QÜESTIIÓ, vol. 25, 3, p. 553-580, 2001

EL CENS DE POBLACIÓ: UN ASSAIG D'INTERPRETACIÓ MITJANÇANT DATA MINING

CRISTINA GUISANDE ALLENDE*
FRANCESC SUBIRADA CURCO**

En aquest article es presenten els resultats d'un treball fet conjuntament entre l'Institut d'Estadística de Catalunya (Idescat), el Centre de Supercomputació de Catalunya (Cesca) i IBM. Les tres institucions van realitzar una anàlisi de les dades del Cens de població i habitatge de 1991 segons la metodologia Data Mining. Del conjunt de tècniques que es poden aplicar amb l'Intelligent Miner, es van fer servir les tècniques de segmentació i d'associació que s'han aplicat als individus segons el tipus de llar al que pertanyen.

Population Census: an interpretation using Data Mining

Paraules clau: Data Mining, estadística oficial, associacions, segmentació, anàlisis de dades, aplicació, mètodes estadístics

Classificació AMS (MSC 2000): 62-07, 62P25, 62H30, 62H20

* Institut d'Estadística de Catalunya. Via Laietana, 58. 08003 Barcelona. E-mail: cguisande@idescat.es
** CEPBA-IBM Research Institute. Jordi Girona, 1-3. 08034 Barcelona. E-mail: frsubirada@es.ibm.com

– Rebut l'abril de 2001.
– Acceptat el novembre de 2001.

1. INTRODUCCIÓ

L'Institut d'Estadística de Catalunya (Idescat), el Centre de Supercomputació de Catalunya (Cesca) i IBM van signar un conveni de col·laboració per tal de realitzar una anàlisi de les dades del Cens de població i habitatge de 1991 segons la metodologia Data Mining.

L'Idescat va aportar l'arxiu estadístic del Cens de població i habitatge, IBM va contribuir amb la tecnologia de tractament de les dades i per la seva banda, Cesca va aportar l'equip informàtic adient per procedir a l'execució del projecte.

L'objectiu d'aquest article és presentar els principals resultats d'aquesta experiència on es va aplicar Data Mining a informació censal. L'article s'estructura de la següent manera: en primer lloc es realitza una breu descripció dels recursos informàtics utilitzats, seguidament es presenten les característiques de les dades censals i el seu tractament i, finalment, s'examinen els principals resultats; en un annex final es recullen els trets bàsics de les operacions de Mineria de Dades.

2. LA MINERIA DE DADES

Durant la col·laboració entre l'Institut d'Estadística de Catalunya, el Cesca i IBM es va utilitzar la següent infrastructura:

1. A nivell de Programari:

El producte IBM anomenat «DB2 Intelligent Miner for Data». Aquest és un producte que permet aplicar operacions de Mineria de Dades als registres per poder trobar en ells informació que prèviament es desconeix.

2. A nivell de Maquinari:

Un ordinador IBM de procés paral·lel anomenat «System Paralel 2» (SP2). Aquest ordinador permet la utilització conjunta de diferents processadors per resoldre un sol problema. En el cas concret de Mineria de Dades dóna la possibilitat de treballar amb grans bases de dades (per exemple el cens de població) en un temps raonable.

Respecte a les operacions que es poden realitzar aquestes són: segmentació, classificació, predicció i anàlisi de relacions. Una breu descripció de les mateixes es presenta en un annex al final d'aquest article. Per les característiques de les dades censals descrites més endavant solament es van poder aplicar dues de les operacions esmentades: la segmentació i dins de l'anàlisi de relacions, l'associació.

3. LES DADES CENSALS

Un cens de població es pot definir com el conjunt d'operacions de recollida, elaboració, valoració i anàlisi de les dades de caràcter demogràfic, cultural, econòmic i socials de tots els habitants d'un país amb referència a un moment o període determinat. Per aquest motiu, acostuma a ser una de les tasques fonamentals dels instituts d'estadística de tots els països. Aquestes dades es recullen cada deu anys aproximadament d'acord amb la normativa internacional de l'ONU i també de la Comunitat Europea.

