

DISSENY I CONSTRUCCIÓ D'UNA MOSTRA ESTRATIFICADA A PARTIR DE DADES CENSALS*

P. LÓPEZ*

C. LOZARES*

Universitat Autònoma de Barcelona

M. DOMÍNGUEZ**

Universitat de Barcelona

L'objectiu de l'article és presentar les principals característiques del disseny i del procés de construcció de la mostra estratificada de l'«Enquesta metropolitana de Barcelona. Condicions de vida i hàbits de la població» de l'edició de l'any 1995. A partir de la informació que prové del cens de població s'agrupen a les persones en seccions censals i es procedeix a la construcció dels estrats d'acord amb un procediment on s'impliquen tècniques d'anàlisi multivariable: d'anàlisi factorial de components principals i de classificació automàtica. Finalment, s'afixa una grandària de mostra segons el criteri òptim de Neyman. El procediment seguit i els criteris i decisions que s'utilitzen van més enllà dels resultats habituals esperats a tota mostra, doncs proporciona conclusions d'interès per a l'anàlisi sociològica i la planificació del territori, al mateix temps que serveix de criteri de validació de les conclusions posteriors a l'anàlisi de l'enquesta.

Desing and construction of a stratified sample from census data

Paraules clau: Enquesta per mostreig, anàlisi de dades, anàlisi de components principals, anàlisi de classificació

Classificació AMS (MSC 2000): 62D05, 62-07, 62H25, 62H30

* Els tres autors són membres del Grup d'Estudis Sociològics sobre la Vida Quotidiana i el Treball (QUIT) del Departament de Sociologia de la Universitat Autònoma de Barcelona.

* Pedro López Roldán (Pedro.Lopez.Roldan@uab.es); Carlos Lozares Colina (Carlos.Lozares@uab.es). Departament de Sociologia. Universitat Autònoma de Barcelona.

** Màrius Domínguez Amorós (marius@riscd2.eco.ub.es). Departament de Sociologia. Universitat de Barcelona.

– Rebut el setembre de 1999.

– Acceptat el novembre de 1999.

1. INTRODUCCIÓ

L'objectiu d'aquest article és presentar les característiques de major rellevància relatives al disseny i al procés de construcció de la mostra estratificada que serveix de base per a la recollida (producció) d'informació de l'*Enquesta Metropolitana de Barcelona. Condicions de Vida i Hàbits de la Població* (EM) de l'edició de l'any 1995, elaborada al sí de l'*Institut d'Estudis Metropolitans de Barcelona*.¹

L'EM constitueix una investigació permanent que va ser projectada l'any 1984 per a analitzar les activitats, condicions i formes de vida de la població de l'Àrea Metropolitana de Barcelona davant l'absència de dades estadístiques sistemàtiques i objectives, de caràcter social i sobre aquest contingut. L'EM s'ha convertit en un instrument periòdic de recollida d'informació que ens ajuda a precisar l'evolució i els canvis de les tendències més estructurals del conjunt de fenòmens socials que s'analitzen a l'enquesta².

Encarregada inicialment per l'antiga Corporació Metropolitana de Barcelona (CMB), la primera edició (1985) considerà els 27 municipis que constituïen la CMB. L'edició de 1990, encarregada per la Mancomunitat de Municipis de la Regió Metropolitana de Barcelona i la Diputació de Barcelona, va estendre l'univers poblacional a l'anomenada Regió Metropolitana de Barcelona (RMB) reunint les 5 comarques que conformen la Regió I dins de la divisió territorial de Catalunya. Finalment, la darrera edició de 1995 va afegir dins la RMB les comarques de l'Alt Penedès i del Garraf³.

L'EM recull informació a partir del disseny d'una mostra estratificada que s'ha mantingut al llarg de les tres edicions, si bé l'àmbit territorial de referència s'ha anat modificant. El punt de partida és la informació que prové del cens de població. A partir de l'agrupació de les persones en seccions censals es disposa d'informació en termes de la proporció de persones que posseeixen determinada característica sobre el total de la secció censal, segons un conjunt d'indicadors seleccionats. A partir d'aquesta infor-

¹Aquest article va ser presentat pels autors en forma de comunicació a les «Jornades internacionals sobre generació d'informació estadística: qualitat i limitacions», organitzades a Barcelona el novembre de 1998 per la xarxa temàtica *Enquestes i qualitat de la informació estadística*. L'exposició detallada del disseny i del procés de construcció de la mostra, per a l'edició de l'Enquesta Metropolitana de 1990, es pot consultar a Lozares i López (1990). Una versió reduïda d'aquest document va ser publicat a Lozares i López (1991b). Posteriorment, prenent com a base aquesta mostra es va considerar el tractament de subpoblacions per a la construcció de zones socials (Lozares i Domínguez, 1993, 1996). A partir d'aquests antecedents fonamentem el contingut d'aquest article, però considerant el material inèdit del procés de construcció de la mostra de l'any 1995. L'Enquesta Metropolitana de l'any 1995 ha estat dirigida per Marina Subirats.

²Són nombroses les publicacions que s'han generat amb les anàlisis de les dades de l'EM, l'edició de les quals ha anat a càrrec de l'Institut d'Estudis Metropolitans de Barcelona.

³En el projecte de la nova edició de l'any 2000 s'està considerant tota la província de Barcelona. Al llarg d'aquests anys, i davant la necessitat de mantenir la sèrie temporal, sempre s'ha considerat el manteniment de submostres representatives que facilitessin la continuïtat i la comparació entre les diferents edicions.

mació es procedeix a la construcció dels estrats d'acord amb un procediment en el que s'impliquen tècniques d'anàlisi multivariàble: d'anàlisi factorial de components principals i de classificació automàtica.

Amb els estrats prèviament analitzats i construïts es procedeix a l'afixació d'una grandària de mostra donada segons el criteri òptim de Neyman. Després de la distribució per estrats es realitza l'assignació de quotes de mostra a cada secció censal de cada estrat segons la seva població. Les unitats finalment s'extreuen de forma aleatòria a partir del marc de la mostra que s'obté del registre del cens electoral.

El procediment seguit i els criteris i decisions que s'utilitzen van més enllà dels resultats habituals esperats a tota mostra, doncs, com destacarem, proporciona conclusions d'interès per a l'anàlisi sociològica, al mateix temps que serveix de criteri de validació de les conclusions posteriors en l'anàlisi de l'enquesta.

D'altra banda, i com a resultat d'una anàlisi específica realitzada amb posterioritat a la recollida d'informació, s'ha reutilitzat la mostra amb l'objectiu de construir el que hem anomenat com a zones socials. Cada estrat de la mostra estratificada pot constituir una base mostral per a dur a terme anàlisis amb totes les garanties de precisió, com a mínim calculables i acceptables, sense augmentar la grandària de la mostra per a l'enquesta general.

Tots aquests aspectes ressenyats són els que articulen el contingut de l'article als apartats següents.

