QÜESTIIÓ, Vol 20, 3 pp. 523-546. 1996

CONTROL DEL RISC DE REVELACIÓ ESTADÍSTICA EN LA DIFUSIÓ DE MICRODADES DE POBLACIÓ, APLICAT AL CAS D'UNA MOSTRA D'INDIVIDUS PROCEDENT DEL CENS DE LA POBLACIÓ DE CATALUNYA DE 1991

ALFONS GARÍN i RAMÍREZ*

Les oficines d'estadística oficial responen al compromís de difondre la informació disponible; aquesta informació s'ha obtingut sota condicions de confidencialitat; a través de l'elaboració d'una mostra de registres individuals anònims (MRA), l'Institut d'Estadística de Catalunya compleix el doble compromís: difondre informació i preservar el secret estadístic. Aquest treball descriu el procés d'anàlisi i control de la revelació estadística aplicat a la mostra de microdades.

En síntesi, el plantejament és el següent: un element de la població que posseeix una combinació única de característiques i és present a la mostra de microdades representa un risc de revelació. Un intrús podria comparar aquesta combinació única de variables categòriques amb la mateixa combinació present en un altre arxiu de dades contenint identificadors formals de l'individu. D'aquesta manera podria lligar ambdós fitxers i accedir a la informació sensible corresponent a l'individu.

L'objectiu és reduir el risc de revelació estadística a un valor que faci pràcticament impossible identificar una persona i, al mateix temps, mantenir el valor informatiu de les dades mitjançant algun model que permeti calcular el risc i aplicar procediments de protecció de les dades.

Control of statistical disclosure risk in dissemination of microdata from population, applied on a sample of individual records from Catalonia's population census 1991.

Keywords: Protecció de Microdades de Població, Mètodes de Control del Risc de Revelació Estadística, Mostra de Registres Anònims d'Individus.

*El treball descrit en aquest article es desenvolupà a la Subdirecció d'Assistència Tècnica Estadística de l'Institut d'Estadística de Catalunya; actualment, Alfons Garín pertany al Gabinet de Planificació i Avaluació de la Universitat Politècnica de Catalunya.

- Article rebut l'abril de 1996.
- Acceptat el setembre de 1996.

1. INTRODUCCIÓ

La difusió de microdades procedents del cens de població està justificada des dels objectius bàsics dels instituts d'estadística oficial, en particular els de posar a disposició dels estudiosos, institucions d'ensenyament universitari, etc., informació de base útil per al millor coneixement de la realitat social i demogràfica del territori, amb el conseqüent desenvolupament d'instruments i mètodes d'anàlisi de les dades. No obstant això, les oficines responsables de la difusió de dades estadístiques s'enfronten al risc de revelació de la confidencialitat, que acompanya sempre el fet de publicar informació de la qual es puguin deduir dades corresponents a individus concrets.

1.1. El concepte de revelació estadística

Tore Dalenius [Dalenius (1977)] establí una definició de revelació estadística que forma part del marc metodològic de la majoria d'operacions d'anàlisi i control de la revelació. En el seu treball dedicat a la recerca d'una metodologia per al control de la revelació, Dalenius arribà a formalitzar aquesta idea: *es produeix revelació estadística quan la difusió de dades estadístiques permet determinar el valor de microdades amb major precisió que en absència de la publicitat de les dades.* Tant en el cas de la difusió de macrodades (taules de dades agregades) com en el cas de microdades (dades individuals de persones, famílies, empreses, etc.) l'escenari que representa aquest concepte és aquell en el qual un usuari de l'estadística identifica una informació determinada com a corresponent a un individu concret (persona, empresa, etc.).

Cal destacar la implicació entre revelació i identificació: no hi ha revelació estadística sense identificació. En general, els models que apliquen mètodes d'anàlisi i control del risc de revelació formulen una aproximació al càlcul de la probabilitat d'identificació; l'esquema de la fase de control de la revelació en una operació de difusió de dades consisteix essencialment en:

i) càlcul de la probabilitat d'identificació p(id.);

ii) obtenció de valors superiors als acceptables;

iii) aplicació de mètodes per minimitzar el valor de p(id.), si es compleix *ii)*.

1.2. Comissions de protecció de dades

Si aquest és l'esquema general (1.1.*i)* – 1.1.*iii)*), sota el qual actua un procés de control del risc, en la pràctica el nivell d'agregació de les dades (micro o macro) han imposat tractaments específics i, en definitiva, metodologies diferents. El cas de la difusió de microdades ha mobilitzat, en diferents països i a diferents nivells nacionals i supra-nacionals [Eurostat (1994)], la creació de comissions de protecció de dades,

acords de normatives i criteris bàsics de difusió, grups de treball [Willenborg (1996a)], seminaris internacionals [Bled (1996)], etc., posant de relleu:

i) una sensibilització respecte al valor informatiu per a la societat de les dades que les institucions d'estadística han elaborat a partir de la informació que els propis individus de la societat subministren;

ii) a la vegada, una sensibilització social del compromís de confidencialitat sota el qual s'ha obtingut la informació;

iii) un reconeixement de les condicions específiques en les quals es manifesta el risc de revelació estadística en el cas de la difusió de microdades[1].

Un exemple d'aquest tipus d'iniciativa el constitueix la *Ponencia de Protección de Datos (PPD)* que, en el cas de difusió de microdades procedents del cens de població, estableix les següents recomanacions bàsiques:

a) Per a fitxers de microdades, únicament es difondran mostres.

b) Els registres individualitzats de persones, llars o habitatges, seran completament anonimitzats. És important fer explícita la recomanació, però no podria ser d'altra manera tenint en compte que en la gravació de dades del cens en suport magnètic s'eliminen els identificadors formals dels individus - DNI, noms i cognoms, adreces, etc.

c) La fracció de mostreig (FM), el nivell de desagregació de les variables geogràfiques (DG) (per exemple, per a la variable *lloc de residència*: municipi) i el nivell de desagregació conceptual (DC) de les variables que no són de localització (per exemple, per a la variable *edat*: estrats de 5 anys), són característiques de la mostra que actuen de forma conjunta respecte al risc de revelació estadística. Aquesta proposta operativa inclou, a nivell paradigmàtic, el concepte *Dalenius* de revelació; en efecte, la identificació és condició necessària de revelació, però la identificació serà possible si l'individu objecte d'identificació és un cas de població única; és acceptable que la característica *d'unicitat* estigui relacionada amb els nivells de desagregació de les dades que, en definitiva, determinen la variació del risc de revelació. Un document de la Ponència presenta relacions entre nivells d'àrea geogràfica (DG) per a la

