PERTORBACIONS ALEATÒRIES AMB COMPENSACIÓ: UNA TÈCNICA PER A LA PROTECCIÓ D'INFORMACIÓ ESTADÍSTICA CONFIDENCIAL

J. TURMO i SOLDEVILA

Institut d'Estadística de Catalunya

Garantir la preservació de la confidencialitat de les dades estadístiques facilitades pels informants és una obligació recollida per l'ordenament jurídic i inherent a la professió estadística. La forma més usual de presentació de resultats en l'estadística oficial són les taules de dades agregades. Després de fixar els objectius que ha de satisfer un mètode de censura òptim per el tractament de les macrodades sensibles, es valoren les diferents tècniques de supressió i d'enmascarament existents. Finalment, es proposa la introducció de pertorbacions aleatòries amb compensació com un mètode de censura, i es presenta una implementació que simplifica la proposta original.

Random perturbations by compensation: a method to protect confidential statistical information .

Key words: confidencialitat estadística, risc de revelació, mètodes de censura per a macrodades, pertorbacions aleatòries, cel·les conflictives.

INTRODUCCIÓ

La finalitat d'aquest article és analitzar l'adequació del mètode de protecció de macrodadades basat en la introducció de pertorbacions aleatòries a les fre-

-Article rebut el juny de 1993.
-Acceptat el febrer de 1994.

qüències originals i comparar-lo amb les altres tècniques existents. La manera més comuna de presentació de les macrodades són les taules estadístiques de freqüències. Es proposa un mètode, la introducció de pertorbacions aleatòries amb compensació, i es presenta, sense entrar en una descripció tècnica detallada, una metodologia i la seva implementació informàtica, que disminueix l'impacte global de la tècnica original, és transparent per als usuaris i aconsegueix resultats reproduïbles.

El document està estructurat en cinc apartats. El primer d'aquests tracta de definir els objectius que ha de satisfer el sistema desitjat, tant des del punt de vista de la preservació de la confidencialitat com de la qualitat de les macrodades resultants. En el segon es descriuen i valoren, succintament, diferents tècniques existents per censurar les macrodades (matrius de freqüències). En el tercer s'exposen diferents criteris utilitzats per considerar que una cel·la de la matriu comporta un risc de revelació. En el quart apartat es descriu el mètode proposat i s'indiquen els punts principals de l'aplicació informàtica que l'implementa. Finalment, en el cinquè punt es valoren els resultats de l'aplicació en un arxiu de macrodades de prova, per mitjà de la presentació de dos llistats breus que mostren els resultats de la seva aplicació: el primer, una taula per a un municipi de 5.000 habitants abans i després del procés de censura; i el segon, els informatius detallats d'aquest procés de censura aplicat al conjunt de tots els municipis de Catalunya.

Finalment, s'inclou una secció amb la bibliografia detallada i actualitzada basada en les diferents metodologies, implementacions i criteris proposats.

1. OBJECTIUS

Es presenten els objectius previstos agrupats en cinc apartats: minimització del risc de revelació indirecta, màxima aproximació als nivells d'informació originals, garantia de la qualitat de les dades, homogeneïtat entre totes les taules del pla de tabulació i inalteració de les previsions d'explotacions.

Garantir la preservació de la confidencialitat de les respostes dels informants originals (siguin persones físiques o jurídiques) i, molt especialment de les seves característiques sensibles, és un requeriment legal i una obligació inherent a la professió estadística. L'objectiu plantejat se centra a minimitzar el risc de revelació indirecta de dades originals dels informants, això implica dues actuacions concretes:

1. Evitar que per mitjà de l'anàlisi de les freqüències molt baixes d'una taula es puguin conèixer característiques privades d'aquests individus o, fins i tot, confrontar-los amb registres d'altres arxius que continguin identificadors directes.

2. Evitar l'existència de subpoblacions úniques dins l'espectre de les ocurrències d'una variable de domini públic per a un àmbit geogràfic concret, ja que la seva existència permetria inferir altres característiques no conegudes, comunes a tots els individus d'aquest conjunt.

El dret a accedir a la màxima informació possible mantenint els criteris de correcció tècnica es pot veure afectat en censurar algunes dades primàries. El segon objectiu tracta de mantenir aquest principi acostant, el més possible, el nivell d'informació disponible i tècnicament vàlida abans i després del procés de censura. La concreció d'aquest objectiu passa per:

3. Garantir la concordança de l'estructura de les distribucions de freqüències marginals i conjuntes prèvia i posterior al procés de censura, fins als nivells de desagregació sectorial i geogràfica significatius per a les dades originals.

4. Assegurar que els mètodes de censura emprats no invalidin les dades resultants per algun dels tipus d'anàlisi temàtica (demogràfica, econòmica, etc.) usuals.

El concepte de "qualitat" aplicat a les dades estadístiques és una matèria que encara avui genera controvèrsia per la seva amplitud i per la complexitat d'una definició precisa. Sense endinsar-nos en el tractament de la qualitat de les dades censurades en sentit global, el fet de garantir l'exactitud de les dades censurades respecte les originals sí que constitueix un objectiu que cal abastar. De manera més concreta:

5. Garantir que el procés de censura afecti el mínim nombre de cel·les, assumint-ne alteracions globals no significatives.

Un altre dels objectius bàsics, dins el conjunt dels encreuaments de les variables previstes en un pla de difusió, és que els totals marginals de totes les variables siguin coherents entre ells i, si és el cas, amb les xifres oficials externes a l'enquesta. Les actuacions lligades a aquest objectiu són:

6. Garantir que els totals de població de les taules censurades coincideixin amb els totals oficials per als diferents nivells de desagregació previstos en el pla de difusió.

7. Aconseguir que els totals marginals per a totes les variables incloses en les taules d'entrada múltiple, en finalitzar el procés de censura, siguin coherents entre ells i amb els recomptes unidimensionals.

L'últim grup d'objectius se centra a evitar que la implementació de les mesures de preservació de la confidencialitat distorsioni els processos d'explotació usuals. En concret, això implica dues actuacions complementàries:

8. Evitar que el desenvolupament i l'aplicació dels mètodes de censura afectin negativament el calendari previst per a la difusió dels productes.

9. Aconseguir que la implementació informàtica dels processos de censura no comporti alteracions ni en l'anàlisi ni en el desenvolupament dels productes informàtics emprats per explotar els arxius estadístics.

2. TÈCNIQUES EXISTENTS

Actualment, en la bibliografia especialitzada es descriuen diversos mètodes enfocats a minimitzar els riscos de revelació indirecta d'informació confidencial en la difusió de dades estadístiques agregades o microdades. Pràcticament tots aquests es poden agrupar en dues categories conceptuals[1]: aquelles que suprimeixen la informació conflictiva i els que l'emmascaren. Dins aquests dos grans grups, les discordances entre les diferents propostes se centren, primordialment, en la implementació i, sobretot, en el refinament dels mètodes d'optimització necessaris per determinar el mínim nombre de cel·les que s'han de tractar, de tal manera que es garanteixi el fet que no es puguin deduir els valors originals a partir de la informació censurada.