2. LA CONSTRUCCIÓ DE LA MOSTRA DE L'ENQUESTA METROPOLITANA

La construcció de la mostra de l'EM per a l'any 1995 manté les mateixes característiques de procediment que en les dues anteriors edicions. Tot seguit especificarem els trets més significatius del disseny mostral i del procés de construcció que s'organitza a partir de la distinció de cinc moments:

1. Definició de l'univers poblacional i selecció del camp de les variables/criteri d'estratificació.
2. Anàlisi de dimensionalització per components principals amb l'objectiu de reduir i sintetitzar els factors principals de variabilitat de la informació original.
3. Anàlisi de classificació automàtica que, en funció dels factors principals, agrupi les seccions censals en grups homogenis, en els estrats de la mostra.
4. Determinació de la grandària i de l'error mostrals amb la corresponent afixació de la mostra entre els estrats.

5. Ponderació a posteriori de la mostra per a restituir l'equiprobabilitat de que un individu sigui elegit a l'atzar.

2.1. Disseny de la mostra: univers poblacional i variables criteri

En el disseny de la mostra es va establir com a objectiu l'extracció d'una mostra aleatòria estratificada representativa de la població de la Regió Metropolitana de Barcelona. L'elecció del procediment de mostreig estratificat es justifica bàsicament per criteris de precisió front altres mètodes i per l'heterogeneïtat social que caracteritza la població objecte d'estudi. Així, la construcció dels estrats ens permet disposar d'una variable d'estratificació que constituirà una variable densa de tipificació social i que ens garantirà la presència a la mostra de tipus socials homogenis característics de la població, tipus o grups socials que estan en la base, es correlacionen, amb els objectius d'estudi de l'EM.

L'univers poblacional es defineix com el conjunt de persones majors de 18 anys de la RMB. A l'edició de 1995 la RMB comprèn les 7 comarques catalanes següents: Barcelonès, Baix Llobregat, Maresme, Vallès Occidental, Vallès Oriental, Alt Penedès i Garraf. Es tracta d'un territori que comprèn 162 municipis i una població total de 3.275.458 persones, segons les dades del Cens de Població de 1991.

La condició indispensable de poder disposar de les dades poblacionals del cens ens ofereix la possibilitat de l'estratificació. De la informació disponible del qüestionari del cens hem seleccionat una sèrie d'indicadors que són els que actuen com a variables/criteri d'estratificació i que caracteritzaran, per l'agregació de les persones, a les seccions censals de la RMB. Així doncs, la unitat elemental a estratificar que hem considerat no són, en primera instància, les persones censades, sinó les seccions censals on aquestes resideixen. Procedir d'aquesta manera es justifica per un doble motiu: per les dificultats que es deriven de tractar una matriu d'individus d'aquesta magnitud; i perquè tan sols l'agregació en seccions censals ens permet un tractament percentual mètric de les variables seleccionades. Tornarem a l'individu en el moment d'aplicar la fórmula per a calcular la grandària de la mostra. Llavors establim el supòsit, per a fer el salt de les seccions als individus, segons el qual els individus d'una mateixa secció censal posseeixen la mateixa condició socioeconòmica.

La selecció final de les variables/criteri que considerarem obeeix en primer terme a la seva disponibilitat com informació recollida al cens de l'any 1991, també respon a criteris de pertinença conceptual d'acord amb els objectius de l'estudi de l'EM, i finalment a criteris de tipus estadístic que tenen a veure amb els resultats del procés d'anàlisi que presentarem a continuació i que ens van conduir a l'eliminació d'alguna de les variables inicialment considerades (per manifesta combinació lineal o per l'escàs valor o dispersió de les variables).

En conseqüència, tenim una matriu de dades amb 3586 seccions censals de la RMB amb la selecció dels 16 d'indicadors o variables socioeconòmiques que apareixen a Taula 1. Es tracta de variables de caracterització demogràfica, cultural-educativa, d'activitat laboral i professional, de mobilitat i de grandària poblacional. Les variables a la matriu original expressen el percentatge de la població de la secció censal que posseeix una determinada característica sobre el total de la població de la secció censal corresponent. Les mitjanes i les desviacions són estadístics computats sobre valors de percentatges de cada secció sobre total de seccions.

Taula 1. Mitjana, desviació i descripció de les variables/criteri de la mostra

Variable	Mitjana	Desviació	Descripció
P1	14,23	5,02	Joves de menys de 15 anys
P2	16,24	6,76	Vells majors de 65 anys
P4	33,59	1,40	Immigració fora Catalunya
P7	2,18	0,38	Analfabets majors de 10 anys
P8	27,63	5,38	Titulats mitjans-superiors majors 20 anys
P9	57,96	3,95	Escolarització 14-24 anys
P11	12,16	4,16	Aturats abans ocupats
P12	3,20	1,93	Atur busca primera feina
P14	38,69	5,79	Dones actives majors 15 anys
P15	17,82	12,82	Professions altes
P16	37,28	17,91	Professions baixes
P17	19,07	4,49	Terciari mitjà/comerç/hosteleria
P18	4,28	3,09	Terciari alt/finances
P19	0,91	0,60	Agropecuari
P20	43,09	13,80	Vehicle privat treball
P23	4,54	4,77	Població Secció/Municipi

La construcció dels estrats homogenis de població és l'objectiu buscat. Es tracta de garantir que a la mostra hi siguin representades una sèrie de característiques de la població —i els fenòmens que en depenen—. El procés de construcció dels estrats es va fer mitjançant la utilització de dues tècniques d'anàlisi multivariable independents i complementàries: l'anàlisi factorial de components principals i l'anàlisi de classificació automàtica.

2.2. L'anàlisi de components principals

Amb l'anàlisi de components principals (ACP) es pretén reduir la informació original per tal d'obtenir un subespai vectorial de menys dimensions o factors, on aquests són

base i per tant linealment independents, i que, ordenats de manera jeràrquica, conserven la major part de la variança total. Obtenim així les dimensions fonamentals de diferenciació de la població de la RMB que estructurin inicialment la realitat social segons la informació introduïda. L'ACP es concep així com una etapa prèvia i complement necessari de la categorització de les unitats, d'obtenció dels estrats.

En el procés de l'ACP distingim cinc moments bàsics que passem a comentar: càlcul dels eixos factorials o components; càlcul dels valors propis i nombre d'eixos a retenir; correlació de les components amb les variables originals, comunalitats i recomposició de la matriu de correlacions; interpretació dels factors amb la rotació dels eixos; i recomposició de la matriu d'unitats en els nous eixos retinguts.

El càlcul dels eixos parteix de les 16 variables originals que apareixen a la Taula 1, la matriu de correlacions de les quals es presenta a la Taula A1 de l'annex⁴. L'examen de les correlacions de la matriu i altres mesures ens confirmen l'adequació de la informació considerada: la significativitat de totes les correlacions, el determinant no nul i pròxim a zero, el comportament de la matriu antiimatge de correlacions, la significativitat del test de Barlett i l'obtenció d'un índex d'adequació mostral de Kaiser-Meyer-Olkin de 0,83.

Essent la matriu de correlacions no singular, el càlcul dels valors propis o variança incorporada a cada eix ens porta a l'obtenció de 16 vectors propis i valors propis associats que es presenten a la Taula 2. Si apliquem els criteris habitualment emprats a l'hora de decidir el nombre d'eixos a retenir, l'espai vectorial original pot reduir-se a només 4 dimensions o variables factorials, acumulant el 76,6 % de la variança total.