A Catalunya l'últim cens es va realitzar prenent per referència les zero hores de l'1 de març de 1991. Comprèn les persones que tenien fixada la seva residència a Catalunya en el moment censal i aquelles que tot i no residir-hi, s'hi trobaven en la data de referència.

El cens de població és una operació quasi insubstituïble per al recompte de la població i disposa de la màxima protecció legal per tal de salvaguardar la confidencialitat de les respostes dels ciutadans al qüestionari censal. La legislació vigent protegeix i garanteix la confidencialitat de les dades censals individuals sotmeses a secret estadístic que no poden ser divulgades, fins i tot a altres administracions o organismes públics no sotmesos al compliment del secret estadístic. La protecció legal assegura el tractament operatiu de la informació mitjançant el qual aquesta no podrà ser tractada de forma personalitzada.

Cal recordar que el cens de població és la principal font de dades demogràfiques i és una eina molt valuosa per a la planificació social i econòmica. Recull informació de la totalitat dels habitants en una data concreta, es a dir, —les dades són de tipus transversal—, dóna com a resultat el que s'acostuma a descriure com «la fotografia» de la població en un moment determinat.

El qüestionari del Cens de 1991 va incloure 31 preguntes relatives a les característiques de les persones i 6 temes en relació a l'habitatge.

Els principals aspectes que investiga es refereixen a:

a) característiques geogràfiques: lloc de residència

b) característiques demogràfiques: sexe, edat, estat civil, lloc de naixement, nacionalitat, relació de parentiu, fecunditat, data del casament, migració, mobilitat.

c) característiques socials o culturals: nivell d'instrucció, estudis en curs, coneixement del català.

d) característiques econòmiques: relació amb l'activitat, sector d'activitat, situació professional, professió o ocupació.

e) característiques de l'habitatge: data de construcció, règim de tinença de l'habitatge, superfície, nombre d'habitacions, instal·lacions de l'habitatge.

4. TRACTAMENT DE LES DADES

La informació recollida en el Cens de població no és uniforme per a la totalitat de la població. Hi ha alguns temes que solament poden tenir resposta per part d'un sector de la població, així per exemple els relatius a l'activitat econòmica es refereixen a la població de 16 anys i més, el nivell d'instrucció a les persones de 10 anys i més, el coneixement del català a les de 2 anys i més, la fecunditat a les dones de 12 anys i més, etc. També hi ha altres que es refereixen al conjunt de la llar com ara les referides a les característiques de l'habitatge.

Per a l'aplicació de les tècniques de la mineria de dades, ha tingut lloc un procés d'anàlisi, de prova i selecció de les variables.

S'han descartat algunes variables perquè presenten un comportament molt semblant en el conjunt de la població, no discriminen, com són algunes de les referides a les instal·lacions de l'habitatge. En altres casos, a pesar del baix pes relatiu de determinades variables, aquestes constituïen un element important a l'hora de determinar grups específics (nínxols), per exemple les referides a la població estrangera que en aquest any representa l'1%.

Per a un millor tractament i posterior anàlisi algunes variables continues es van agrupar en intervals (edat, any d'arribada a Catalunya, data de construcció de l'habitatge) és a dir, es van transformar en discretes i altres en variables categòriques com és el canvi de municipi de residència en els últims deu anys, com indicador de migració en el decenni anterior al cens.