[1]Respecte a les condicions específiques que se senyalen a l'últim punt 1.2.*iii)*, cal destacar que són de doble naturalesa: qüestions tècniques, de les quals ens ocupem en aquest paper, i una qüestió política. L'especifitat a la qual es refereix aquest aspecte polític està relacionat amb la percepció especial amb la qual pot rebre l'opinió pública, a qualsevol nivell, la notícia de difusió de microdades per part de les institucions oficials d'estadística. Aquest risc pressumpte de revelació pot comportar un cost polític. La línia d'actuació més eficaç per minimitzar aquest risc, consisteix en informar obertament sobre els mètodes de control aplicats a reduir el risc *real* de revelació, en determinar els usuaris destinataris de la informació i en definir les condicions contractuals sota les quals es lliuren les dades. Aquest treball s'ocupa del control del risc real de revelació.

variable *lloc de residència* (per exemple: Comunitats Autònomes, Províncies, Municipis > 100.000 habitants, etc.), fraccions de mostreig (FM) acceptables (mai superiors al 5%), afegint recomanacions generals per als nivells DC en funció de DG i FM. Els resultats de l'operació que es descriu en aquest article ofereixen un contrast empíric d'aquesta interrelació que la Ponència, encertadament, establí.

1.3. El projecte de l'Institut d'Estadística de Catalunya: MRA

L'Institut d'Estadística de Catalunya ha portat a terme el projecte de producció i difusió d'una mostra de registres anonimitzats (MRA) d'individus, procedent de la població de Catalunya, registrada al cens de 1991. La fase principal del projecte ha tingut com a objectiu el control de la revelació estadística de les dades finals. S'incorporen a la metodologia les dues propostes: el concepte *Dalenius* de revelació i les recomanacions bàsiques de la *Ponencia de Protección de Datos* per a difusió de microdades, considerades aquelles com una proposta inicial de criteris operatius:

a) es produirà una mostra de registres anònims,

b) s'utilitzaran les variacions de DG i DC com a instruments de control de la revelació i, per tant, no s'utilitzaran mètodes de distorsió o alteració de les dades.

En aquest article es presenta el desenvolupament de l'operació de control de la revelació, adoptant un model conceptual, utilitzant procediments de càlcul dels components del risc de revelació i oferint els resultats obtinguts i l'estructura final de la mostra. Els diferents apartats descriuen les diferents fases de l'operació:

2. Model conceptual per a l'estudi del risc de revelació estadística.

3. Mètode de la submostra per al càlcul de poblacions úniques; fases del procés.

4. Resultats i conclusions.

5. Estructura de l'arxiu final.

2. MODEL CONCEPTUAL PER A L'ESTUDI DEL RISC DE REVELACIÓ ESTADÍSTICA

Sabem que el compliment de les recomanacions *a)* i *b)* de PPD (difondre una mostra de la població i anonimitzar els registres) no cancel·la el risc de revelació. També tenim, com a hipòtesi general, la relació entre risc de revelació i probabilitat d'identificació individual. En conseqüència, el primer objectiu de l'operació és arribar

a un model que permeti quantificar la probabilitat d'identificació, d'alguna manera fiable, i proposi accions de control del risc. Aquest objectiu es complirà amb un model basat en l'anàlisi de freqüències a partir de les dades contingudes a la mostra, prèviament produïda; no serà, doncs, un model predictiu [Duncan i Lambert (1986)] de fonament baiesià, sinò que estarà inclòs en un marc identificatiu [Skinner (1994)] en el qual la se.lecció de les variables útils per a la identificació serà crucial en l'anàlisi final. Les fases del procés seran:

A) Producció d'una mostra de registres anònims d'individus del Cens 1991 de la població de Catalunya;

B) Anàlisi i càlcul de la p(identificació) a partir de les dades contingudes en la mostra (freqüències de determinades variables).

C) Modificacions dels paràmetres de la mostra (DC) i (DG) i repetició del pas **B)** fins obtenir una p(identificació) òptima.

El camí per arribar a una expressió formal del risc en termes de p(identificació) —fase **B)**—, s'inicia tractant de donar resposta a la pregunta: quines són les condicions que fan possible la identificació d'un individu concret, a partir d'informació anònima? Farem una hipòtesi sobre motivació, regles i mètodes d'identificació que defineixen un escenari d'identificació i reconeixerem els components del risc de revelació, obtenint el procediment de valoració de cada un dels components.

2.1. Motius de l'espionatge estadístic

Hi ha dues situacions que podrien motivar l'intent d'identificació:

a) les dades contingudes a l'arxiu de microdades tindrien una utilitat real, si s'estableix un vincle fiable amb individus concrets formalment identificats; es a dir, que l'usuari de les dades podria obtenir un benefici comercial, polític, personal, etc., si pogués afegir al coneixement previ i formal d'alguns individus (noms, cognoms, adreces, DNI, etc.) la informació inicialment anònima, no coneguda fins al moment de la publicació de les dades,

b) l'èxit en l'intent de revelació estadística pot comportar un desprestigi institucional per a l'autoritat de l'estadística oficial.

Si les dades són anònimes, en ambdós casos (2.1.*a)* i 2.1.*b)*) la primera condició perquè sigui possible la identificació és l'existència d'informació prèvia o al marge de la publicació de les microdades, a l'abast de l'usuari de l'estadística, en forma d'arxius o directoris de persones (clients bancaris, professionals, ciutadans al cens

electoral, etc.), bases de dades o qualsevol font de coneixement personal (cercles de confiança, per exemple).

2.2. Regles d'identificació

Sota la primera condició (existència d'informació prèvia), i sota el supòsit que es produeix l'intent de revelació, establim les regles que s'han de complir per acceptar que s'ha arribat a una identificació correcta.

a) L'usuari de les microdades ha de comptar amb alguna informació que identifiqui formalment algun o diversos individus (noms, adreces, etc.), conjuntament amb dades que determinen característiques personals; per exemple: edat, sexe i professió.