El principi bàsic dels mètodes de *supressió*[2] consisteix a eliminar aquelles cel·les que determinem que són conflictives, que presenten un risc no acceptable de la revelació de la informació confidencial. A aquest conjunt de dades eliminades es denomina "supressions primàries". També serà necessari, com a mínim, eliminar alguna altra cel·la de la mateixa filera i de la mateixa columna —en el cas de les matrius bidimensionals— a què pertanyen cadascuna de les cel·les suprimides en el procés primari, ja que en cas contrari per simple diferència del total marginal es podria deduir la dada original objecte de censura. El conjunt de cel·les eliminades per aquest motiu s'anomena "supressions secundàries". Fer que el nombre d'aquestes supressions secundàries o la seva inèrcia —suma dels valors de les cel·les afectades— sigui mínima, representa resoldre un problema

[1] Vegeu Cox (1988).

[2] Vegeu Börjesson (1987); Citteur i Willenborg (1991); Cox (1984); Cox (1986); Dalenius (1983); Griffin, Navarro i Flores-Baez (1989); Lougee-Heimer (1989); Repsilber (1992); Wolf (1990).

complex de programació lineal entera[3], especialment si les dimensions de les taules que s'han de censurar són grans. S'han proposat mètodes alternatius basats en la teoria de xarxes[4] que han estat aplicats amb èxit en operacions pilot, si bé la complexitat dels algoritmes fa que les implementacions actuals siguin notablement complexes i la seva execució comporti un elevat consum de temps d'ordinador.

Els mètodes d'emmascarament[5], com el seu nom indica, consisteixen a alterar els valors originals de les dades conflictives, de tal manera que a partir de les taules de difusió pública no es puguin deduir els valors originals d'aquestes cel·les.

Les dues tècniques més comunes, dins els mètodes d'emmascarament, són la d'arrodoniment i la que consisteix a introduir pertorbacions aleatòries a certs elements de la taula. La primera es fonamenta en arrodonir tots els elements de la matriu de freqüència a valors múltiples d'un valor base donat (normalment entre 5 i 10). Un refinament d'aquesta tècnica és l'anomenat arrodoniment controlat[6] que pretén que les diferències positives i negatives es compensin per a cada filera i columna. L'aplicació d'aquestes tècniques implica recalcular els totals marginals.

La segona consisteix a pertorbar els valors conflictius, sumar-los-hi un nombre aleatori calculat dins un interval determinat (simètric, asimètric, constant o variable) al voltant de zero. Es pot fer públic o no quines cel·les han estat modificades. De la mateixa manera que en els mètodes de supressió, podrà ser necessari censurar altres cel·les secundàries, encara que utilitzant aquestes tècniques només serà imprescindible per no revelar els valors de les cel·les originals pertorbades, en el cas que es faci públic a quines cel·les afecta la censura. De tota manera, és aconsellable realitzar compensacions o emmascaraments secundaris perquè els totals marginals de les dades censurades siguin coincidents amb els dels originals. De no fer-ho s'han de recalcular els totals marginals. Un enfocament presentat recentment[7], anomenat pertorbació amb compensació, intenta abordar aquest objectiu aconseguint que la suma de les pertorbacions i compensacions aplicades als elements de cada filera o columna (en el cas de taules bidimensionals) sigui nul·la. Aquesta proposta no altera els totals marginals originals, si bé la seva implementació pot ser complexa. També existeixen propostes que postulen pertorbar tots els elements de la matriu.

---

[3] Vegeu Cox, Fagan, Greenberg i Hemmig (1987); Zayatz (1992,a) i (1992,b).

[4] Vegeu (1992); Kelly, Golden i Assad (1992); Sullivan i Zayatz (1991).

[5] Vegeu Börjesson (1987); Cox (1984); Dalenius (1986); Greenberg (1986) i (1988); Heer (1992).

[6] Cox i Ernst (1980).

[7] Apple i Hoffmann (1992).

Actualment, existeix una línia d'investigació incipient que té com a finalitat aportar una metodologia coherent que pugui ser aplicada en la localització de les subpoblacions conflictives, tant per a les dades agregades com per a les microdades. Aquest enfocament està basat en els principis de la teoria de la informació i el seu objectiu central és minimitzar la pertorbació global (o "soroll" en la terminologia emprada) que és necessària d'introduir en l'arxiu original per aconseguir que no es pugui deduir informació confidencial essencialment nova, respecte a la informació de coneixement públic legal. Un pre-requisit és disposar de la relació d'arxius informatitzats considerats de caràcter públic i de la descripció detallada de les variables que contenen. Aquest enfocament promet reduir l'impacte global dels processos de censura en les dades estadístiques, mantenint i fins i tot millorant, en certs casos, el nivell de protecció de les característiques privades dels informants.

3. CRITERIS PER DETERMINAR LES CEL·LES CONFLICTIVES

Independentment de la tècnica triada per efectuar el procés de censura, cal utilitzar uns criteris reproduïbles per determinar quines són les cel·les conflictives —llevat dels casos en què s'utilitzen procediments que afecten totes les dades—. Els criteris[8], no excloents, utilitzats majoritàriament per decidir que una cel·la és conflictiva, són de tres tipus:

a. Que la seva freqüència sigui inferior a un cert valor llindar, normalment entre 3 i 15 (valor mínim).

b. Que la freqüència d'un element representi, per ella mateixa, un percentatge sobre el total d'alguna de les dimensions de la matriu superior a un límit fixat, normalment entre el 75 i 90% (predominança).

c. Que el conjunt dels efectius d'una filera o columna estigui assignat a una única cel·la (poblacions úniques).

Una proposta anomenada en ocasions $N - K$ i utilitzada especialment en estadístiques de tipus econòmic és la que consisteix en una combinació dels criteris a i b on N representa la freqüència i K el percentatge de predominança a partir dels quals se censuren les dades.

La fixació dels valors i percentatges llindars i de si cal analitzar o no els casos de poblacions úniques pot dependre de la tipologia de l'arxiu o, fins i

[8]Blien i Wirth (1992); Citteur i Willenborg (1991); Dalenius (1983); Duncan i Lambert (1986); Greenberg (1990); Repsilber (1992).

tot, del tipus de variable[9]. L'ordenament jurídic d'alguns països fixa, a nivell reglamentari, el valor d'aquests paràmetres, en d'altres estan inclosos en els codis interns sobre secret estadístic dels seus organismes estadístics oficials o la praxi consolidada.

Un altre paràmetre que cal tenir en compte és si el procés de censura pot afectar o no els valors nuls —zeros—. Aquesta consideració és especialment important en el cas de les supressions o emmascaraments secundaris, ja que el fet de poder modificar ocasionalment alguna cel·la de freqüència zero, permet que es puguin anonimitzar les subpoblacions úniques en tots els casos. També pot ser necessari en certs casos modificar alguna cel·la a zero si es desitgen obtenir conjunts de taules amb els totals marginals coherents sense distorsions significatives en l'estructura de les distribucions de freqüències. Normalment els zeros no són considerats com a valors primaris que s'haurien de censurar.