La relació entre les variables primitives i les components obtingudes ens permetrà simultàniament recompondre les variables originals en els nous eixos, el que ens mostrarà l'estructura el primer espai d'atributs, i donar identitat a les components. A la Taula A2 de l'annex es presenta la matriu de saturacions que mostra aquesta relació, en aquest cas després d'haver aplicat una rotació varimax. Als Gràfics 1 i 2 es representen igualment aquestes dades que complementaran la interpretació de la identitat dels factors o components.

El primer eix o component, que acumula el 41 % de la variança total⁵, té un pes cabdal, sent per tant una dimensió determinant de l'estructura de relacions de les variables originals. L'eix o la dimensió es defineix a partir de la polaritat que oposa, d'una banda, la presència de professions baixes, d'immigració de fora de Catalunya, de l'analfabetisme

⁴Tots els resultats de les anàlisis presentades per a la construcció de la mostra s'han realitzat mitjançant el programari estadístic SPSS.

⁵Considerem el percentatge de varianza explicat pels factors inicialment, sense considerar a l'explicació la redistribució que implica la rotació dels eixos.

i de l'atur; de l'altra banda, la dimensió recull en aquest pol el fet de tenir una professió alta i una ocupació en el terciari alt/finances, de posseir un nivell de titulació mitjà o superior, amb altes taxes d'escolarització entre 14 i 24 anys. Es tracta doncs d'una dimensió de *categoria social entesa com un compost de categoria professional, nivell educatiu, inserció laboral i origen immigrant*.

El segon eix, amb el 17,7 % del total de la varianza, reflecteix una dimensió que oscil·la entre la presència d'altas proporcions de població jove de menys de 15 anys, on hi ha una major taxa de dones actives i s'empra el vehicle privat per anar a treballar, en un extrem, i el predomini de població vella de més de 65 anys, juntament amb un certa presència del terciari mitjà/comerç/hosteleria, en l'altre extrem. És una dimensió de *cicle vital o d'edat*, que oposa a les seccions censals de població més jove, resident a la perifèria metropolitana i dels municipis front al nuclis urbans més centrals on la proporció de població vella és més important.

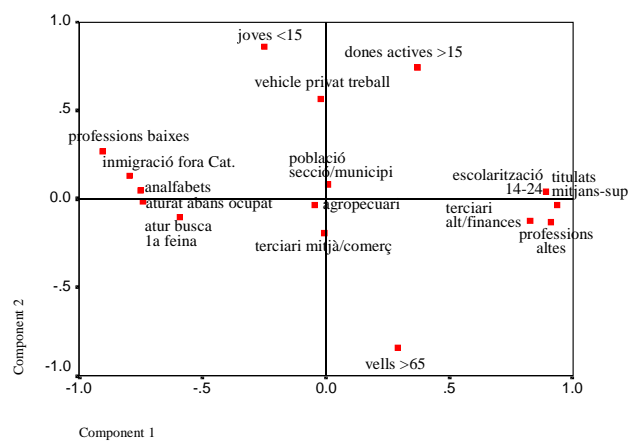
Taula 2. Valors propis (autovalors) obtinguts a l'ACP

Component	Autovalors inicials			Sumes de les saturacions al quadrat de l'extracció		
	Total	% de la varianza	% acumulat	Total	% de la varianza	% acumulat
1	6,561	41,0	41,0	6,561	41,0	41,0
2	2,828	17,7	58,7	2,828	17,7	58,7
3	1,891	11,8	70,5	1,891	11,8	70,5
4	,978	6,1	76,6	,978	6,1	76,6
5	,822	5,1	81,8			
6	,586	3,7	85,4			
7	,476	3,0	88,4			
9	,324	2,0	92,9			
10	,299	1,9	94,8			
11	,255	1,6	96,4			
12	,234	1,5	97,8			
13	,152	,9	98,8			
14	,130	,8	99,6			
15	,038	,2	99,8			
16	,028	,2	100,0			

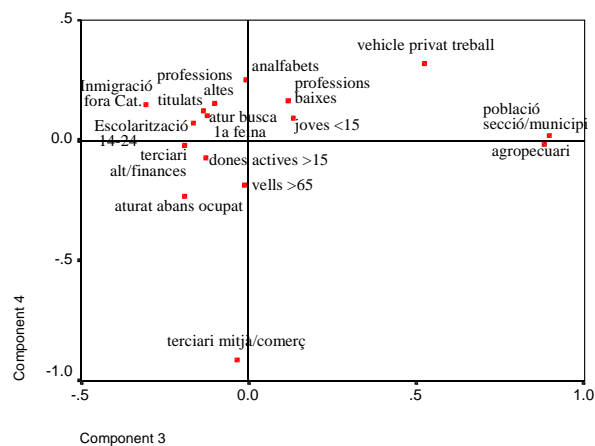
El tercer eix, que té l'11,8% del total de la varianza, es defineix con una dimensió d'*identitat metropolitana*, d'oposició rural-urbà o grau de «metropolització», doncs ens mostra la polaritat entre una alta proporció de població ocupada al sector agropecuari i valors alts en el quocient poblacional secció/municipi, i pràcticament la resta de variables.

El quart eix, finalment acumula el 6,1 % de la varianza, essent una dimensió definida netament per l'oposició entre la variable terciari mitjà/comerç/hosteleria i la resta; és per tant una dimensió de *terciarització* que diferencia les seccions censals on hi ha una important presència d'activitat de comercial i hostelera de la resta.

Amb aquesta caracterització dels eixos factorials, finalment, la recomposició de les unitats seccions censals en el subespai vectorial ens porta a obtenir les puntuacions factorials en l'espai reduït de quatre dimensions que seran reutilitzades com a punt de partida de l'anàlisi de classificació automàtica.



Gràfic 1. Components 1 i 2 en l'espai rotat



Gràfic 2. Components 3 i 4 en l'espai rotat

2.3. L'anàlisi de classificació automàtica

L'estratificació de la mostra descansa sobre el principi de que la població en estudi careix d'homogeneïtat a efectes estadístics. L'homogeneïtat estadística fa referència a la relació que s'estableix entre les distintes característiques d'una mateixa població, quant més correlacionades estiguin aquestes característiques més homogènia serà la població considerada. L'objectiu de l'estratificació serà doncs la classificació de les seccions censals en estrats que, a efectes del mostreig, de guany en la precisió, seran l'expressió de conjunts de seccions el més homogènies possible a dintre de cada estrat, i el més heterogènies entre elles, segons les variables/criteri factorials derivades de l'anàlisi anterior.

Aquesta classificació es realitza sense tenir en compte cap restricció de contigüitat territorial, per la qual cosa s'obindrà com a resultat un mapa de seccions de la RMB de diferent caracterització social segons la seva pertinença als estrats, la presència dels quals tindrà per tant una distribució diferenciada al territori.

A partir de la matriu reduïda que es deriva de l'ACP, de 4 variables de puntuacions factorials que identifiquen les 3586 seccions censals, hem considerat afegir, per la importància desigual de la grandària de les seccions en relació al municipi i els seus efectes sobre la mostra, una cinquena variable que manifestés explícitament aquesta relació.