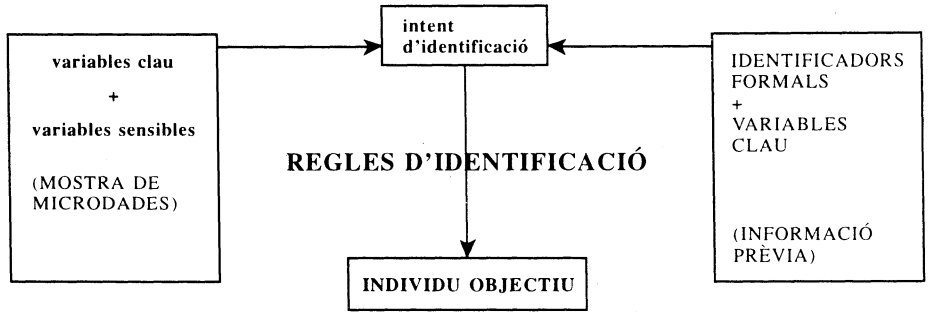
b) De l'arxiu estadístic de microdades (anònim), l'usuari triarà les variables que mantinguin coherència conceptual i de nivell de desagregació amb el seu arxiu d'identificadors formals, amb l'objectiu de realitzar una comparació de continguts entre les dues informacions.

c) D'aquest acarament d'arxius, únicament acceptarà com a èxit una correspondència exacta de tots i cada un dels valors de les variables que ha utilitzat en la comparació.

d) Haurà de comprovar que cap altre registre de l'arxiu de microdades compleix amb la correspondència exacta, per afirmar que es tracta de la identificació de l'individu objectiu. Si l'arxiu de microdades estadístiques és una mostra, ha de poder afirmar, amb confiança, que l'individu identificat no només és únic a la mostra, sinó que també és únic a la població.

2.3. Escenari d'identificació

De les regles 2.2.*a)*–2.2.*d)*, es desprèn la importància del rol de les variables clau d'identificació. Com és fàcil veure, la determinació de les variables útils per a l'acarament d'informació és crucial en l'èxit de la identificació; aquestes variables clau corresponen a les dades que figuren en els arxius utilitzats com a informació prèvia, acompanyant els identificadors formals, i presents també a la mostra de microdades. La resta d'informació, absent a la informació prèvia però present a MRA, es considera informació sensible i és l'objecte real de revelació. L'esquema 1 representa aquest escenari en el qual l'intrús selecciona i utilitza variables clau, identificatives d'un individu objectiu de revelació, i compara els seus valors amb els valors que conté l'arxiu de microdades:



Esquema 1

2.4. Components del risc de revelació

En el context català, marc poblacional d'MRA, s'han de considerar les bases de dades, arxius de registres personals, etc., externs a MRA, que poden actuar com a informació que faci possible la identificació d'individus. D'una banda l'existència d'aquests arxius és la primera condició d'identificació i, de l'altra, la seva estructura informa sobre les possibles variables clau disponibles i els nivells DG i DC d'aquestes variables. Un factor important, que afecta el valor del risc, és la distància temporal entre els períodes de referència de la mostra de microdades i les dades externes en mans de l'usuari. Respecte d'aquest factor, l'estat de maduració de les dades del cens dona un valor informatiu a nivell agregat molt diferent del valor individual de les dades. Totes aquestes consideracions sobre l'existència d'arxius externs a MRA, ens porta a traslladar a la fase d'execució del procés l'anàlisi de la composició dels arxius més o menys públics, en termes de variables clau d'identificació.

En el procés de càlcul del risc de revelació ens interessa estudiar els següents components:

a) Idèntica codificació i qualitat de les dades. En el procés d'anàlisi de la correspondència entre els valors de les variables clau d'un i altre conjunt de dades, adquireixen importància:

   i) els sistemes de codificació emprats en ambdós casos,

   ii) el resultat de la imputació de dades en la fase de validació i depuració del cens i

   (iii) taxes d'error i no-resposta en l'edició final dels arxius. Aquest component afecta negativament la precisió de la correspondència exacta de les variables clau.

b) Presència de l'individu objectiu d'identificació en MRA. El factor de mostreig (FM) és determinant en el valor d'aquest component.

c) L'individu objectiu és població única. Si els dos components anteriors *((a)*: idèntica codificació de les variables clau, i *(b)*: presència en la mostra de l'individu objectiu) es manifestessin positivament quant al risc de revelació, no són condicions suficients per a la identificació; manca un tercer component, nuclear en el procés d'identificació, que és la condició de població única per part de l'individu objectiu. Si l'individu no és un cas únic en la població, respecte als valors de variables clau analitzades, la identificació no és possible. En aquest component (unicitat de l'individu) són factors importants: el factor geogràfic (DG) i els nivells de desagregació de les variables clau en MRA.

e) El procés de comparació d'arxius ha de trobar un registre amb combinació única de variables clau i amb correspondència exacta amb la informació prèvia disponible. Si hi ha èxit, l'espia estadístic troba un cas únic a la mostra; l'últim component del risc de revelació és el nivell de confiança amb el qual es pot afirmar que un individu únic en la mostra ho és també a la població. Trobarem dificultats en el càlcul d'aquest component i triarem un estimador que sobrevalora el risc (que sempre és més segur), però sabem que hi ha factors de risc evident com són els àmbits professionals minoritaris per determinades categories territorials o poblacions rares o de coneixement públic (president de govern, per exemple). Serà necessari aplicar un criteri de seguretat a l'estudi general del risc, consistent en l'anàlisi de freqüències de cada una de les categories (DC) de les variables clau, creuades amb les categories de la variable LLOC DE RESIDÈNCIA, rebutjant aquelles que donen freqüències per sota d'un llindar de seguretat.

