_i més propers a $a_i, i = 1, \ldots, n$, que satisfan

$Ax = b$. Suposem que qualsevol atacant extern coneix els límits inferiors i superiors, l_{a_i} i u_{a_i} respectivament, de cada cel·la. En cas de no existir aquest coneixement previ, $l_{a_i} = -\infty$ ($l_{a_i} = 0$ si considerem taules positives $a \geq 0$) i $u_{a_i} = +\infty$. Considerem segur el valor d'una cel·la sensible si donats uns nivells inferiors i superiors de protecció per a cada cel·la sensible $i \in \mathcal{P}$, lpl_i i upl_i respectivament, els valors publicats satisfan $x_i \geq a_i + upl_i$ o $x_i \leq a_i - lpl_i$. Podem formular el problema d'optimització CTA (d'acord a una determinada distància L) com:

$$\begin{array}{ll} \min\limits_{x} & ||x - a||_L \\ \text{s.a} & Ax = b \\ & l_{a_i} \leq x_i \leq u_{a_i} \quad i = 1, \ldots, n \\ & x_i \leq a_i - lpl_i \text{ o } x_i \geq a_i + upl_i \quad i \in \mathcal{P}. \end{array} \tag{1}$$

A la taula 2 es mostra un petit exemple de protecció amb el mètode CTA.

	z_1	z_2	z_3	$Total$
E_1	20	24	28	72
E_2	38	38	**40**	116
E_3	40	39	42	121
$Total$	98	101	110	309

(a)

	z_1	z_2	z_3	$Total$
E_1	15	24	33	72
E_2	43	38	**35**	116
E_3	40	39	42	121
$Total$	98	101	110	309

(b)

	z_1	z_2	z_3	$Total$
E_1	25	24	23	72
E_2	33	38	**45**	116
E_3	40	39	42	121
$Total$	98	101	110	309

(c)

Figura 2: Exemple protecció d'una taula amb el mètode CTA. En negreta les cel·les sensibles, amb nivells de protecció superior/inferiors iguals a 5(lpl,upl). (a) Taula original. (b) Taula protegida en sentit "lower, lpl=5". (c) Taula protegida en sentit "upper, upl=5". Es pot observar com el valor de la cel·la sensible té, en tots dos casos, una desviació de 5 unitats.

El problema d'optimització CTA es por formular com a un problema MILP (Mixed Integer Linear Programming) incorporant, com a variables binàries del problema, la decisió de protegir en sentit "lower" o "upper". Aquest és un problema molt més difícil de resoldre i actualment s'està treballant en la recerca de tècniques heurístiques [13] i exactes [4] per millorar la eficiència en la cerca de solucions òptimes. A continuació detallem la formulació del problema MILP per a CTA:

CTA com a MILP: Si definim les desviacions dels nous valors respecte els originals com $z_i = x_i - a_i, \quad i = 1, \ldots, n$ — de manera similar $l_{z_i} = l_{x_i} - a_i$ i $u_{z_i} = u_{x_i} - a_i$ — podem reformular el problema (1) com:

$$\begin{array}{ll} \min\limits_{z} & ||z||_L \\ \text{s.a} & Az = 0 \\ & l_{z_i} \leq z_i \leq u_{z_i} \quad i = 1, \ldots, n \\ & z_i \leq -lpl_i \text{ o } z_i \geq upl_i \quad i \in \mathcal{P}, \end{array} \tag{2}$$

on $z \in \mathbb{R}^n$ representa el vector de desviacions. Aquesta formulació no utilitza els valors originals $a_i, i = 1, \ldots, n$, per tant, dues taules diferents però amb les mateixes relacions lineals representades per la matriu A i els mateixos límits l_z, u_z, lpl i upl, poden ser protegides amb les mateixes perturbacions z i per tant només cal resoldre una vegada el problema de minimització. Utilitzant la distància L_1 i uns costos $w \geq 0$ associats a cada cel·la reformulem (2) com:

$$\begin{array}{ll} \min\limits_{z} & \sum\limits_{i=1}^{n} w_i|z_i| \\ \text{s.a} & Az = 0 \\ & l_{z_i} \leq z_i \leq u_{z_i} \quad i = 1, \ldots, n \\ & z_i \leq -lpl_i \text{ o } z_i \geq upl_i \quad i \in \mathcal{P}. \end{array} \tag{3}$$