En el procés de classificació automàtica, l'aplicació de l'anàlisi factorial previ ens proporciona dues avantatges d'interès: l'obtenció de quatre components factorials que són les variables reduïdes que més discriminen a les seccions censals i la incorrelació d'aquestes, dues característiques essencials en la construcció dels grups o estrats. El mètode de classificació emprat, a partir de considerar com a mesura de proximitat la distància quadràtica euclidiana, es desglossa en dues etapes: primer s'aplica una classificació jeràrquica ascendent mitjançant el procediment ward (o de mínima pèrdua d'inèrcia), en un segon moment, amb el nombre de grups determinat i els centres inicials definits, s'opera una classificació no jeràrquica per l'agregació al voltant de centres mòbils amb l'objectiu d'optimitzar l'assignació als estrats. D'aquesta forma finalment s'obtenen 8 estrats, la distribució de freqüències i els centres finals dels quals es presenten a la Taula 3.⁶

⁶El procés de classificació comporta a més diversos procediments de validació: una anàlisi sistemàtica de les classificacions entre 15 i 5 estrats, la comparació amb el resultat obtinguts en edicions anteriors de l'enquesta, la comparació del procediment de classificació *ward* amb altres jeràrquics ascendents (distàncies màximes, mínimes, distàncies promig, mediana i centroide), la replicació de la classificació a partir de submostres així com la utilització de criteris teòrics-interpretatius.

Taula 3. Distribució final de les seccions censals per estrat i centres finals

Estrat	Seccions finals		Centres finals				
			FSC1	FSC2	FSC3	FSC4	ZP23
1	159	4,5 %	0,1163	0,5210	1,8965	0,2306	1,6172
2	661	18,5 %	-0,6742	0,8633	-0,0267	0,2349	-0,0464
3	539	15,1 %	0,5725	1,0295	-0,2933	-0,5932	-0,2002
4	71	2,0 %	0,2549	-0,3544	5,5014	-0,0575	6,1028
5	642	18,0 %	-1,1779	-0,3439	-0,2648	0,7265	-0,2211
6	449	12,6 %	-0,2770	-0,8112	-0,1391	-1,5785	-0,2509
7	647	18,1 %	0,5030	-0,8780	-0,2312	0,0141	-0,2832
8	400	11,2 %	1,6390	-0,0754	-0,3357	0,9127	-0,2850
Total	3568	100,0 %	0,0000	0,0000	0,0000	0,0000	0,0000

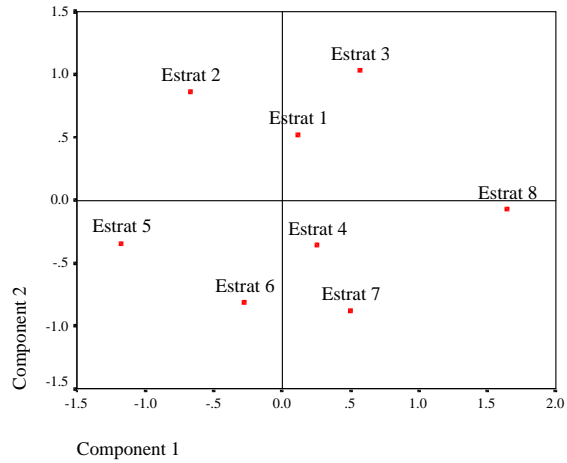
La caracterització sociològica dels vuit estrats és la que presentem breument tot seguit, de manera descriptiva i esquemàtica⁷. Les dades de referència es poden observar a la Taula A3 de l'annex, amb les mitjanes i desviacions de les diferents variables, i visualitzar en els Gràfics 3 i 4, d'ubicació dels estrats en els eixos factorials. La caracterització dels estrats es presenta a continuació en ordre decreixent d'acord amb la gradació de trets que es configuren amb la categoria socioprofessional, el nivell d'estudis, l'ocupació i l'origen de naixement.

L'estrat 8 identifica a la població de seccions censals on predominen els grups o les posicions socials de classe alta, són persones formades i ocupades, nascudes a Catalunya, és una població madura i urbana, i amb una taxa d'activitat femenina alta.

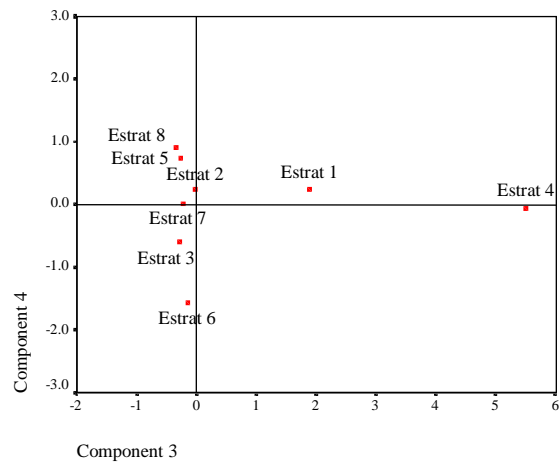
L'estrat 7 es correspon amb una posició social de classe mitjana, mitjana-alta, de persones formades, d'origen català, major índex d'envelliment, de medi urbà i amb menor activitat femenina.

L'estrat 3 també es caracteritza per una posició social de classe mitjana, mitjana-alta, formada i d'origen català, si bé es tracta d'una població més jove i on es dona la major taxa d'ocupació femenina; es correspon a una situació mitjana segons el grau d'urbanització.

⁷Insistim en que es tracta d'una identificació bàsicament descriptiva que no incorpora consideracions de naturalesa teòrica ni es vincula amb dades externes a les aquí tractades. De la mateixa forma no s'ha realitzat una anàlisi comparativa de desigualtats i disparitats entre els estrats si bé s'apunta en certa mesura en la pròpia descripció.



Gràfic 3. Representació dels estrats a les components 1 i 2



Gràfic 4. Representació dels estrats a les components 3 i 4

L'estrat 1 recull seccions censals amb posicions socials de classe mitjana-baixa i amb un nivell educatiu mitjà-baix. És població d'origen català, d'edats intermèdies o joves, localitzada als municipis més petits amb una presència important de l'activitat agropecuària.

L'estrat 4 coincideix amb l'anterior en el predomini de classe mitjana-baixa i nivells educatius mitjans-baixos, així com per la localització en petits nuclis de població. És

una població ocupada, d'edats intermèdies sobre tot, on l'origen català és més accentuat i on també s'intensifica l'activitat al sector agropecuari.

L'estrat 6 presenta nivells professionals i educatius intermedis, en canvi l'atur és dels més alts. L'origen immigrant es combina amb el català, essent un dels grups amb una població més vella, ubicada en nuclis urbans.

L'estrat 2 agrupa a la població de categoria professional baixa i nivells formatius baixos, amb xifres d'atur per sobre de la mitjana. Es tracta de les seccions amb major proporció de població jove, amb alts nivells de persones d'origen immigrant, ubicades a municipis de diferents grandàries.