2.5. El risc de revelació estadística com a probabilitat d'identificació

Els quatre components del punt anterior (2.4.*a)* – 2.4.*d)*) els interpretem com a les condicions que es requereixen, simultàniament, per tal d'assolir una correcta identificació. El risc total d'identificar correctament un individu, comparant la informació continguda a MRA amb la informació prèvia de qualsevol origen, es calcula a partir del producte d'una successió de probabilitats condicionades:

$$(1) \qquad p(\text{identificació}|\text{intent}) = p(a) \cdot p(b|a) \cdot p(c|a,b) \cdot p(d|a,b,c)$$

p(intent): considerarem que l'intent es produeix i acceptarem: p(intent) = 1;

p(identificació|intent) = p(identificació);

$p(a)$: probabilitat d'homogeneïtat entre els sistemes que han produït la informació prèvia, que s'utilitza com a patró d'identificació, i la mostra de registres de població.

$p(b|a)$: probabilitat que l'individu objectiu estigui present a la mostra, donada la condició a.

$p(c|a,b)$: probabilitat que la combinació de valors de les variables clau de l'individu objectiu sigui única a la població, donades les condicions a i b.

$p(d|a,b,c)$: probabilitat de verificar que una combinació única de variables clau, present a la mostra i idèntica a la de l'individu objectiu, ho és també a la població, donades les condicions a,b i c.

Expressat en aquests termes el concepte *risc de revelació*, com a solució estadística al problema del risc real de violació de la confidencialitat que acompanya la difusió de microdades de població, es proposen estimadors dels quatre components de l'equació (1):

i) Considerem que les variables que produeixen els resultats a,b,c i d, són independents, i l'equació (1) queda:

$$p(\text{identificació}) = p(a) \cdot p(b) \cdot p(c) \cdot p(d) \tag{2}$$

ii) $p(a) = 1$; podríem recollir estadístiques d'errors en els processos d'enregistrament i codificació d'arxius d'origen administratiu per a inferir un valor acceptable de $p(a)$; sobre càlculs d'altres sistemes estadístics, i referents als arxius del cens, es recullen taxes d'errors que estimarien $p(a)$ de l'ordre de 0,20 [Marsh (1986)]; s'opta per una condició desfavorable (màxim risc) com a criteri de seguretat.

iii) $p(b)$ = FM, és a dir, la probabilitat d'estar present a la mostra és igual al factor de mostreig;

iv) per al càlcul de $p(c)$ i $p(d)$ s'ha d'optar per algun mètode d'inferència estadística, un cop descartada la possibilitat de comptar, a nivell poblacional, els casos d'unicitat respecte a diferents combinacions de DC i DG per a les variables clau, i els cops en què un cas de població única a la mostra ho és també a la població.

2.5.1. Mètodes de càlcul de la proporció de poblacions úniques

La proporció de poblacions úniques és l'estadístic que representa $p(c)$; necessitem un estimador de la proporció de poblacions úniques. A partir de la literatura metodològica sobre el tema s'han considerat tres mètodes que subministren estimadors del paràmetre buscat.

i) El mètode proposat per Bethlehem [Bethlehem *et alt.* (1990)] que utilitza un model Poisson-gamma de distribució de freqüències de combinacions de variables clau.

ii) L'anàlisi de distribució de les classes d'equivalència de la mostra de dades disponible, una vegada determinades les claus d'identificació [Voshell (1991)].

iii) Un mètode que dóna resultats similars al ii), millorant la precisió a partir d'un cert nombre de registres, basat en un procés de submostreig. L'anàlisi també necessita la definició prèvia de les variables clau d'identificació [Voshell (1991)].

Per calcular $p(c)$ optem pel mètode de submostreig. Aquest mètode es basa en la utilització d'una submostra obtinguda a partir de la mostra de registres individuals del cens. La idea és obtenir un estimador de la proporció de casos únics a la població, a partir de les dades empíriques que ens subministra l'anàlisi del comportament de les observacions de la submostra respecte a la mostra, tal com es desenvolupa a l'apartat 3.

2.5.2. *Probabilitat que un individu sigui un cas únic a la mostra i a la població, simultàniament*

Per explicar $p(d)$, considerem que, un cop l'usuari ha trobat a MRA un registre únic que concorda exactament amb l'individu objectiu, encara ha de verificar que aquest registre no pertanyi a un altre individu, és a dir, ha d'afirmar, amb confiança, que, a més de cas únic a la mostra és també únic a la població. Perquè això sigui possible, l'usuari ha de tenir informació complementària, sobretot referida a subpoblacions de dimensió petita (àrees geogràfiques petites, professions minoritàries, etc.), que permetin, pràcticament, una identificació espontània. A fi de controlar aquest component del risc de revelació, és bàsic determinar correctament el nivell DG que eviti localitzacions fines. En el nostre cas s'ha controlat que la distribució de cada una de les variables, en totes les seves categories desagregades, creuada amb la variable de localització de residència, no proporcioni en cap cas freqüències per sota d'un llindar de control.

A més, i malgrat la dificultat d'estimar $p(d)$, s'ha optat per un estimador que ofereix uns valors molt alts en comparació amb el que s'ha acceptat en altres treballs similars; es tracta d'utilitzar la proporció d'únics vertaders respecte a únics observats en la submostra com la probabilitat d'afirmar encertadament que un cas únic a la mostra és únic a la població.

3. MÈTODE DE LA SUBMOSTRA PER AL CÀLCUL DE POBLACIONS ÚNIQUES; FASES DEL PROCÉS

3.1. Fases principals del procés

I. Obtenció d'una mostra de registres de persones del Cens utilitzant un factor de mostreig f.

II. Amb el mateix factor f, s'obté una submostra a partir dels registres de la mostra.

Així doncs, tenim:

N : dimensió de la població

f : factor de mostreig

n_1 : dimensió de la mostra = $N \cdot f$

n_2 : dimensió de la submostra = $n_1 \cdot f$

III. Determinació de les variables clau.

IV. Determinació dels nivells de desagregació de les variables clau (DG i DC).

V. En base als registres de la submostra, obtenció de les classes d'equivalència de dimensió 1 (poblacions úniques a la submostra en funció de les combinacions possibles de valors de les variables clau).

VI. En base als registres de la mostra, obtenció de les classes d'equivalència de dimensió 1 (poblacions úniques a la mostra en funció de les combinacions possibles de valors de les variables clau).

VII. Comprovació de les poblacions úniques obtingudes a **(V)** que són també úniques a **(VI)**.

VIII. Inferència de p(identificació).

3.2. Estimadors dels components de p(identificació)

L'anàlisi està basat en un procés iteratiu per a diferents escenaris definits a (III) i (IV).