Per convertir (3) en un problema de programació lineal equivalent substituïm cada z_i per la diferència de dues variables no negatives, z_i^+ i z_i^-, associades a les desviacions positives i negatives respectivament:

$$z_i = z_i^+ - z_i^-, \quad i = 1, ..., n, \tag{4}$$

i formulem (3) com

$$\begin{array}{ll} \min\limits_{z^+, z^-} & \sum\limits_{i=1}^{n} w_i(z_i^+ + z_i^-) \\ \text{s.a} & A(z^+ - z^-) = 0 \\ & l_{z_i} \leq z_i^+ - z_i^- \leq u_{z_i} \quad i = 1, \ldots, n \\ & z_i^+ - z_i^- \leq -lpl_i \text{ o } z_i^+ - z_i^- \geq upl_i \quad i \in \mathcal{P}, \\ & z^+ \geq 0, \quad z^- \geq 0. \end{array} \tag{5}$$

Podem definir variables binàries $y \in \mathbb{R}^p$ que indiquen el sentit de protecció de les cel·les sensibles i transformen el problema (5) en un (difícil) problema de programació lineal entera mixta (MILP)

$$\begin{array}{ll} \min\limits_{z^+, z^-, y} & \sum\limits_{i=1}^{n} w_i(z_i^+ + z_i^-) \\ \text{s.a} & A(z^+ - z^-) = 0 \\ & 0 \leq z_i^+ \leq \quad u_{z_i} \quad i \notin \mathcal{P} \\ & 0 \leq z_i^- \leq -l_{z_i} \quad i \notin \mathcal{P} \\ & upl_i \, y_i \leq z_i^+ \leq u_{zi} \, y_i \quad i \in \mathcal{P} \\ & lpl_i(1 - y_i) \leq z_i^- \leq -l_{zi}(1 - y_i) \quad i \in \mathcal{P}. \end{array} \tag{6}$$

Quan $y_i = 1$ el sentit de protecció de la cel·la a_i és "upper" i la restricció considerada serà $upl_i \leq z_i^+ \leq u_{zi}$ amb $z_i^- = 0$; quan $y_i = 0$ el sentit de protecció serà "lower" amb $z_i^+ = 0$ i $lpl_i \leq z_i^- \leq -l_{z_i}$. Més informació del mètode CTA a [1, 5].

Mètode CSP: Donat una taula i un conjunt S de cel·les sensibles o primàries, el mètode CSP troba el mínim conjunt addicional de cel·les secundàries C que cal eliminar per tal de garantir la privacitat del conjunt S de cel·les sensibles o primàries. El fet de trobar el mínim conjunt C afavoreix l'objectiu d'aconseguir la mínima pèrdua d'informació en la publicació posterior de la taula. El mètode CSP es pot plantejar també com a un problema MILP (Mixed Integer Linear Programming). També existeixen tècniques exactes ([10, 11, 17]) i heurístiques ([2, 6, 12, 15, 21]). En aquest article, no donarem més detalls de la modelització del problema MILP. Podem trobar més informació a les referències citades anteriorment. Si mostrem però, a la figura 3, un petit exemple molt il·lustratiu del mètode.

	z_1	z_2	z_3	$Total$
E_1	20	24	28	72
E_2	38	38	**40**	116
E_3	40	39	42	121
$Total$	98	101	110	309

(a)

	z_1	z_2	z_3	$Total$
E_1	*	24	*	72
E_2	*	38	*	116
E_3	40	39	42	121
$Total$	98	101	110	309

(b)

Figura 3: Exemple protecció d'una taula amb el mètode CSP. En negreta les cel·les primàries/sensibles, amb nivells de protecció superior/inferiors iguals a 5 (lpl,upl). (a) Taula original. (b) Taula protegida. Amb *, les cel·les primàries i secundàries eliminades.