L'estrat 5 finalment reproduïx, accentuant-la, la imatge de l'estrat anterior: major proporció de categories professionals baixes, mínims nivells educatius, màximes taxes d'atur i amb major presència de persones d'origen immigrant. En aquest cas es tracta de població menys jove i que resideix a municipis de més grandària.

A continuació l'exposició seguirà amb les dues etapes finals del disseny de mostra proposat: el càlcul de la grandària de la mostra i l'afixació, i la ponderació a posteriori de la mostra.

2.4. Grandària mostral, afixació de la mostra i ponderació a posteriori

Una vegada determinats els estrats amb la caracterització socioeconòmica obtinguda i com a expressió de conjunts homogenis de seccions censals, correspon determinar la distribució dels individus de la mostra en aquests estrats a partir de la determinació d'un nombre total. El càlcul de la grandària mostral s'inscriu en la fixació dels paràmetres bàsics: grandària de la població, estimacions de la mitjana i de la variabilitat, nivell de significació i error mostral.

Com a mesura de variabilitat es considera la distància quadràtica euclidiana de cada secció al centre global del núvol de punts que forma l'espai vectorial dimensionalitzat i reduït, i com a paràmetre la mitjana d'aquesta distància. D'aquesta manera s'aconsegueix estimar i reflectir a la mostra, no tan sols una característica determinada d'interès de l'estudi, sinó tot un conjunt, atès que es pren un punt mitjà i una desviació d'aquest conjunt de característiques dimensionalitzades a partir de les dades censals.

Si considerem un nivell de significació de 2σ i un error mostral relatiu de l'1,76 % per a dades globals de la RMB, la grandària mostral n surt d'aplicar la fórmula següent:

$$n = \frac{z^2 \cdot \sigma_y^2}{e^2 \cdot \bar{Y}^2}$$

- on:
- z és el nombre de sigmes de nivell de significació.
 - σ_y^2 és la variància de la distància quadràtica euclidiana de les seccions censals al centre de la totalitat del núvol (valor obtingut de 0,2794).
 - e és l'error mostral relatiu.
 - \bar{Y} és la mitjana de les distàncies quadràtiques euclidianes (valor obtingut de 1,0654).

i on s'obté un total de 5200 individus.

Aquest nombre d'individus va ser distribuït entre els estrats amb el criteri d'afixació òptima de Neyman. Amb aquest criteri s'opera l'efecte de l'estratificació de la mostra segons el qual quan més gran i variable és un estrat major proporció de mostra se l'assigna. Per tant no es tracta d'una distribució estrictament proporcional a la població de cada estrat, sinó que a efectes d'optimització, de guany en la precisió de les estimacions, s'adopta aquest doble criteri que s'expressa a la fórmula:

$$n_h = \frac{N_h \cdot \sigma_h}{\sum_{h=1}^K N_h \cdot \sigma_h} \cdot n$$

- on:
- n_h és la grandària mostral de l'estrat h ($h = 1 \dots 8$)
 - N_h és la població major de 15 anys de l'estrat h
 - σ_h és la desviació de la distància quadràtica euclidiana de les seccions censals de l'estrat h al centre del seu estrat.

Els valors que s'obtenen de l'afixació es presenten a la Taula 4.

Amb el nombre mostral de cada estrat es procedeix a l'assignació proporcional de quotes de mostra en termes d'individus corresponents a cada secció censal de l'estrat. Amb aquest repartiment es garanteix l'acompliment de l'aleatorietat de la mostra en l'elecció d'un individu que pertany a una secció determinada. L'assignació de quotes es fa segons la fórmula:

$$n_{sh} = \frac{N_{sh}}{N_h} \cdot n_h$$

- on:
- n_{sh} és la quota de mostra de la secció s de l'estrat h
 - N_{sh} és la població major de 15 anys de la secció s l'estrat h
 - N_h és la població major de 15 anys de l'estrat h
 - n_h és la grandària mostral de l'estrat h

Taula 4. Valors d'afixació mostral òptima de Neyman per a cada estrat

Estrat	N_h	σ_h	$N_h \cdot \sigma_h$	Coefficient d'afixació	n_h
1	223.835	0,6025869	134.880,0	0,0820091	426,45
2	834.715	0,3801628	317.327,6	0,1929400	1.003,29
3	588.235	0,4386580	258.034,0	0,1568886	815,82
4	64.298	1,1315886	727.58,88	0,0442385	230,04
5	627.659	0,4814569	302.190,7	0,1837366	955,43
6	323.600	0,5777149	186.948,5	0,1136676	591,07
7	519.470	0,4496771	233.593,8	0,1420286	738,55
8	331.165	0,4196152	138.961,9	0,0844909	439,35
Total	3.512.977	0,5285464	1.644.695,4	1,0000000	5.200,00

L'assignació que s'obté dona quotes de mostra no enteres per a cada secció censal, per la qual cosa, a efectes d'elecció dels individus, es computa l'arrodoniment generant una grandària mostral final de 5263 enquestats/des que varen ser extrets aleatòriament del marc mostral que es deriva del Cens Electoral corresponent a l'any 1994.

Cal assenyalar per tant que aquest disseny mostral estratificat suposa l'afixació no proporcional de les quotes de mostra de cada estrat. Segons la grandària poblacional de l'estrat i la variabilitat de les característiques socioeconòmiques pròpies d'aquest, hi haurà individus majors de 18 anys que tindran una probabilitat major de ser elegits a partir de la quota que s'assigna a l'estrat on s'ubica la secció censal a la qual pertany, és a dir, no es garanteix el criteri d'equiprobabilitat quan un individu és elegit a l'atzar. Aquest criteri té un sentit instrumental doncs ens assegura la presència a la mostra d'aquelles característiques menys freqüents a la població, però, alhora, sobredimensiona la presència dels individus que les posseeixen. En conseqüència, una vegada obtinguda la mostra, es procedeix a la restitució del valor real de les freqüències ponderant el seu pes en el conjunt i així garantir una mostra estrictament aleatòria.

Aquesta ponderació és una magnitud que transforma la probabilitat real que un individu hagi estat escollit en la probabilitat teòrica sota hipòtesi d'estricta aleatorietat, i que es pot expressar amb la relació següent:

$$PES = \frac{\text{Probabilitat Teòrica}}{\text{Probabilitat Real}} = \frac{N_h/N}{n_h/n}$$

És a dir, s'atorga un menor pes a aquells individus que tenen una probabilitat major de ser elegits, i un major pes a aquells altres amb una probabilitat menor.

3. COMENTARIS FINALS

El procediment seguit per a la construcció de la mostra estratificada de l'Enquesta Metropolitana de Barcelona i els criteris i decisions que s'utilitzen van més enllà dels resultats habituals esperats del disseny mostral, ja que el procés de construcció proporciona conclusions d'interès per a l'anàlisi sociològica o pel coneixement social del territori, el que pot ser també de gran ajut en la tasca de planificació de l'Administració. Al mateix temps l'anàlisi realitzada serveix de criteri de control i validació de les conclusions posteriors a la pròpia anàlisi de les dades obtingudes a l'enquesta.