Definim:

u_2 : nombre de poblacions úniques observades a n_2 (submostra)

u_1 : nombre de poblacions úniques observades en n_1 (mostra)

u : poblacions úniques observades a la submostra que ho són també a la mostra.

$p(a) = 1$;

$p(b) = f$; la probabilitat d'estar present a la mostra és igual a la fracció de mostreig;

estimador de $p(c) = (u_1/n_1) \cdot (u/u_2)$; la probabilitat de que un individu sigui un element únic a la població, s'infereix a partir de la proporció de poblacions úniques vertaders a la mostra;

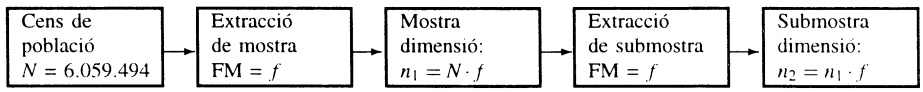
estimador de $p(d) = u/u_2$; la proporció de poblacions úniques vertaders a la submostra respecte a les observades a la submostra, s'utilitza com a estimador de la probabilitat de que una població única, observada a la mostra, sigui registre únic de la població; l'estimador de l'expressió (2) serà:

(3) estimador de p(identificació) $= 1 \cdot f \cdot ((u_1/n_1) \cdot (u/u_2)) \cdot u/u_2$;

4. RESULTATS I CONCLUSIONS

4.1. Aplicació del procés

La primera fase d'aplicació del procés consisteix en la construcció dels arxius de dades a partir de les quals es realitzarà l'anàlisi. L'esquema del procés d'obtenció d'aquests arxius, d'acord amb les fases 3.1.**I** i 3.1.**II** és:



Un objectiu intermedi és obtenir una dimensió n_2 de la submostra aproximadament igual a 10.000 unitats, que ens permet fer hipòtesis d'error acceptables en el càlcul dels estimadors puntuals de $p(c)$ i $p(d)$. El factor de mostreig (FM), queda determinat, ja que és funció de N i n_2.

Amb FM = f = 0,0407 es procedeix a una extracció de mostra a partir de l'arxiu del Cens de Població i Habitatges 1991, mitjançant un procediment Bernoulli, es construeix un arxiu de registres individualitzats i per a cada registre es recullen les dades de l'habitatge on resideix l'individu.

Amb el mateix factor f i amb el mateix mètode d'extracció s'obté una submostra a partir de la mostra.

D'acord amb el mètode d'extracció utilitzat, els resultats esperats són:

$E(n_1) = f \cdot N$

$E(n_2) = f^2 \cdot N$; la dimensió de la mostra (n_1) resulta ser de 245.944 unitats.

4.2. Determinació de les variables clau

Les variables que mostren una major potència en la identificació d'un individu són aquelles que representen les dades que típicament apareixen en els arxius personals

de tota classe d'institucions (bancàries, esportives, registres administratius, etc.). Les variables clau que s'han considerat base de l'anàlisi del risc de revelació són:

variables de localització:

Localització de residència

Lloc de naixement

Altres variables:

Edat (respecte a 1995, data d'extracció de la mostra)

Sexe

Estat civil

Ocupació o Ofici o Professió

Nivell d'estudis acabats

Relació amb l'activitat.

4.3. Determinació dels nivells de desagregació de les variables clau (DG i DC)

Dels components del risc de revelació, descrits a 2.4, la proporció de poblacions úniques és l'element principal a controlar. Els intruments disponibles que actuen directament sobre aquest component de risc són DG i DC de les variables clau d'identificació.

DG implica la dimensió geogràfica de referència de les variables de localització. Per determinar DG acceptables necessitem conèixer la dimensió mínima de la població de referència, per sota la qual el risc (poblacions úniques) és inacceptable. S'ha considerat que el factor geogràfic principal és el lloc de residència. És a dir, la variable clau: *Localització de residència* serà la variable geogràfica amb major nivell de desagregació, present a la mostra. La resta de variables de localització supeditaran la seva distribució per categories als nivells òptims d'informació del lloc de residència, dintre de valors mínims de risc. El càlcul del factor geogràfic [Greenberg i Voshell (1990a)] exigeix conèixer FM i tenir en compte les poblacions mínimes de les diferents particions territorials de Catalunya: municipis, comarques, regions, províncies, Catalunya. L'alta variació en la població comarcal de Catalunya (la més petita no supera els 4.000 habitants), fa impossible que DG es presenti a nivell comarcal, amb un FM del 4% aproximadament. En conseqüència, es dissenya una distribució per regions que representen agregacions comarcals i són coherents amb una proposta oficial de partició administrativa de Catalunya; mantenint la població de referència per sobre del llindar mínim, calculat en 40.000 habitants aproximadament, s'obtenen 14 regions que agrupen un nombre determinat de comarques cadascuna. La població mínima d'aquestes regions és de 146.778 habitants que corresponen a la de la Regió

11 (PENEDÈS) que agrupa les comarques de l'Alt Penedès i Garraf (segons dades del Cens 1991).

Per altra banda, el control de DC s'inicia aplicant un criteri de seguretat [Greenberg i Voshell (1990b)], consistent en no acceptar categories que donin freqüències, creuades amb el màxim DG del lloc de residència, que presentin un risc de població única.

4.4. L'anàlisi dels tres estrats

El mètode adoptat exigeix fer un càlcul iteratiu de *p*(identificació), partint de diferents combinacions de DG i DC, obtenint una sèrie de resultats (diferents quantificacions del risc de revelació) que ha de permetre seleccionar una combinació centrada en el compromís d'un mínim risc de revelació i un màxim contingut informatiu dels registres de la mostra.