A continuació, es presenta un quadre il·lustratiu de la qüestió, el qual incorpora una primera valoració del grau de satisfacció dels objectius proposats en l'apartat 1 d'aquest article, per part de cadascuna de les tècniques descrites anteriorment:

	Objectius								
Tècniques	1	2	3	4	5	6	7	8	9
Supressió									
secundàries directes	A	B/N	M/B	M/B	B	A	A	A	A
optimitzant secun.	A/M	B/N	M	M	M	A	A	B	A
Arrodoniment	A	N	M	M	B	N	N	A	A
Pertorb. aleatòries									
(informant cel·les)	M	N	A	A/M	A/M	____	____	A	M
(sense informar)	A/M	N	A	M	A/M	____	____	A	A
també secundàries	A/M	A/N	A/M	M	M	A	A/N	A/B	A/M
totes les cel·les	A/M	N	A/M	M/B	B	A/N	A/N	A	A

Notes: grau de satisfacció dels objectius: A (alt), M (mitjà), B (baix), N (nul).
Els números indicats es corresponen amb els objectius tipificats a l'apartat 1.

En alguns casos es presenta una doble valoració que depèn de les diferents especificacions de detall possibles per a cada mètode; en especial en l'objectiu 2 —evitar l'existència de subpoblacions úniques— depèn de si l'aplicació de cada tècnica pot o no pot afectar les caselles de freqüència zero. En el cas del grau

[9] Gusfield (1990).

d'acompliment dels objectius 6 i 7 —garantir l'homogeneïtat dels totals globals i marginals— dependrà de si s'apliquen o no mesures d'ajustament de les matrius. Per a aquests objectius 6 i 7, a més, és indiferent que el mètode aplicat prevegi informar o no de quines són les cel·les pertorbades.

S'estudien els resultats de la valoració dels diferents mètodes, el ventall de possibles tècniques idonies es pot restringir a dues: la de supressió (optimitzant les cel·les secundàries afectades) i la consistent a introduir pertorbacions aleatòries compensant els seus efectes mitjançant pertorbacions secundàries, sense informar els usuaris de quines han estat les cel·les afectades.

Valorant la incomoditat i poca experiència dels potencials usuaris dels productes resultants en analitzar taules agregades amb informació suprimida, s'optaria per proposar la pertorbació aleatòria d'aquelles caselles que comporten un risc de revelació indirecta. Tanmateix es proposa aplicar dos refinaments: tractar prèviament les poblacions úniques en l'àmbit de filera/columna i ajustar els totals marginals de les taules mitjançant mètodes de compensació.

4. DESCRIPCIÓ DEL MÈTODE PROPOSAT

Com s'ha avançat en l'apartat anterior, es proposa aplicar un mètode mixt, la base del qual sigui la introducció de pertorbacions aleatòries en aquelles caselles que presentin una freqüència igual o menor a un valor donat. No es considera el fet d'informar de quines han estat les caselles afectades pel procés de censura. Addicionalment, en el cas de matrius de dimensió dos o superior, s'ha cregut convenient implementar un procediment auxiliar per als casos en què tots els individus d'una ocurrència d'una variable de coneixement públic estiguin inclosos en una, i únicament una, ocurrència d'una altra de les variables d'encreuament i aquesta sigui del tipus sensible. Les possibles diferències que es derivessin de l'aplicació d'aquestes pertorbacions primàries es compensaran mitjançant mètodes d'ajustament de matrius multidimensionals.

Amb l'aplicació conjunta d'aquestes dues tècniques i fixant valors adequats de la freqüència màxima que s'ha de pertorbar i de l'amplitud del nombre aleatori que s'ha d'addicionar, es creu que es pot abastar un nivell de preservació de la confidencialitat satisfactori que en la pràctica s'ha de traduir en fer irrellevant el risc de revelació de la "identitat" de l'informant original, fins i tot aplicant tècniques d'investigació sofisticades[10].

[10] Blien i Wirth (1992).

L'aplicació pràctica comprèn quatre fases seqüencials, encara que en la implementació informàtica, desenvolupada per fer les proves que es presenten més endavant, constitueixin una única aplicació:

a. Detecció de les cel·les conflictives: freqüències menors que N i poblacions úniques.

b. Redistribució d'una part de la freqüència de les poblacions úniques segons el model previst per a cadascuna de les variables.

c. Introducció de pertorbacions aleatòries en les caselles amb valor igual o inferior a N.

d. Aplicació dels procediments de compensació multidimensional fins a obtenir els totals marginals originals de les taules.

En aquest context, per població única s'entén aquell element del conjunt dels que configuren un vector en qualsevol dimensió de la matriu de freqüència que sigui l'únic que presenti un valor diferent de zero per a aquest vector. La detecció d'aquestes caselles i de les que presenten valors menors del llindar fixat, es fa conjuntament.

La redistribució d'una part d'aquestes freqüències, s'efectua tenint en compte una màscara lògica que recull les regles d'incompatibilitat i les distribucions conjuntes globals per a les variables presents en la taula que s'estigui tractant. Es preveu repartir un percentatge dels efectius de la cel·la que constitueix la població única, amb una cota inferior mínima. Aquest procés no ha d'afectar les freqüències per sota del valor llindar N.

La introducció de pertorbacions aleatòries s'efectua sumant una variable aleatòria centrada en zero i amb una dispersió fixada externament a les cel·les conflictives. Es consideren dos refinaments: un per als casos en què el valor resultant de l'aplicació d'aquesta pertorbació sigui un valor inferior a zero, l'altre per quan sigui zero.

Els procediments de compensació que s'han aplicat són de dos tipus: el primer l'aplicat solament en una dimensió de la matriu (la que tingui més elements), està basat en el mètode de la ponderació entera de les restes resultants de l'expansió dels totals dels vectors censurats sobre els originals. Per a la resta de les dimensions s'aplica la compensació dels parells de valors recíprocs. A cada compensació es veuran implicades $2^{**}\ d$ cel·les, on d és la dimensió de la matriu de dades. En una visió geomètrica representarà actuar en els valors situats en els vèrtexs d'un paral·lelogram; en el cas d'una matriu de dues dimensions, en els d'un paral·lelepípede per a dimensió 3, etc.

5. VALORACIÓ DELS RESULTATS EXPERIMENTALS

Per avaluar la bondat del mètode proposat i la viabilitat de la seva aplicació experimental, s'ha efectuat una prova exhaustiva amb una taula encreuada de freqüències (lloc de naixement de la població per nivell d'instrucció) derivada de l'explotació dels padrons municipals d'habitants de 1986 de Catalunya, de la qual es presenten els resultats concrets per a un municipi de 5.000 habitants i les diferents dades informatives que s'han obtingut al llarg del procés de censura si s'aplica a tots els municipis de Catalunya.

Si comparem les Taules A i B, abans i després del procés de censura, es pot observar i valorar l'impacte del procés. Per dur a terme el procés, s'ha considerat el tractament de les poblacions úniques (redistribuint un percentatge del 10% amb un mínim de 2 individus) i pertorbar els elements de freqüència 5 i inferior amb una pertorbació d'amplitud 3.

Els informatius derivats d'aquest procés, que es presenten a les Taules C.1 a C.5, recullen el resultat global d'aplicar el mètode, amb els criteris exposats en el paràgraf anterior, a les taules dels 940 municipis de Catalunya. S'ha volgut aplicar en aquestes condicions extremes (677 municipis tenen menys de 2.000 habitants i es tracta d'una taula extensa) per valorar l'impacte, fins i tot en condicions molt desfavorables. Els totals de Catalunya han estat recalculats a partir de la suma de les taules de tots els municipis.