En aquest sentit i com a resultat d'una anàlisi específica que es pot efectuar amb posterioritat a la recollida d'informació, és possible reutilitzar la mostra amb l'objectiu de construir el que hem anomenat com a *zones socials* (Lozares i Domínguez, 1993, 1996), i considerar les necessitats presents habitualment en els estudis que plantegen el tractament de submostres per àmbits territorials més restringits amb errors mostrals que siguin acceptables donades les limitacions econòmiques. El que aquí proposem és que cada estrat de la mostra estratificada pot constituir una base mostral per a dur a terme anàlisis amb totes les garanties de precisió, com a mínim calculables i acceptables, sense augmentar la grandària de la mostra per a l'enquesta general.

El procediment consisteix en considerar els estrats homogenis generats en la construcció de la mostra com a zones socials, geogràficament localitzables i sense restricció de contigüitat territorial, als quals se'ls hi apliquen les dades obtingudes a l'enquesta. Per tant, les zones socials són caracteritzades separadament per la informació de l'enquesta amb un error mostral acceptable que en molts casos garanteix la representativitat. D'altra banda en aquestes zones socials es poden realitzar les anàlisis que es considerin adients, en particular de dimensionalització i estructuració de grups socials a l'interior de les zones on es poden observar noves diversitats i perfils. Posteriorment, la comparació de les diferents zones genera models generals de composició de grups i la idea d'una certa dinàmica social.

BIBLIOGRAFIA

- Cochran, W.G. (1971). *Técnicas de muestreo*. México: CECOSA.
- Grosbas, J.M. (1987). *Méthodes statistiques des sondages*. Paris: Economica.
- Lébart, L.; Morineau, A.; Fénélon, J.P. (1985). *Tratamiento estadístico de datos*. Barcelona: Marcombo.
- López Roldán, P. (1996). «La construcción de tipologías: metodología de análisis». *Papers. Revista de Sociologia*, 48, 9-29.
- Lozares, C.; López, P. (1990). *Enquesta Metropolitana de la Regió Metropolitana de*

- Barcelona. Construcció de la mostra estratificada. Sèrie Documents de Treball 90/1. Bellaterra: Institut d'Estudis Metropolitans de Barcelona.*
- Lozares, C.; López, P. (1991). «El análisis de componentes principales. Aplicación al análisis de datos secundarios». *Papers. Revista de Sociologia*, 37, 31-63.
- Lozares, C.; López, P. (1991). «El muestreo estratificado por análisis multivariado». En: *El pluralismo metodológico en la investigación social: ensayos típicos*, editado por M. Latiesa. Granada: Universidad de Granada, 107-160.
- Lozares, C.; Domínguez, M. (1993). *Enquesta de la Regió Metropolitana de Barcelona 1990. Territori i realitat social: les zones socio-demogràfiques de la Regió Metropolitana de Barcelona*. Barcelona: Mancomunitat de Municipis de l'Àrea Metropolitana de Barcelona y Diputació de Barcelona.
- Lozares, C.; Domínguez, M. (1996). «Tratamiento multivariado de subpoblaciones en una gran encuesta social: la construcción de zonas sociales». *Papers. Revista de Sociologia*, 48, 71-87.
- Nel-lo, O.; Recio, A.; Solsona, M.; Subirats, M. (1998). *Enquesta Metropolitana de Barcelona. Condicions de Vida i Hàbits de la població. La transformació de la societat metropolitana*. Barcelona: Mancomunitat de Municipis de l'Àrea Metropolitana de Barcelona i Diputació de Barcelona.
- Sánchez Carrión, Juan Javier (Ed.) (1984). *Introducción a las técnicas de análisis multivariante aplicadas a las ciencias sociales*. Madrid: Centro de Investigaciones Sociológicas.

ANNEX

Taula A1. Matriu de correlacions de les variables emprades a l'ACP

	P1	P2	P4	P7	P8	P9	P11	P12	P14	P15	P16	P17	P18	P19	P20	P23
P1	1,000	-,791	,248	,250	-,276	-,178	,102	,108	,430	-,316	,472	-,241	-,323	,109	,553	,165
P2	-,791	1,000	-,417	-,271	,258	,161	-,092	-,145	-,405	,371	-,517	,290	,336	-,030	-,495	-,091
P4	,248	-,417	1,000	,571	-,705	-,597	,520	,373	-,206	-,719	,756	-,128	-,622	-,211	-,074	-,222
P7	,250	-,271	,571	1,000	-,607	-,660	,530	,506	-,183	-,557	,665	-,133	-,577	,056	,161	-,023
P8	-,276	,258	-,705	-,607	1,000	,845	-,647	-,427	,335	,955	-,884	-,062	,767	-,151	-,034	-,107
P9	-,178	,161	-,597	-,660	,845	1,000	-,672	-,520	,281	,801	-,781	-,045	,724	-,170	-,078	-,138
P11	,102	-,092	,520	,530	-,647	-,672	1,000	,441	-,155	-,626	,566	,137	-,526	-,107	-,091	-,166
P12	,108	-,145	,373	,506	-,427	-,520	,441	1,000	-,234	-,406	,413	,024	-,401	-,040	-,035	-,101
P14	,430	-,405	-,206	-,183	,335	,281	-,155	-,234	1,000	,267	-,189	-,158	,245	-,096	,269	-,016
P15	-,316	,371	-,719	-,557	,955	,801	-,626	-,406	,267	1,000	-,888	-,073	,739	-,125	-,045	-,091
P16	,472	-,517	,756	,665	-,884	-,781	,566	,413	-,189	-,888	1,000	-,214	-,818	,098	,280	,122
P17	-,241	,290	-,128	-,133	-,062	-,045	,137	,024	-,158	-,073	-,214	1,000	,029	-,039	-,327	-,085
P18	-,323	,336	-,622	-,577	,767	,724	-,526	-,401	,245	,739	-,818	,029	1,000	-,169	-,204	-,146
P19	,109	-,030	-,211	,056	-,151	-,170	-,107	-,040	-,096	-,125	,098	-,039	-,169	1,000	,341	,667
P20	,553	-,495	-,074	,161	-,034	-,078	-,091	-,035	,269	-,045	,280	-,327	-,204	,341	1,000	,458
P23	,165	-,091	-,222	-,023	-,107	-,138	-,166	-,101	-,016	-,091	,122	-,085	-,146	,667	,458	1,000

Determinant de la Matriu de Correlacions = 0,0000004

Mesura d'Adequació Mostrat de Kaiser-Meyer-Olkin = 0,83302

Test d'Esfericitat de Bartlett = 52574,674 (Significació = 0,00000)

Taula A2. Matriu factorial o de saturacions de l'ACP

	Component			
	1	2	3	4
P8 Titulats mitjans-superiors majors 20 anys	,935	-,032	-,134	,121
P15 Professions altes	,911	-,128	-,101	,155
P16 Professions baixes	-,904	,268	,117	,164
P9 Escolarització 14-24 anys	,891	,044	-,166	,072
P18 Terciari alt/finances	,828	-,125	-,190	-,019
P4 Immigració fora Catalunya	-,798	,133	-,306	,148
P7 Analfabets majors de 10 anys	-,749	,045	-,010	,252
P11 Aturats abans ocupats	-,742	-,011	-,192	-,230
P12 Atur busca primera feina	-,592	-,101	-,126	,103
P1 Joves de menys de 15 anys	-,248	,863	,133	,092
P2 Vells majors de 65 anys	,293	-,844	-,013	-,186
P14 Dones actives majors 15 anys	,371	,747	-,127	-,073
P20 Vehicle privat treball	-,023	,569	,525	,320
P23 Població Secció/Municipi	,008	,081	,896	,020
P19 Agropecuari	-,047	-,033	,880	-,015
P17 Terciari mitjà/comerç/hosteleria	-,009	-,192	-,036	-,915