Amb l'objectiu de fer operativa l'aplicació del procés s'han acotat les combinacions possibles de DG i DC:

- ***Lloc de residència***: tres nivells DG diferents,
  - **(1)** 1 estrat: Catalunya
  - **(2)** 4 estrats: nivell Provincial
  - **(3)** 14 estrats: distribució de Catalunya en 14 regions que agrupen un nombre determinat de comarques cadascuna. La població mínima d'aquestes regions és de 146.778 habitants que corresponen a la de la Regió 11 (Penedès) que agrupa les comarques de l'Alt Penedès i Garraf (segons dades del Cens 1991).
- ***Edat***: tres nivells de DC diferents,
  - **(1)** estrats d'1 any (1...) (estrat = edat
  - **(2)** 13 estrats de 5 anys (1...13) (13 = 65 anys i més
  - **(3)** estrats de 10 anys (1...8) (8 = 80 anys i més.
    S'ha fixat un nivell de desagregació per a la resta de variables, en funció dels criteris de seguretat.
- ***Lloc de naixement***: tres estrats (1...3); per al càlcul de l'estrat s'utilitza sempre la distribució en 14 Regions: s'obté la Regió que correspon al municipi de residència i la Regió corresponent al municipi de naixement; l'estrat 1 = nascut a la mateixa regió de residència; 2 = nascut en una altra regió de Catalunya; 3= nascut fora de Catalunya.
- ***Sexe***: dos estrats (1,6); estrat 1= Home; estrat 6 = Dona.
- ***Estat civil***: cinc estrats (1...5) segons la codificació del qüestionari del Cens.

- ***Ocupació o Professió***: vint estrats d'acord amb la codificació de la pregunta 23 del Cens (1...20).
- ***Nivell d'estudis acabats***: quinze estrats (1...15) d'acord amb la codificació de la pregunta 19 del Cens.
- ***Relació amb l'activitat***: deu estrats (1...10), corresponents al codi de la relació principal, d'acord amb la pregunta 22 del Cens.

4.5. Resultats del càlcul de *p*(identificació) i conclusions

Per a cada nivell DG de la variable **Lloc de residència** s'ha calculat la *p*(identificació) dels tres nivells DC de la variable **Edat**. La clau d'identificació dels registres (tant de la mostra com de la submostra), utilitzats per al càlcul, és el valor resultant de la combinació:

lloc residència	edat	lloc naixement	sexe	estat civil	professió	nivell estudis acabats	relació amb l'activitat

Els resultats obtinguts es mostren a la taula 1. Les files de la taula representen els valors obtinguts a diferents nivells de DG de la variable **Lloc de residència**. La lectura de les columnes ofereix el resultat de l'anàlisi a diferents nivells de desagregació de la variable **Edat**.

Resultats del càlcul de p(identificació):

		edat 1 any DC(1)	**edat 5 anys DC(2)**	**edat 10 anys DC(3)**
Catalunya DG(1)	u_1	33831	11737	9585
	u_2	4771	2460	2104
	u	1362	463	378
	p(identif.) × 100	**0,0455%**	**0,00686%**	**0,00511%**
Províncies DG(2)	u_1	50528	20073	16678
	u_2	5586	3309	2960
	u	2022	802	679
	p(identif.) × 100	**0,1093%**	**0,0194%**	**0,0144%**
14 regions DG(3)	u_1	83816	38349	32519
	u_2	6853	4847	4539
	u	3328	1537	1305
	p(identif.) × 100	**0,3264%**	**0,0636%**	**0,0443%**

Taula 1

4.5.1. Sensibilitat de les variables en funció de DG i DC

(a) La variable **Edat**, en el nivell DC(1) (estrats d'un any), és responsable dels valors més alts del risc de revelació.

(b) La variable **Referència geogràfica de residència**, en combinació amb la variable **Edat (DC(1))**, és responsable, tanmateix, d'un increment significatiu del risc de revelació al passar del nivell DG(1) —Catalunya— al (2) —Províncies— i al (3) —14 regions—. Nogensmenys, l'increment en el pes de DG(2) al DG(3) és molt menys significatiu quant a l'increment del risc.

(c) Els valors percentuals estimats de la probabilitat d'identificació són dins els rangs que s'han calculat en altres operacions del mateix tipus, tenint present que, per a nosaltres, el valor de $p(d)$ és molt més elevat del que s'accepta en altres casos (0.001, per a una combinació DG(3) DC(2) equivalent).

A la vista dels resultats, una combinació de nivells DG i DC per a les variables Lloc de residència i Edat que compleixi amb l'objectiu proposat de mantenir el compromís entre màxima informació i mínim risc, pot ser DG(3) i DC(2) amb una FM de 0,040625.

5. ESTRUCTURA FINAL I CARACTERíSTIQUES DE LA MOSTRA DE REGISTRES ANONIMITZATS (MRA) PROCEDENT DEL CENS 91

Característiques de la mostra

La mostra és aleatòria simple, amb procediment Bernoulli d'extraccions d'unitats
Unitat mostral: individu de la població, amb domicili habitual a l'habitatge (residents presents i absents), de Catalunya amb data 1 de març de 1991.
Població: 6.059.494
Probabilitat teòrica d'inclusió: 0,0407
Dimensió final de la mostra: 245.944
Factor de mostreig resultant: 0,0406
Error màxim absolut esperat, al 95% de confiança, per a un estimador de proporció de variància màxima: 0,02%

La mostra és representativa de la distribució de la població de Catalunya respecte als diferents estrats o categories en què es presenten les variables que contenen la informació.

El nivell de desagregació o estratificació de les variables ha estat subjecte a un control per tal d'eliminar el risc de revelació estadística.

En la mostra, per tant, hi ha un límit físic en la desagregació de les dades i un altre lògic consistent en la pèrdua de fiabilitat en l'intent d'analitzar subpoblacions o dominis a partir de la intersecció de categories en un nivell de desagregació més baix.

Variables presents en la mostra i nivells de desagregació

Es distingeixen dos tipus de variables:

(1) Les variables que contenen informació corresponent a l'individu, unitat mostral primària.

(2) Les variables amb la informació referent a l'habitatge de residència de l'individu present a la mostra.