Agraïments

Al Dr. Jordi Castro (Universitat Politècnica de Catalunya, UPC) i al Sr. Enric Ripoll (Institut d'Estadística de Catalunya, Idescat) per les seves observacions i suggeriments.

Referències

[1] J. Castro (2006), Minimum-distance controlled perturbation methods for large-scale tabular data protection, European Journal of Operational Research, 171 39-52

[2] J. Castro (2007), A shortest paths heuristic for statistical disclosure control in positive tables, INFORMS Journal on Computing 19(4) 520-533

[3] J. Castro and D. Baena (2006). Automatic Structure Detection in Constraints of Tabular Data, Lecture Notes in Computer Science, 4302 12-24.

[4] J. Castro and D. Baena (2008). Using a mathematical programming modeling language for optimal CTA, Lecture Notes in Computer Science, 5262 1-12.

[5] J. Castro, A. González and D. Baena (2009), Users and programmers manual of the RCTA package, Technical Report DR 2009/01, Dept. of Statistics and Operations Research, Universitat Politècnica de Catalunya, 2009.

[6] L.H. Cox (1995). Network models for complementary cell suppression, Journal of the American Statistical Association 90 1453-1462.

[7] J.Domingo-Ferrer, L.Franconi, (eds.) (2006). Lecture Notes in Computer Science. Privacy in Statistical Databases, 4302, Springer, Berlin.

[8] J.Domingo-Ferrer, Y.Saigin, (eds.) (2008). Lecture Notes in Computer Science. Privacy in Statistical Databases,5262, Springer, Berlin.

[9] J. Domingo-Ferrer and V. Torra (2002), A Critique of the Sensitivity Rules Usually Em- ployed for Statistical Table Protection, International Journal of Uncertainty Fuzziness and Knowledge-Based Systems, 10(5), 545-556.

[10] M. Fischetti and J.J. Salazar-González (1999), Models and algorithms for the 2-dimensional cell suppression problem in statistical disclosure control. Mathematical Programming, 84, 283-312.

[11] M. Fischetti and J.J. Salazar-González (2001), Solving the cell suppression problem on tabular data with linear constraints, Management Science 47,1008-1026.

[12] S. Giessing and D. Repsilber (2002), Tools and strategies to protect multiple tables with the GHQUAR cell suppression engine, Lecture Notes in Computer Science 2316 181 192.

[13] A. González and J. Castro (2011), A heuristic block coordinate descent approach for controlled tabular adjustment, Computers and Operations Research 38, 1826-1835.

[14] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, R. Lenz, J. Naylor, E. Schulte- Nordholt, G. Seri and P.P. de Wolf (2010), Handbook on Statistical Disclosure Control (v. 1.2), Network of Excellence in the European Statistical System in the field of Statistical Disclosure Control. Available on-line at http://neon.vb.cbs.nl/casc/SDC_Handbook. pdf.

[15] J.P. Kelly, B.L. Golden and A.A. Assad (1992), Cell suppression: disclosure protection for sensitive tabular data. Networks, 22 28-55.

[16] D.A. Robertson and R. Ethier (2002), Cell suppression: experience and theory, Lecture Notes in Computer Science, 2316 8-20.

[17] J.J. Salazar-González (2004), Mathematical models for applying cell suppression methodology in statistical data protection, European Journal of Operational Research 154 740-754.

[18] J.A.Sànchez, J.Urrutia, E.Ripoll (2004), Trade-Off between Disclosure Risk and Information Loss Using Multivariate Microaggregation: A Case Study on Business Data, Privacy in Statistical Databases,307-322.

[19] J.Urrutia, E.Ripoll (2002), Empirical Evidences on Protecting Population Uniqueness at Idescat, Inference Control in Statistical Databases, 213-230.

[20] L. Willenborg and de T. Waal (2000), Elements of Statistical Disclosure Control, Lecture Notes in Statistics. Elements of Statistical Disclosure Control, 155, Springer, New York.

[21] P.P de Wolf (2002), HiTaS: A heuristic approach to cell suppression in hierarchical tables. Lecture Notes in Computer Science, 2316 74-82.