Mètode de rotació: Normalització Varimax amb Kaiser

Taula A3. Descripció dels estrats

	Mitjanes de cada estrat								Total RMB
	Estrat 1	Estrat 2	Estrat 3	Estrat 4	Estrat 5	Estrat 6	Estrat 7	Estrat 8	
Joves de menys de 15 anys	19,61	21,18	19,71	17,22	16,02	12,31	11,56	14,23	16,24
Vells majors de 65 anys	11,70	8,25	10,25	14,72	12,03	20,57	20,69	16,35	14,23
Index d'envelliment > 65/ < 15	,63	,41	,57	,91	,84	1,87	1,99	1,31	1,10
Immigració fora Catalunya	25,66	41,84	30,29	16,34	46,53	32,17	28,17	20,18	33,59
Estrangers	1,62	1,31	1,95	1,37	1,12	2,53	1,92	3,60	1,91
Analfabets majors de 10 anys	1,87	3,11	,95	1,41	5,16	1,71	,91	,37	2,18
Titulats mitjans-superiors majors 20 anys	23,57	17,19	33,61	21,74	14,31	22,01	33,40	57,84	27,63
Escolarització 14-24 anys	53,91	50,66	65,79	49,94	44,93	52,15	64,62	79,15	57,96
Aturats abans ocupats	9,51	13,85	10,92	8,03	15,26	14,58	10,74	7,43	12,16
Atur busca primera feina	2,28	3,63	2,22	2,18	4,94	3,45	2,63	2,17	3,20
Atur total	11,80	17,48	13,14	10,21	20,20	18,03	13,37	9,60	15,36
Dones actives majors 15 anys	39,29	40,47	45,10	35,72	34,53	35,19	36,34	41,80	38,69
Professions altes	15,44	8,81	20,75	13,85	6,97	13,80	23,77	42,68	17,82
Professions baixes	43,77	54,58	29,94	41,06	55,36	32,94	25,30	10,59	37,28
Terciari mitjà/comerç/hosteleria	17,41	17,23	20,34	17,77	16,91	26,55	19,33	15,93	19,07
Terciari alt/finances	3,21	1,91	5,91	2,67	1,61	3,79	6,19	8,43	4,28
Agropecuari	4,72	,80	,34	12,59	,60	,57	,28	,15	,91
Vehicle privat treball	65,36	51,76	44,37	68,98	41,61	30,35	34,33	44,44	43,09
Vehicle privat estudi	6,70	2,21	2,91	9,47	1,15	1,12	1,78	4,49	2,51
Vehicle privat treball + estudi	31,91	21,04	20,70	36,18	15,19	11,69	14,47	22,29	18,49
Població Secció/Municipi	28,43	3,85	1,58	94,68	1,28	,83	,36	,33	4,54

ENGLISH SUMMARY

DESIGN AND CONSTRUCTION OF A STRATIFIED SAMPLE FROM CENSUS DATA*

P. LÓPEZ*

C. LOZARES*

Universitat Autònoma de Barcelona

M. DOMÍNGUEZ**

Universitat de Barcelona

The purpose of the article is to show the most important features of the design and construction process of the stratified sample from the «Enquesta metropolitana de Barcelona. Condicions de vida i hàbits de la població» (Metropolitan survey of Barcelona. Life conditions and habits of the population) 1995 edition. Persons are gathered in census sections according to the information from the population census and strata are constructed by means of a procedure in which techniques of multivariate analysis are implied: principal components factor analysis and cluster analysis. Finally, a sample size is determined and allocated according to Neyman's optimum criterium. The procedure followed and criteria and decisions used go beyond the usual results expected in any sample since it provides interesting conclusions for sociological analysis and territorial planning as well as it may be used as validation criteria for later conclusions in the survey's analysis.

Keywords: Sample surveys, data analysis, factor analysis and principal components, cluster analysis

AMS Classification (MSC 2000): 62D05, 62-07, 62H25, 62H30

* The three authors are members of the *Grup d'Estudis Sociològics sobre la Vida Quotidiana i el Treball* (QUIT) of the Departament de Sociologia of the Universitat Autònoma de Barcelona.

* Pedro López Roldán (Pedro.Lopez.Roldan@uab.es); Carlos Lozares Colina (Carlos.Lozares@uab.es). Departament de Sociologia. Universitat Autònoma de Barcelona.

** Màrius Domínguez Amorós (marius@riscd2.eco.ub.es). Departament de Sociologia. Universitat de Barcelona.

– Received September 1999.

– Accepted November 1999.

1. INTRODUCTION

The purpose of this article is to show the most relevant features regarding the design and construction process of the stratified sample which is the basis for information collection of the «Enquesta Metropolitana de Barcelona. Condicions de Vida i Hàbits de la Població» (EM, Metropolitan Survey of Barcelona. Life Conditions and Habits of the Population) 1995 edition, carried out by the «Institut d'Estudis Metropolitans de Barcelona» (Institute for Metropolitan Studies of Barcelona).

The EM is a permanent research which was designed in 1984 in order to analyse activities, life conditions and lifestyles of population in the Metropolitan Area of Barcelona, facing the lack of systematic and objective statistic data, of social character and about these contents. The EM has become a periodical tool for collecting information which helps us to specify the evolution and most structural tendency changes of all social phenomena which are analysed in the survey. Up to now, there have been three editions (1985, 1990, 1995) and the organisation of a fourth edition is being planned for 2000.

The EM collects information from the design of a stratified sample which has been kept through the three editions, however, the specific territorial scope has been modified. Next, the most relevant features of the sampling design and construction procedure are specified.

2. CONSTRUCTION OF THE SAMPLE FROM «METROPOLITAN SURVEY»

2.1. Sample design: population universe and criteria variables

Extracting a stratified random sample which was representative of the population in the Metropolitan Area of Barcelona was the objective established in the design of the sample. Choosing the stratified sample procedure is basically justified by precision criteria, compared to other methods, and by social heterogeneity which is a feature of the target population. Thus, strata construction enables us to have a stratification variable which will be a dense variable of social typification and which will guarantee the presence in the sample of homogeneous social types, which are characteristic of the population, types or social groups defining a basic variable which correlates with other targets in the EM.

The population universe is defined as the collection of persons above 18 in the RMB (Metropolitan Area of Barcelona), a territory including 162 municipalities and a total population of 3,275,458 people according to data from the Population Census in 1991. From the information available in the census survey, we have selected some indicators

which are the ones acting as variables/stratification criteria, and which will define, as a result of people addition, the census sections in the RMB. Thus, we have not considered individuals in the census as basic units to be stratified at first stage, but census sections where these individuals live. This procedure is justified by a double reason: because of difficulties deriving from dealing with such a big matrix of individuals; and because only addition in census sections enables us to use a metric percentage treatment of the selected variables. Consequently, we have a data matrix with 3,586 census sections in the RMB with the selection of the 16 indicators-variables of demographic characterising, cultural-educative, of labour and professional activity, of mobility and of population extension.