La Taula C.1 mostra el nombre de caselles amb una diferència determinada entre els valors originals i cada fase de procés; i, també, el nombre de caselles conflictives (poblacions úniques i freqüències baixes). La Taula C.2 ofereix el resultat des del punt de vista dels valors de les cel·les afectades, de les inèrcies i de les diferències originades. En les quatre taules següents, s'il·lustren les diferències per a cada casella bàsica corresponent a la taula "Total Catalunya", per a cada fase del procés, i en valor absolut i percentatge sobre el valor de la casella.

Vistos els resultats, es pot afirmar, que l'aplicació del mètode proposat suposa que el risc de revelació d'informació confidencial esdevingui irrellevant, tot mantenint uns nivells d'informació òptims. Així mateix, no distorsiona la forma de treballar usual amb taules de dades estadístiques per part de l'usuari final.

Taula A **Població de dret per lloc de naixement i nivell d'instrucció. (1/2)**
Municipis
Selecció

Àmbit	total	nascuts Catal	nascuts Barcel	nascuts Girona	nascuts Lleida	nascuts Tarrag	nascuts r estat	nascuts Andal	nascuts Aragó	nascuts Astur.	nascuts Balears	nascuts Canàries	nascuts Cantàb
Fix													
total	5 003	3 808	219	3	165	3 521	1 033	302	250	10	1	1	12
dificult llegir/escriure	532	382	26	2	20	334	150	43	24	5	0	0	0
primària incompleta	1 501	1 124	70	1	47	1 006	364	115	91	0	0	0	1
e g b primera etapa	1 463	1 163	53	0	49	1 061	283	91	83	1	1	0	3
e g b segona etapa	476	381	17	0	18	346	85	22	16	0	0	0	4
f p primer grau	371	333	10	0	4	319	33	7	13	0	0	0	1
f p segon grau	127	112	5	0	6	101	14	3	4	0	0	0	0
b.u.p. / c.o.u.	263	215	18	0	6	191	42	9	5	0	0	1	1
títol mitjà	144	116	7	0	12	97	26	8	8	1	0	0	1
títol superior	115	74	10	0	3	61	33	3	8	3	0	0	1
no hi consta	11	8	3	0	0	5	3	1	0	0	0	0	0

Taula A **Població de dret per lloc de naixement i nivell d'instrucció. (2/2)**
Municipis
Selecció

Àmbit	nascuts Cast - M	nascuts Cast - II	nascuts P.Valen	nascuts Extrem	nascuts Galicia	nascuts Madrid	nascuts Murcia	nascuts Navarra	nascuts P.Basc	nascuts Rioja	nascuts Ceu/Mel	nascuts estran	no hi consta
Fix													
total	64	99	62	65	17	21	96	7	25	0	1	58	4
dificult llegir/escriure	9	10	9	13	7	7	18	5	0	0	0	0	0
primària incompleta	20	28	19	23	5	5	50	1	6	0	0	12	1
e g b primera etapa	19	25	14	16	3	2	20	0	5	0	0	17	0
e g b segona etapa	8	18	2	8	0	1	1	0	5	0	0	9	1
f p primer grau	1	1	3	0	0	0	5	0	2	0	0	5	0
f p segon grau	5	0	2	0	0	0	0	0	0	0	0	1	0
b.u.p. / c.o.u.	2	6	7	3	0	4	0	0	3	0	1	5	1
títol mitjà	0	4	1	1	1	2	1	0	0	0	0	2	0
títol superior	0	7	5	1	1	0	1	1	2	0	0	7	1
no hi consta	0	0	0	0	0	0	0	0	2	0	0	0	0

Taula B **Població de dret per lloc de naixement i nivell d'instrucció -dades censurades-. (1/2)**
Municipis
Selecció

Àmbit	total	nascuts Catal.	nascuts Barcel.	nascuts Girona	nascuts Lleida	nascuts Tarrag.	nascuts r. estat	nascuts Andal.	nascuts Aragó	nascuts Astúr	nascuts Balears	nascuts Canàries	nascuts Cantàb.
Fitx													
total	5 003	3 908	219	3	185	3 521	1 033	302	250	10	1	1	12
dificult llegir/escriure	532	385	27	2	21	335	147	44	25	4	0	0	0
primària incompleta	1 501	1 127	70	1	48	1 008	362	116	80	0	0	0	2
e g b primera etapa	1 463	1 187	54	0	49	1 064	279	91	83	2	1	0	2
e g b segona etapa	478	381	17	0	18	346	86	20	18	0	0	0	2
f p primer grau	371	330	10	0	1	319	40	7	13	0	0	0	1
f p segon grau	127	111	4	0	6	101	13	5	4	0	0	0	0
b u p / c o u	263	212	17	0	5	190	43	8	6	0	0	1	2
titol mitjà	144	113	6	0	11	96	27	7	5	1	0	0	2
titol superior	115	78	11	0	6	81	29	1	8	3	0	0	1
no hi consta	11	4	3	0	0	1	7	3	0	0	0	0	0

Taula B **Població de dret per lloc de naixement i nivell d'instrucció -dades censurades-. (2/2)**
Municipis
Selecció

Àmbit	nascuts Cast.-M	nascuts Cast.-Ll	nascuts P. Valen.	nascuts Extrem.	nascuts Galícia	nascuts Madrid	nascuts Múrcia	nascuts Navarra	nascuts P. Basc	nascuts Rioja	nascuts Ceu/Mel	nascuts estran	no hi consta
Fitx													
total	64	99	62	65	17	21	96	7	25	0	1	58	4
dificult llegir/escriure	9	10	9	13	6	6	17	4	0	0	0	0	0
primària incompleta	20	29	21	22	6	3	47	2	5	0	0	11	0
e g b primera etapa	19	25	14	17	1	3	19	0	2	0	0	17	0
e g b segona etapa	8	18	4	8	0	2	4	0	3	0	0	8	2
f p primer grau	3	4	3	0	0	0	7	0	2	0	0	1	0
f p segon grau	1	0	3	0	0	0	0	0	0	0	0	3	0
b u p / c o u	4	5	6	1	0	4	0	0	5	0	1	7	1
titol mitjà	0	1	1	3	3	3	1	0	0	0	0	4	0
titol superior	0	7	1	1	1	0	1	1	4	0	0	7	1
no hi consta	0	0	0	0	0	0	0	0	4	0	0	0	0

Taula C.1 Informatius de les proves de censura de dades agregades

Úniques (f/c): percentatge per redistribuir: 10 Valor mínim per redistribuir: 2 També freq per pertorbar N
Pertorbacions: valor màxim per pertorbar 5 Amplitud de la pertorbació: 3