(1) Variables de l'individu

VARIABLE	DEFINICIÓ	ESTRATS/CODIFICACIÓ
REG_R	Regió de residència	**14 estrats / 1...14**
		Codificació pròpia, distribució de Catalunya en 14 regions
REG_N	Regió de naixement	**3 estrats / 1...3**
		1= nascut a la mateixa regió REG_R
		2= nascut a catalunya, diferent REG_R
		3= nascut fora de Catalunya
REG_T	Regió de treball	**15 estrats / 1...14, 99**
		Codificació pròpia, distribució de Catalunya en 14 regions
		99 = treballa fora de Catalunya
PREG_2	Relació amb la persona	**12 estrats / 1...12**
	principal	Codificació original del qüestionari Cens-91
		Pregunta 2 del qüestionari del Cens-91
PREG_6	Sexe	**2 estrats / 1, 6**
		Codificació original del qüestionari Cens-91
		Pregunta 6 del qüestionari del Cens-91
PREG_7	Edat (l'any 1995)	**13 estrats / 1...13**
		Estrats de 5 anys
		13 = 65 o més.
		Pregunta 7 del qüestionari del Cens-91
PREG_12	Estat civil	**5 estrats / 1...5**
		Codificació original del qüestionari Cens-91

VARIABLE	DEFINICIÓ	ESTRATS/CODIFICACIÓ
PREG_14	Lloc de residència fa 1 any,	**5 estrats / 1...5**
PREG_15	5 i 10 respectivament	Codificació original del qüestionari Cens-91
PREG_16		Preguntes 14, 15 i 16 del qüestionari del Cens-91
PREG_17	Any d'arribada a Catalunya	**Valor absolut**
		Pregunta 17 del qüestionari del Cens-91
PREG_17_2	Procedència	**2 estrats / 1, 6**
		1 = d'un altre municipi de l'Estat espanyol
		6 = de l'estranger
		Pregunta 17 del qüestionari del Cens-91
PREG_19	Nivell d'estudis	**15 estrats / 1...15**
		Codificació original del qüestionari Cens-91
		Pregunta 19 del qüestionari del Cens-91
PREG_22	Relació amb l'activitat	**10 estrats / 1...10**
		De les tres situacions permeses en el qüestionari del Cens, es recull la principal
		Pregunta 22 del qüestionari del Cens-91
PREG_23	Ocupació, professió o ofici	**20 estrats / 1...20**
		Codificació de la pàgina 33 del qüestionari
		Pregunta 23 del qüestionari del Cens-91
PREG_24	Situació professional	**7 estrats / 1...7**
		Codificació original del qüestionari Cens-91
		Pregunta 24 del qüestionari del Cens-91
PREG_25	Activitat principal de l'establiment on treballa	Codificació automàtica a partir del literal en termes de la CNAE-74 (3 dígits).
		Pregunta 25 del qüestionari del Cens-91
PREG_26_1	Codi del lloc de treball	**5 estrats / 1...5**
		Codificació original del qüestionari Cens-91
		Pregunta 26 (part primera) del qüestionari del Cens-91
PREG_27	Mitjà de transport	**9 estrats / 1...9**
		Codificació original del qüestionari Cens-91
		Pregunta 27 del qüestionari del Cens-91

(2) Variables de l'habitatge

VARIABLE	DEFINICIÓ	ESTRATS/CODIFICACIÓ
HABI	Nombre d'habitatges de l'edifici	**3 estrats / 1...3** Codificació original del qüestionari Cens-91
TIPV	Tipus d'habitatge	**7 estrats / 1...7** Codificació original del qüestionari Cens-91
SUPF	Superfície	**Valor absolut**
TINEN	Règim de tinença	**8 estrats / 1...8** Codificació original del qüestionari Cens-91
INFRA	Informació sobre les característiques materials de la llar[2]	La codificació és directa del qüestionari del Cens-91
	Aigua corrent	**1-1 (1...3)**
	Aigua calenta	**2-2 (1,6)**
	Refrigeració	**3-3 (1,6)**
	Cuina	**4-4 (1,6)**
	Electricitat	**5-5 (1,6)**
	Gas	**6-6 (1,6)**
	Telèfon	**7-7 (1,6)**
	Calefacció	**8-8 (1...4)**
	Combustible	**9-9 (1...6)**
	Vàter	**10-10 (1...3)**
	Nombre de vàters	**11-12 (XX)**
	Bany	**13-13 (1,6)**
	Nombre de banys	**14-15 (XX)**

REFERÈNCIAS METODOLÒGIQUES CONSULTADES

[1] **Bethlehem, J.G., Keller, W.J.** i **Pannekoek, J.** (1990). «Disclosure Control of Microdata». *ASA*, **85, 409**, *Application and Case Studies*, 38–45.

[2] **Bled** (1996). *3rd. International Seminar: Statistical Confidentiality*. Bled / Slovenia, 2-4 Octubre 1996.

[2] Aquesta variable està representada per una cadena de dígits. Cadascuna de les posicions correspon als conceptes que s'indiquen.

[3] **EUROSTAT** (1994). *Protection of Confidential Data in Eurostat.* Document presentat a: TES-Statistical Disclosure Control Seminar, Voorburg, Juny 1995.

[4] **Dalenius, T.** (1977). «Towards a methodology of statistical disclosure control». *Statistik tidskrift*, **5**, 429–444.

[5] **Duncan G.T.** i **Lambert D.** (1986). «Disclosure-Limited Data Dissemination». *ASA*, **81, 393**, Applications, 10–18.

[6] **Greenberg, B.** i **Voshell Zayatz, L.** (1990). *The Geographic Conponent of disclosure risk for microdata. Bureau of the census.* Statistical research division report series. Census/SRD/RR-90/13.

[7] **Greenberg, B.** i **Voshell Zayatz, L.** (1990). *Strategies for measuring risk in public use microdata files.* Symposium on Statistical Disclosure Avoidance, Voorburg.

[8] **Marsh, C., Skinner, C.** i **altres** (1991). «The Case for Samples of anonimized records from the 1991 Census». *J.R. Statist. Soc. A* **154**, Part 2, 305–340.

[9] **Marsh, C., Dale, A.** I **Skinner, C.** (1991). «Safe data versus safe settings: access to customized results from the British Census». *48th. I.S.I. Session. El Cairo.* Septembre 9–17.

[10] **Skinner, C., Marsh, C., Openshaw, S.** i **Wymer, C.** (1994). «Disclosure Control for Census Microdata». *Journal of Official Statistics*, **1**, 31–51.

[11] **Voshell Zaiatz, L.** (1991). *Estimation of the percent of unique population elements on a microdata file using the sample.* Bureau of the Census. Statistical research division report series. Census/SRD/RR-91/08.