2.2. Strata construction

From this information we proceed to construct strata in a procedure in which techniques of multivariate analysis are involved: principal components factor analysis and cluster analysis. The stratification of the sample is based on the principle of considering the studied population without homogeneity as far as statistics is concerned. Statistic homogeneity refers to the relation established amongst different features in the same population, the more correlated these features are, the more homogeneous the population will be considered. The purpose of stratification will then be the classification of census sections in strata which, with the object to sampling, and improving in precision, will be the expression of as much as possible homogeneous groups of sections inside every stratum, and as much as possible heterogeneous amongst them according to variables/factor criteria deriving from the previous analysis.

This classification is made without considering any restriction in territorial continuity, so the result will be a map of RMB sections divided into a number of strata with different socio-economic features.

With the principal components factor analysis (ACP), we attempt to reduce original information in order to obtain a vector subspace where factors are bases and therefore lineally independent, and which, in a hierarchical order, keep most of the total variation. By this means, we obtain the fundamental dimensions of population difference in the RMB, which initially structure social reality according to the information introduced. So the ACP is thought as a previous stage and necessary complement of categorisation of units, of strata obtaining. In the analysis presented, we finally retain 4 components.

In the reduced matrix deriving from the ACP, from 4 factor scoring variables identifying the 3,586 census sections, we proceed to cluster analysis. The classification process must be divided into two different stages: in the first one, we apply a forward hierarchical clustering (ward's method), in which a first classification in 8 classes or strata is established; in a second stage, with the number of groups determined and initial centres

defined, a non-hierarchical classification operates out of addition around mobile centres in order to optimise strata assignments.⁸

2.3. Sample size, allocation and sample weighting

Once the strata have been defined as the expression of uniform sets of censal sections, the individuals of the sample are distributed among the strata on the basis of the pre-determined sample size. The sample size is calculated by determining the basic parameters: the size of the population, the estimations of mean and its variability, the significance level, and the sampling error.

We compute the Euclidean distance of each section to the global centroid of the cluster of points in space of the dimensionalised factors as a measure of variability, and the mean of this distance as the parameter. In this way, it was possible to estimate and reflect in the sample not only a specific characteristic of interest of study, but also a whole set of characteristics, given that a midpoint was taken as well as a deviation of this set of dimensionalised population characteristics.

If we take a significance level of 2 sigma and a relative sampling error of 1,76 %, the calculation of the sample size was carried out using the following formula:

$$n = \frac{z^2 \cdot \sigma_y^2}{e^2 \cdot \bar{Y}^2}$$

where: z Number of sigma of the significance level.

σ_y^2 Variance of the squared Euclidean distance from the censal sections to the centroid of the cluster as a whole (the resulting value was 0,2794).

e Relative sampling error.

\bar{Y} Mean of the squared Euclidean distances (the resulting value was 1,0654).

This resulted in a total of 5,200 individuals. Using this figure we distributed the individuals among the different strata. The homogeneity of these strata facilitate the proper allocation of the individuals who belong to the censal sections of a specific stratum, using Neyman's optimum allocation criterion. It was the use of this optimum allocation process that allowed us to obtain the real effect of the stratification of the sample. According to Neyman's optimum allocation criterion, the distribution of the 5,200 individuals

⁸Classification process also means different validation procedures: a systematic analysis of classification between 15 and 5 strata, comparison between results obtained in previous editions of the survey, comparison between ward classification procedure and other upward hierarchical methods (average linkage between groups, average linkage within groups, complete linkage, centroid clustering and median clustering), repetition of the cluster analysis from sub-samples as well as use of theoretic-interpretative criteria.

is not carried out in strict proportion of the population of each strata, but with a view to optimising the result, that is, in order to increase the accuracy of the estimations. To this end, a two-fold strategy is adopted in which the larger and the more variable the strata, the larger the proportion of the sample that is allocated to it. This is expressed in the formula:

$$n_h = \frac{N_h \cdot \sigma_h}{\sum_{h=1}^K N_h \cdot \sigma_h} \cdot n$$

where: n_h Sample size of the strata h ($h = 1 \dots 8$)

N_h Population over 15 in strata h

σ_h Standard deviation of the squared Euclidean distances from the censal sections of strata h to the centroid of its strata.

n Resulting sample size.

Once the sample size for each strata was obtained, the sample quotas were allocated to the different censal sections that make up each strata. This procedure ensures the randomness of the sample in the selection of an individual belonging to any given selection. Quotas were allocated according to the following formula:

$$n_{sh} = \frac{N_{sh}}{N_h} \cdot n_h$$

where: n_{sh} Sample quota of section s of the strata h .

N_{sh} Population over 15 in section s of strata h .

N_h Population over 15 strata h .

n_h Sample size of strata h .

The resulting allocation yielded sample quotas for each population section that were not integers, and consequently, for the purposes of selecting individuals, these figures were rounded off to the nearest integer. After rounded-off, the sample size was ultimately established at 5,263 individuals, who were selected at random from the 1994 Electoral List.

It should be noted that this sample design involved non-proportional allocation of the sample quotas of each strata. Depending on the population size of the strata, and the variability of the socio-economic characteristics of the strata, there will be individuals of over 18 who have a greater probability of being selected from the quota allocated to the strata containing the censal section to which that individual belongs, i. e. equal probability is not guaranteed when an individual is selected at random. This criteria make sense in this context because it ensure the presence in the sample of characteristics that are less common in the population, since these are the more variable phenomena. At the same time, however, this criteria gives rise to overrepresentation of individuals with such characteristics. Consequently, once we obtained the sample, we restored the

real values of the frequencies by weighting them according to the proportion of the overall sample they represented in order to guarantee a truly random sample.

This weighting is a figure that transforms the real probability that an individual has been selected into the theoretical probability governed by the hypothesis of strict randomness, and can be expressed as follows:

$$WEIGHT = \frac{\textit{Theoretical probability}}{\textit{Real probability}} = \frac{N_h/N}{n_h/n}$$

where: N_h Population over 18 in censal section h .

N_h Population over 18 in the Region.

n_h Sample quota of section h .

n Total sample.

3. FINAL COMMENTS

The procedure followed to construct the sample and criteria and decisions used go beyond the usual expected results in the sample design since it provides interesting conclusions for the sociological analysis or of social knowledge of the territory which might be of a great help in the planning task of the Administration. At the same time, it might be used as control and validation criteria for later conclusions in the survey's analysis.

On the other hand, and as a result of a specific analysis carried out after information collection, the sample has been used again with the purpose of constructing what we have called «social areas». Every stratum in the stratified sample may constitute a sampling base in order to carry out analysis with all precision guarantees, at least calculable and acceptable, without enlarging the extension of the sample for the general survey. Analysis of these areas with the survey's data also enables to obtain a social profile, a structural analysis of dimensionalisation inside the homogeneous stratum without corresponding with an administration unit.