Freqüències de les diferències entre les dades originals i les diferents fases

			Úniques	Censurades	Elevades	Arrodonides
Diferència:	-6	Freqüències:	1 (0.0 %)	1 (0.0 %)	10 (0.0 %)	0 (0.0 %)
Diferència:	-5	Freqüències:	2 (0.0 %)	2 (0.0 %)	20 (0.0 %)	11 (0.0 %)
Diferència:	-4	Freqüències:	4 (0.0 %)	499 (0.2 %)	557 (0.3 %)	349 (0.2 %)
Diferència:	-3	Freqüències:	23 (0.0 %)	1039 (0.5 %)	1389 (0.6 %)	1086 (0.5 %)
Diferència:	-2	Freqüències:	329 (0.2 %)	2327 (1.1 %)	3834 (1.8 %)	3445 (1.6 %)
Diferència:	-1	Freqüències:	152 (0.1 %)	4076 (1.9 %)	9593 (4.4 %)	12729 (5.9 %)
Diferència:	0	Freqüències:	214887 (99.4 %)	189165 (87.5 %)	185393 (85.8 %)	181443 (83.9 %)
Diferència:	1	Freqüències:	700 (0.3 %)	6920 (3.2 %)	9075 (4.2 %)	11519 (5.3 %)
Diferència:	2	Freqüències:	95 (0.0 %)	6156 (2.8 %)	4458 (2.1 %)	4305 (2.0 %)
Diferència:	3	Freqüències:	7 (0.0 %)	6015 (2.8 %)	1717 (0.8 %)	1112 (0.5 %)
Diferència:	4	Freqüències:	0 (0.0 %)	0 (0.0 %)	118 (0.1 %)	152 (0.1 %)
Diferència:	5	Freqüències:	0 (0.0 %)	0 (0.0 %)	26 (0.0 %)	41 (0.0 %)
Diferència:	6	Freqüències:	0 (0.0 %)	0 (0.0 %)	4 (0.0 %)	6 (0.0 %)
Diferència:	7	Freqüències:	0 (0.0 %)	0 (0.0 %)	2 (0.0 %)	2 (0.0 %)
Diferència:	8	Freqüències:	0 (0.0 %)	0 (0.0 %)	2 (0.0 %)	0 (0.0 %)
Diferència:	9	Freqüències:	0 (0.0 %)	0 (0.0 %)	2 (0.0 %)	0 (0.0 %)
Fora valors		Freqüències:	0	0	0	0

Municipis tractats: 940 Cel.les tractades: 216200

Cel.les modificades: Valors únics (files) 482, Valors únics (columnes) 48, Pertorbades 35404

Taula C.2 Informatius de les proves de censura de dades agregades

Úniques (f/c): percentatge per redistribuir: 10 Valor mínim per redistribuir: 2 També freq. per pertorbar: N
Pertorbacions: valor màxim per pertorbar: 5 Amplitud de la pertorbació: 3

Mèrcies de les diferències entre les dades originals i les diferents fases

			Úniques		Censurades	
Diferència:	-6	Inèrcies:	62	(0.0 %)	62	(0.0 %)
Diferència:	-5	Inèrcies:	94	(0.0 %)	94	(0.0 %)
Diferència:	-4	Inèrcies:	170	(0.0 %)	3010	(0.1 %)
Diferència:	-3	Inèrcies:	680	(0.0 %)	5696	(0.1 %)
Diferència:	-2	Inèrcies:	3622	(0.1 %)	11811	(0.2 %)
Diferència:	-1	Inèrcies:	1631	(0.0 %)	13372	(0.2 %)
Diferència:	0	Inèrcies:	5972379	(99.9 %)	5918396	(99.0 %)
Diferència:	1	Inèrcies:	0	(0.0 %)	13475	(0.2 %)
Diferència:	2	Inèrcies:	0	(0.0 %)	12722	(0.2 %)
Diferència:	3	Inèrcies:	0	(0.0 %)	0	(0.0 %)
Diferència:	4	Inèrcies:	0	(0.0 %)	0	(0.0 %)
Diferència:	5	Inèrcies:	0	(0.0 %)	0	(0.0 %)
Diferència:	6	Inèrcies:	0	(0.0 %)	0	(0.0 %)
Diferència:	7	Inèrcies:	0	(0.0 %)	0	(0.0 %)
Diferència:	8	Inèrcies:	0	(0.0 %)	0	(0.0 %)
Diferència:	9	Inèrcies:	0	(0.0 %)	0	(0.0 %)
Diferència:	10	Inèrcies:	0	(0.0 %)	0	(0.0 %)
Diferència:	11	Inèrcies:	0	(0.0 %)	0	(0.0 %)

Taula C.3 Informatius de les proves de censura de dades agregades

Úniques (f/c): percentatge per redistribuir: 10 Valor mínim per redistribuir: 2 També freq. per perforbar: N
Perforbacions: valor màxim per perforbar: 5 Amplitud de la perforbació: 3

Matriu de diferències absolutes per al total dels municipis (redistribució dades úniques -fil/col-)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)
(1)	-20	-7	-13	0	1	4	0	0	0	0	2	0	4	0	0	0	1	0	0	0	0	16	19
(2)	-10	6	-2	1	-1	-5	2	0	0	0	-3	3	-1	-1	-1	0	0	0	0	0	0	0	8
(3)	3	-1	-15	1	-6	0	-2	0	0	0	-2	-8	-1	-1	-1	0	-4	0	0	0	0	0	0
(4)	-17	-8	-22	5	0	7	0	0	0	0	1	2	3	0	0	0	2	0	0	0	0	9	25
(5)	-40	6	-27	-2	0	6	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	23	32
(6)	-33	14	-39	7	0	9	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	9	30
(7)	-34	11	-25	-4	0	7	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	12	27
(8)	-15	2	-18	6	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	16
(9)	-13	9	-12	1	0	3	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	2	9
(10)	-29	16	-9	-14	6	9	0	0	0	0	2	3	6	2	2	0	1	0	0	0	0	4	28

Suma de les diferències de totes les cel.les: 0

Matriu de diferències percentuals per al total dels municipis (redistribució dades úniques -fil/col-)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)
(1)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	3.8
(2)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.8
(3)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
(4)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	4.1
(5)	0.0	0.1	-0.2	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	25.0
(6)	0.0	0.3	-0.6	0.1	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	25.9
(7)	0.0	0.1	-0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	8.3
(8)	0.0	0.0	-0.1	0.1	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	11.5
(9)	0.0	0.1	-0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	6.5
(10)	-0.3	2.6	-0.2	-0.6	0.6	4.4	0.0	0.0	0.0	0.0	1.6	2.1	5.8	0.9	2.2	0.0	1.1	0.0	0.0	0.0	0.0	0.6	9.2

Taula C.4 Informatius de les proves de censura de dades agregades

Úniques (f/c) percentatge per redistribuir: 10 Valor mínim per redistribuir 2 També freq per pertorbar: N
Pertorbacions: valor màxim per pertorbar: 5 Amplitud de la pertorbació: 3