[12] **Willenborg, L.C.R.** (1996). *Establishment of a work session on statistical disclosure control: a proposal.* Paper presentat al Seminar on Integrated Statistical Information Systems and Related Matters, Bratislava, Slovakia, 21–24.

[13] **Willenborg, L.C.R.** i **De Waal, T.** (1996). *Statistical Disclosure Control in Practice.* Springer-Verlag.

El número 1 del volum 46-1992, de la revista *Statistica Neerlandica*, monogràfic sobre el tema del risc de revelació per a microdades, és una antologia de contribucions metodològiques.

ENGLISH SUMMARY

CONTROL OF STATISTICAL DISCLOSURE RISK IN DISSEMINATION OF MICRODATA FROM POPULATION, APPLIED ON A SAMPLE OF INDIVIDUAL RECORDS FROM CATALONIA'S POPULATION CENSUS 1991

ALFONS GARÍN i RAMÍREZ*

Official statistical offices are committed to disseminate available information obtained under conditions of confidentiality. By devising a sample of anonymised records the Institut d'Estadística de Catalunya fulfils a double commitment: to disseminate information while keeping statistical secret. This paper describes the process of statistical disclosure analysis and control applied to the microdata sample.

In short, the approach is the following: a single element of the population in the microdata sample, possessing a unique combination of characteristics, involves disclosure risk. An intruder could compare this unique combination of categorical variables with the same combination found in another data file containing formal identifiers of that individual. Thus both files could be matched in order to again access to sensitive information belonging to that individual.

The goal is to limit disclosure risk to make practically impossible to identity a person while keeping data information-worthy, by means of some model which allows calculation of the risk, and data protection procedures.

Keywords: Protection of Census Microdata, Statistical Disclosure Control, Sample of Anonimized Records.

---

*El treball descrit en aquest article es desenvolupà a la Subdirecció d'Assistència Tècnica Estadística de l'Institut d'Estadística de Catalunya; actualment, Alfons Garín pertany al Gabinet de Planificació i Avaluació de la Universitat Politècnica de Catalunya.

-Received april 1996.

-Accepted september 1996.

1. INTRODUCTION

A common problem that official statistical offices have for long found themselves faced with is that of protecting statistical records from disclosure when they are disseminating data regarding population censuses by means of a probabilistic sample in either of its two traditional forms, which are:

- Flat files, in which every record corresponds to a single individual.
- Hierarchic files, in which the sample unit is people living in the same household.

In both cases, the observed rule (Ponencia de Protección de Datos) is not giving formal identifiers such as names, adresses, I.D. numbers, etc. In other words, records are anonymized.

It is a widely known fact that record anonymization doesn't cancel statistical disclosure risk. Data that wouldn't be available without the dissemination of the microdata sample may now be used to obtain information about a certain individual (Dalenius' statistical disclosure concept). Disclosure may be achieved by different methods, the one more deeply studied being that of identifying the target individual by matching the values of certain variables with those contained in a list or a database that the statistical spy has, which contains the formal identifiers.

Given this situation, and taking into account the ethical code that binds them (obligation to disseminate as much information as possible for public use while complying with the law regarding personal data protection), official statistics offices need to develop methods of control for minimising disclosure risk.

The Institut d'Estadística de Catalunya has worked on this problem by applying the most adequate procedures to each statistical disclosure component in each case. These are based on Catalonia's population characteristics. The following are the main guidelines regarding data protection and disclosure control strategies that have been aplied.

Microdata dissemination policy

Public institutions devoted to the study and statistical analysis of population's characteristics are the main users of the information published by the IEC. The aim of these studies is to enhance the general knowledge on both the sociodemographic reality and the information analysis' methods and procedures.

Data protection methodology

There are two general methods:

- Making probabilistic modifications in data to alter the information given.

- Supressing information by means of supressing variables, aggregating categories, top coding, etc., whenever these items present a high disclosure risk in their original form.

In its sample of individual records, the IEC follows the method of supressing information, restraining from any kind of alteration of data.

Model for studying disclosure risk

Two models have been historically used to develop a estimation method that will enable to quantify disclosure risk. They are:

- The predictive model proposed by Duncan and Lambert [Duncan i Lambert (1986)]; a Bayesian model, in wich the identification of the target individual is not crucial. It is based on both the current and the previous predictive distributions, connected by an uncertainty function.

- The model, which we might call of identification , based on the analysis of the frequency of the variables that may be used to identify a certain individual. This method uses data from the available sample and drastically separates identificative variables from variables containing sensitive information, which won't be used as an identificative key.

The IEC follows the second method.

Identification rules

An identification rule is the procedure followed by an investigator trying to match a record from a microdata sample and the target individual, making sure that the equivalence is correct. The rule used to study the IEC's sample establishes two conditions:

- The equivalence must be exact.

- The target individual must constitute a unique case (The combination of values in the identification key has to be unique) in the population as a whole, not only in the sample.

Main problem in estimating disclosure risk: population uniques

Depending on the identification rule followed, the probability of identification is a function of the proportion of unique cases in population given a combination of key variables. The three most operational methods to make this estimation are:

- The Poisson-Gamma model, developed, amongst others, by Bethlehem i altres [Bethlehem *et alt.* (1990)].

- Analyzing the distribution of the equivalence classes found in the sample [Voshell (1991)].
- The subsample method [Voshell (1991)].

In calculating the proportion of unique populations, the IEC follows the subsample method.

The problem of variables containing geographic references

Along with the variable AGE, location variables have proved very influential in the variation of disclosure risk. These are the variables that provide information about the residence, place of birth, place of work, and so on. It is necessary to determine the minimum size (under wich disclosure risk is unacceptable) of the population contained in the geographic reference zones. It is also necessary to decide which variable is going to be allowed a higher degree of geographic reference detail, aggregating categories in the rest of location variables. This prevention is based on empirical analysis as well as on the knowledge of subpopulations with singular characteristics with very accurate location data (such as lists of professionals).

The estimation of the disclosure risk undertaken by the IEC has been done by fixing three territorial division levels in Catalonia for the PLACE OF RESIDENCE variable, along with three different strata for the AGE variable, thus obtaining a complete table of values which has been used to asses the ideal combination. That is, the one that offers enough information to study Catalonia's population characteristics, protecting at the same time confidential data.