Matriu de diferències absolutes per al total dels municipis (+ pertorbacions aleatòries)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)
(1)	-18	20	29	22	-4	21	18	26	10	21	44	64	4	41	66	33	48	49	15	31	29	55	51
(2)	9	40	19	12	15	12	47	42	16	47	55	17	67	-37	22	41	37	63	29	29	48	13	54
(3)	-1	22	-8	-23	-19	6	40	33	30	34	-19	69	69	53	32	72	29	53	58	44	45	3	51
(4)	-63	19	9	25	3	33	48	4	22	29	61	32	44	37	56	27	53	42	42	17	37	72	50
(5)	7	17	-23	33	72	63	42	22	12	25	35	-11	19	41	37	44	21	13	21	22	17	80	50
(6)	7	6	-36	22	42	41	7	16	10	17	24	44	30	11	15	34	27	11	12	10	19	60	37
(7)	-43	41	5	-5	50	92	3	55	33	24	82	70	54	36	39	59	39	38	52	15	15	99	59
(8)	67	35	-6	63	37	-8	27	37	32	35	33	47	45	55	53	45	10	37	50	27	9	62	20
(9)	48	44	36	46	31	69	24	40	25	30	66	46	47	50	73	40	44	51	28	22	34	95	37
(10)	-1	49	-38	17	58	17	8	5	6	5	27	37	11	22	6	20	11	10	7	-2	0	31	50

Suma de les diferències de totes les cel.les: 6975

Matriu de diferències percentuals per al total dels municipis (+ pertorbacions aleatòries)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)
(1)	0.0	0.1	0.1	0.1	0.0	0.2	2.8	4.5	3.9	4.9	0.3	1.1	0.0	0.2	1.3	1.7	0.4	8.5	2.4	7.6	3.6	1.1	10.2
(2)	0.0	0.0	0.0	0.0	0.0	0.0	1.2	1.9	1.5	1.6	0.1	0.0	0.2	-0.1	0.1	0.5	0.1	1.7	0.8	1.0	1.2	0.1	5.2
(3)	0.0	0.0	0.0	0.0	0.0	0.0	0.9	1.4	2.4	1.2	0.0	0.1	0.3	0.1	0.1	0.7	0.1	1.5	1.4	1.7	0.9	0.0	4.8
(4)	0.0	0.0	0.0	0.0	0.0	0.2	2.0	0.3	2.6	2.1	0.4	0.1	0.5	0.2	0.5	0.4	0.8	2.7	1.6	1.5	1.4	0.4	8.2
(5)	0.0	0.2	-0.2	0.3	0.4	1.6	8.0	8.8	7.2	8.9	1.0	-0.2	1.0	1.1	1.7	3.3	1.4	3.3	3.0	9.2	3.8	1.7	39.1
(6)	0.0	0.1	-0.6	0.3	0.4	1.4	2.0	9.1	9.5	6.3	1.2	1.3	2.4	0.6	1.2	3.4	3.0	4.0	2.6	5.3	6.8	1.7	31.9
(7)	0.0	0.2	0.0	0.0	0.3	1.3	0.3	6.6	6.5	3.3	2.1	0.8	1.5	1.0	1.0	1.3	2.2	4.5	2.8	2.5	1.7	0.6	18.2
(8)	0.1	0.3	0.0	0.6	0.4	-0.1	3.3	6.3	11.8	5.9	1.1	0.6	1.4	3.0	2.2	2.1	0.8	3.5	3.9	4.0	1.8	1.0	14.4
(9)	0.1	0.5	0.3	0.5	0.5	1.1	3.1	3.4	5.8	5.3	3.6	0.7	1.4	4.2	3.3	1.4	4.5	5.4	1.8	3.4	8.3	0.9	26.6
(10)	0.0	8.0	-1.0	0.8	5.9	8.4	32.0	27.8	37.5	31.3	22.0	25.9	10.6	9.6	6.5	21.5	12.5	58.8	15.9	-33.3	0.0	4.9	16.3

Taula C.5 Informatius de les proves de censura de dades agregades

Uniques (f/c): percentatge per redistribuir: 10 Valor mínim per redistribuir: 2 També freq. per pertorbar: N
Pertorbacions: valor màxim per pertorbar: 5 Amplitud de la pertorbació: 3

Matriu de diferències absolutes per al total dels municipis (dades elevades)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)
(1)	-1	-11	30	11	-10	8	7	29	6	14	20	44	-9	24	53	9	14	23	-1	19	15	29	2
(2)	35	29	110	28	-85	-66	2	-24	-3	-11	-98	-53	-16	-128	-64	-12	-12	7	-18	-8	14	-68	2
(3)	52	5	68	-54	-135	-95	-32	-46	-22	-37	-157	-82	-56	-45	-104	-24	-97	-42	-15	-18	-58	-118	-26
(4)	-35	-15	-1	7	-17	-22	9	-43	-10	-2	24	-19	-12	-13	5	-40	10	-11	-11	-12	-3	11	6
(5)	-12	-12	-39	7	69	50	33	5	2	13	30	-20	9	32	22	29	12	1	8	13	8	36	15
(6)	-7	-17	-56	8	38	34	-2	13	4	8	18	34	17	3	6	28	19	-5	-3	7	10	21	7
(7)	-78	-9	-32	-40	39	68	-21	22	3	-7	59	26	27	17	10	11	20	-1	11	-12	-1	30	11
(8)	44	-2	-40	20	27	-23	-4	20	16	11	30	13	19	53	24	14	-4	6	25	10	0	18	-1
(9)	22	3	8	7	19	34	6	24	1	7	51	23	12	37	45	-19	32	14	1	3	15	36	10
(10)	-21	29	-48	6	55	12	2	0	3	4	23	34	9	20	3	4	6	8	3	-2	0	5	-26

Suma de les diferències de totes les cel.les: 0

Matriu de diferències percentuals per al total dels municipis (dades elevades)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)
(1)	0.0	0.0	0.1	0.0	0.0	0.1	1.1	5.0	2.4	3.3	0.1	0.7	-0.1	0.1	1.0	0.5	0.1	4.0	-0.2	4.7	1.9	0.6	0.4
(2)	0.0	0.0	0.1	0.0	0.0	-0.1	0.0	-1.1	-0.3	-0.4	-0.2	-0.1	-0.1	-0.2	-0.2	-0.1	0.0	0.2	-0.5	-0.3	0.3	-0.4	0.2
(3)	0.0	0.0	0.1	0.0	-0.1	-0.2	-0.7	-1.9	-1.8	-1.3	-0.4	-0.1	-0.2	-0.1	-0.4	-0.2	-0.4	-1.2	-0.4	-0.7	-1.1	-0.6	-2.4
(4)	0.0	0.0	0.0	0.0	0.0	0.0	-0.1	0.4	-3.2	-1.2	-0.1	0.2	-0.1	-0.1	-0.1	0.0	-0.6	0.2	-0.7	-0.4	-1.1	-0.1	0.1	1.0
(5)	0.0	-0.1	-0.3	0.1	0.4	1.3	6.3	2.0	1.2	4.6	0.9	-0.4	0.5	0.8	1.0	2.1	0.8	0.3	1.2	5.4	1.8	0.8	11.7
(6)	0.0	-0.3	-0.9	0.1	0.4	1.2	-0.6	7.4	3.8	3.0	0.9	1.0	1.3	0.2	0.5	2.8	2.1	-1.8	-0.6	3.7	3.6	0.6	6.0
(7)	0.0	0.0	-0.2	-0.2	0.2	1.0	-1.8	2.6	0.6	-1.0	1.5	0.3	0.7	0.5	0.3	0.2	1.1	-0.1	0.6	-2.0	-0.1	0.2	3.4
(8)	0.0	0.0	-0.3	0.2	0.3	-0.3	-0.5	3.4	5.9	1.9	1.0	0.2	0.6	2.9	1.0	0.7	-0.3	0.6	1.9	1.5	0.0	0.3	-0.7
(9)	0.0	0.0	0.1	0.1	0.3	0.6	0.8	2.0	0.2	1.2	2.8	0.4	0.4	3.1	2.0	-0.6	3.3	1.5	0.1	0.5	3.6	0.4	7.2
(10)	-0.2	4.7	-1.2	0.3	5.5	5.9	8.0	0.0	18.8	25.0	18.7	23.8	8.7	8.7	3.3	4.3	6.8	47.1	6.8	-33.3	0.0	0.8	-8.5

Taula C.6 Informatius de les proves de censura de dades agregades

Úniques (f/c) percentatge per redistribuir: 10 Valor mínim per redistribuir: 2 També freq. per pertorbar: N
Pertorbacions: valor màxim per pertorbar: 5 Amplitud de la pertorbació: 3

Matriu de diferències absolutes per al total dels municipis (elevades i arrodonides)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)
(1)	-77	-32	-12	-42	-70	-8	9	29	7	15	10	36	-21	20	48	11	1	23	-1	19	15	19	1
(2)	117	61	33	84	-2	-28	4	-22	-3	-7	-67	-22	23	-97	-44	-4	9	5	-15	-8	15	-35	3
(3)	242	116	161	40	63	-3	-27	-45	-22	-34	-94	-23	-26	2	-82	-5	-84	-35	-10	-16	-57	-44	-17
(4)	-2	-3	30	27	-8	-13	7	-42	-8	-2	22	2	3	-13	16	-38	11	-9	-12	-12	-1	44	1
(5)	-55	-49	-31	-22	-3	30	36	4	1	13	11	-29	-2	9	9	16	11	0	6	12	7	13	13
(6)	-33	-35	-49	1	2	18	-3	11	4	7	4	17	11	4	0	19	13	-7	-3	6	10	-4	7
(7)	-72	-2	-35	-60	6	41	-22	21	1	-8	42	11	21	2	5	14	11	-1	11	-11	-3	20	8
(8)	8	-32	-72	0	10	-43	-9	20	16	8	19	-3	1	34	13	7	-5	7	23	9	-1	-7	-3
(9)	-47	-43	-29	-13	-16	1	4	24	1	6	34	-16	-12	25	36	-20	30	10	-3	3	15	0	10
(10)	-81	19	4	-15	18	5	1	0	3	2	19	27	2	14	-1	0	3	7	4	-2	0	-6	-23

Suma de les diferències de totes les cel.les: 0

Matriu de diferències percentuals per al total dels municipis (elevades i arrodonides)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)
(1)	0.0	-0.1	0.0	-0.1	-0.1	-0.1	1.4	5.0	2.8	3.5	0.1	0.6	-0.3	0.1	0.9	0.6	0.0	4.0	-0.2	4.7	1.9	0.4	0.2
(2)	0.0	0.1	0.0	0.1	0.0	0.0	0.1	-1.0	-0.3	-0.2	-0.1	0.0	0.1	-0.1	-0.1	0.0	0.0	0.1	-0.4	-0.3	0.4	-0.2	0.3
(3)	0.0	0.1	0.1	0.0	0.0	0.0	-0.6	-1.9	-1.8	-1.2	-0.2	0.0	-0.1	0.0	-0.3	-0.1	-0.4	-1.0	-0.2	-0.6	-1.1	-0.2	-1.6
(4)	0.0	0.0	0.1	0.1	0.0	-0.1	0.3	-3.1	-0.9	-0.1	0.1	0.0	0.0	-0.1	0.2	-0.6	0.2	-0.6	-0.5	-1.1	0.0	0.3	0.2
(5)	0.0	-0.4	-0.2	-0.2	0.0	0.8	6.9	1.6	0.6	4.6	0.3	-0.6	-0.1	0.2	0.4	1.2	0.7	0.0	0.9	5.0	1.5	0.3	10.2
(6)	0.0	-0.7	-0.8	0.0	0.0	0.6	-0.8	6.3	3.8	2.6	0.2	0.5	0.9	0.2	0.0	1.9	1.4	-2.6	-0.6	3.2	3.6	-0.1	6.0
(7)	0.0	0.0	-0.2	-0.3	0.0	0.6	-1.9	2.5	0.2	-1.1	1.1	0.1	0.6	0.1	0.1	0.3	0.6	-0.1	0.6	-1.8	-0.3	0.1	2.5
(8)	0.0	-0.3	-0.6	0.0	0.1	-0.6	-1.1	3.4	5.9	1.3	0.6	0.0	0.0	1.8	0.5	0.3	-0.4	0.7	1.8	1.3	-0.2	-0.1	-2.2
(9)	-0.1	-0.5	-0.3	-0.1	-0.3	0.0	0.5	2.0	0.2	1.1	1.9	-0.3	-0.4	2.1	1.6	-0.7	3.1	1.1	-0.2	0.5	3.6	0.0	7.2
(10)	-0.8	3.1	0.1	-0.7	1.8	2.5	4.0	0.0	18.8	12.5	15.4	18.9	1.9	6.1	-1.1	0.0	3.4	41.2	9.1	-33.3	0.0	-1.0	-7.5

REFERÈNCIES

[1] **Appel, G.** i **Hoffmann, D.J.** (1977). "Perturbation by Compensation". *Proceedings of the International Seminar on Statistical Confidentiality.* Dublin.

[2] **Blien, U.** i **Wirth H.** (1992). "Empirical tests of the Anonymity of Data from Official Statistics". *Proceedings of the International Seminar on Statistical Confidentiality.* Dublin.

[3] **Börjesson, M.** (1987). "Swedish Data Protection in Practice". *Seminar on Openness and Protection of Privacy in the Information Society. Proceedings*, 73–77. Netherlands Central Bureau of Statistics.

[4] **Citteur, C.A.W.** i **Willenborg, L.C.R.J.** (1991). *Public Use Files: Current Practices at National Statistical Bureaus.* Netherlands Central Bureau of Statistics.

[5] **Cox, L.H.** (1982). "Suppression Methodology and Statistical Disclosure Control". *INFOR: Canadian Journal of Operational Research and Information Processingt.*, volum **20**, nº **4**, 423–432. (Novembre).

[6] **Cox, L.H.**, i **Ernst, L.R.** (1980). "Controlled Rounding". *Journal of the American Statistical Association.*, volum **75**, nº **370**, 377–385.

[7] **Cox, L.H.** (1984). *Methods for controlling statistical disclosure in aggregate magnitude data.* Comunicació personal.

[8] **Cox, L.H., McDonald, S-K.** i **Nelson, D.** (1986). "Confidentiality Issues at the United States Bureau of the Census". *Journal of Official Statistics,* volum **2**, nº **2**, 135–160.

[9] **Cox, L.H., Fagan, J.T., Greenberg, B.** i **Hemmig, R.** (1987). *Research at the Census Bureau into Disclosure Avoidance Technique for Tabular Data.* Informe intern. U.S. Bureau of the Census SRD/RR-87/06. (Febrer).

[10] **Cox, L.H.** (1988). "Statistical Confidentiality". *Encyclopedia of Statistical Sciences,* Volum **8**, 641–642.

[11] **Cox, L.** (1992). "Solving Confidentiality Protection Problems in Tabulations Using Network Optimization: A Model for Cell Supression in U.S. Economic Censuses". *Proceedings of the International Seminar on Statistical Confidentiality.* Dublin.

[12] **Dalenius, T.** (1983). "Informed Consent or R.S.V.P.". *Incompleted Data in Sample Surveys. Proceedings of the Symposium Panel on Incomplete Data,* Volum **3**, sessió **II**. Academic Press. Nova York.

[13] **Dalenius, T.** (1986). "Finding a Needle In a Haystack". *Journal of Official Statistics,* volum **2**, nº **3**, 329–336.

[14] **Duncan, G.T.** i **Lambert, D.** (1986). "Disclosure-Limited Data Dissemination". *Journal of the American Statistical Association,* volum **81**, nº **393**, 10–28.

[15] **Greenberg, B.** (1986). *Designing a Disclosure Avoidance Methodology for the 1990 Decennial Censuses.* 1990 Census Data Products Fall. U.S. Bureau of the Census. Conference. Arlington, Virgínia.

[16] **Greenberg, B.** (1988). *Disclosure Avoidance Research for Economic Data.* Joint Advisory Committee Meeting. Comunicació privada. U.S. Bureau of the Census. (Octubre)

[17] **Greenberg, B.** (1990). *Disclosure Avoidance Research at the Census Bureau.* 1990 Annual Research Conference. U.S. Bureau of the Census. Session on Disclosure Avoidance. Arlington, Virgínia. (Març).

[18] **Griffin, R.A., Navarro, A.** i **Flores-Baez, L.** (1989). "Disclosure Avoidance for the 1990 Census". *Proceedings of the Section on Survey Research Methods.* ASA. U.S. Bureau of the Census. Washington.

[19] **Gusfield, D.** (1990). "A little knowledge goes a long way: faster detection of compromised data in 2-D tables". *Proceedings of the 1990 IEEE Computer Society Symposium on Research in Security and Privacy*, 86–94. IEEE.

[20] **Heer, G. A.** (1992). "Bootstrap Procedure to Preserve Statistical Confidentiality in Contingency Tables". *Proceedings of the International Seminar on Statistical Confidentiality.* Dublin.

[21] **Kelly, J.P., Golden, B.L.** i **Assad, A.A.** (1992). *Cell Suppression: Disclosure Protection for Sensitive Tabular Data. Networks,* Volum **22**, 397–417.

[22] **Lougee-Heimer, R.** (1989). *Guarantying Confidentiality the protection of tabular data.* Comunicación personal. Department of Mathematical Sciencies of Clemson University. (Abril).

[23] **Repsilber, D.** (1992). "Safeguarding Secrecy in Aggregated Data". *Proceedings of the International Seminar on Statistical Confidentiality.* Dublin.

[24] **Rowe, E.** (1991). *Some Considerations in the Use of Linear Networks to Suppress Tabular Data.* U.S. Bureau of the Census.

[25] **Sullivan, C.M.** i **Zayatz, L.** (1991). "A Network Flow Disclosure Avoidance System Applied to the Census of Agriculture". *Proceedings of the Section on Survey Research Methods,* 357–362. ASA.

[26] **Wolf, M.K.** (1990). *Microaggregation and Disclosure Avoidance for Economic Establishment Data.* U.S. Bureau of the Census.

[27] **Zayatz, L.V.** (1992,a). "Linear Programming Methodology Used for Disclosure Avoidance Purposes at the Census Bureau". *Proceedings of the Section on Survey Research Methods.* ASA (en premsa).

[28] **Zayatz, L.V.** (1992,b). "Using Linear Programming Methodology for Disclosure Avoidance Purposes". *Proceedings of the International Seminar on Statistical Confidentiality.* Dublin.

ENGLISH SUMMARY:

RANDOM PERTURBATIONS BY COMPENSATION: A METHOD TO PROTECT CONFIDENTIAL STATISTICAL INFORMATION

J. Turmo i Soldevila

INTRODUCTION

The aim of this paper is to analyse which are the most effective methods to guarantee statistical confidenciality of aggregated data or macrodata and to preserve accurately the original data structure.

We introduce a method that applies random perturbations by compensation and also a methodology with its computering system that reduces the global effect of the original technique. This method is clear for users and it obtains results to be reproduced.

1. OBJECTIVES

The adequate method must satisfy these five requests: minimization of indirect disclosure risk, maximum approach to the original levels of information, guarantee of data quality, all tables of the tabulation plan should be homogenized and operating methods must be preserved.

2. ACTUAL METHODS

Actual methods can be divided into two conceptual categories: the ones that eliminate confidential information and the ones that mask it.

The methods that delete information are based on the elimination of risky cells —cells with an unacceptable risk of disclosure of confidential information—.

This group of eliminated data is called "primary suppressions". In the case of bivariate matrices some other cell of the same row and column of the cells deleted in primary supressions should be eliminated. If not the original data censured could be deduced from the difference with the marginal total. The eliminated cells are called "secundary supressions".

The masked methods change the original value of risky —cells— data in order to avoid disclosure of original data of the tables to be disseminated. There are two methods mainly used: the round-off methods and the one of random perturbations that adds to risky cells a random number calculated within a determined interval (symmetrical or asymmetrical, constant or variable) around zero.

3. DETERMINATION OF RISKY CELLS

There are 3 non-exclusive criteria mainly used to determine risky cells:

a) The frequency must be less than some threshold value.

b) The frequency of an element must be a percentage over the total of one of the matrix dimensions above a fixed limit.

c) The whole row or column must be assigned to a cell.

4. DESCRIPTION OF THE METHOD PROPOSED

The mixted method proposed applies random perturbations to cells with frequency less than or equal to a given value.

The application has four stages:

a) Risky cells are selected: frequencies less than N and unique populations.

b) A part of the frequency of unique populations is re-allocated according to the model for each one of the variables.

c) Random perturbations are applied to cells with frequency less than or equal to N.

d) The multidimensional compensating methods must be applied until marginal totals that were in the original tables are obtained.

Two types of compensating methods are used: the first method is only applied to a matrix dimension (the one with more elements) and it is based on the weighting method of the remainder resulting from the expansion of vectorial totals that were originally censured. In the case of other dimensions it is applied the compensation of pairs of reciprocal values. In each compensation $2^{**} d$ cells are affected (d is the dimension of data matrix).

5. EVALUATION OF THE EXPERIMENTAL RESULTS

The method and its application has been evaluated with a detailed test using a frequency cross-table. We use data obtained from a municipality of 5000 inhabitants and data obtained from the censoring method applied to all municipalities of Catalonia.

In this process, unique populations are used (by re-allocating 10%, with a minimum of 2) and elements of frequency 5 or less are perturbed with a perturbation of amplitude 3.

By evaluating results of this method, the disclosure risk of this method is irrelevant and it keeps a high level of information. Also users are allowed to work with methods they are used to.