També es van combinar variables com per exemple, relació amb l'activitat amb situació professional i es van crear noves variables com el nombre de persones per llar. Finalment, les variables utilitzades, presentades segons els temes investigats pel cens, són les següents:

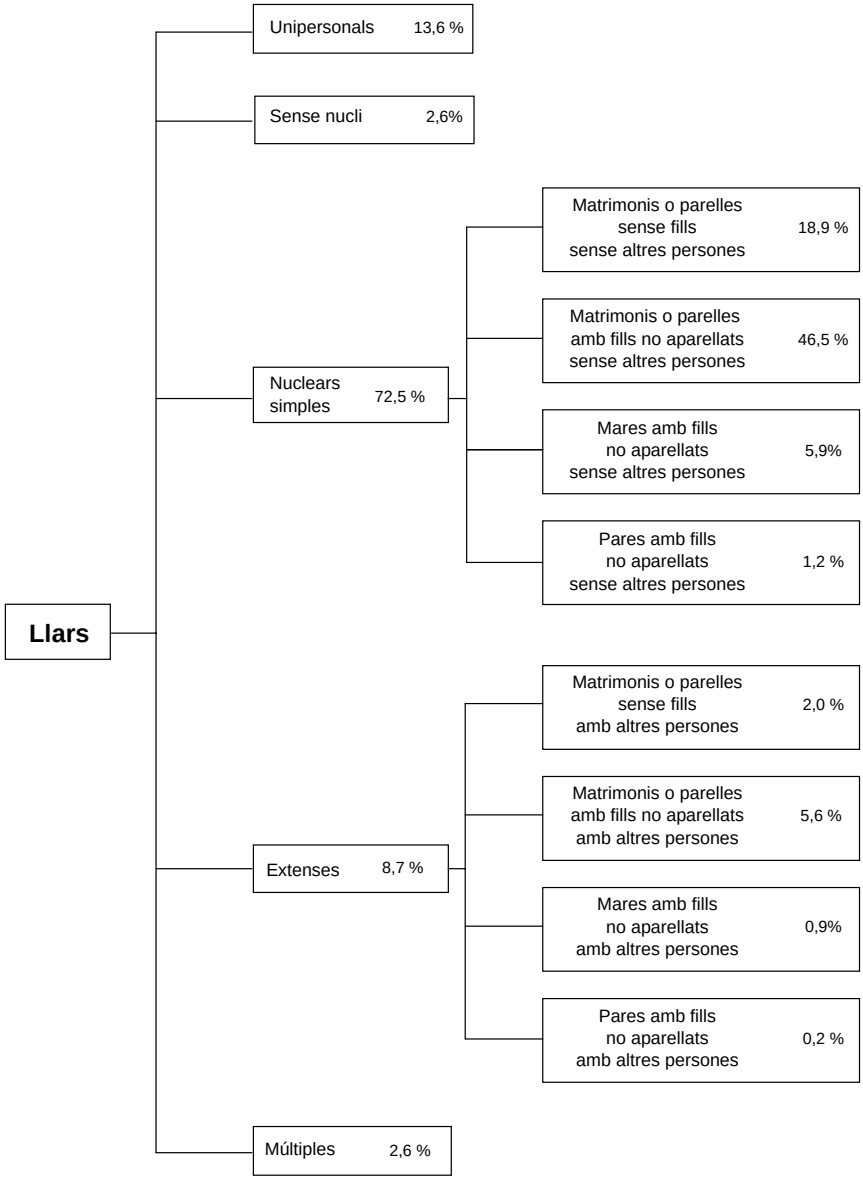
a) característiques geogràfiques i demogràfiques: lloc de naixement, edat, sexe, estat civil, nacionalitat, província de residència, canvi de residència als últims deu anys, any d'arribada a Catalunya, formes de convivència, nombre de persones per llar.

b) característiques socials o culturals: nivell d'instrucció, estudis en curs, coneixement del català.

c) característiques econòmiques: relació amb l'activitat, sector d'activitat, situació professional, professió o ocupació, mitjà de transport utilitzat per anar a treballar o estudiar.

d) característiques de l'habitatge: superfície de l'habitatge, nombre d'habitacions per llar, any de construcció de l'habitatge, règim de tinença de l'habitatge.

Quadre 1. Tipologia de llars.



Les variables referides a llars i famílies no es poden aplicar de manera directa com es recull al cens, per tant requereixen abans de ser utilitzades una ordenació, agrupació i classificació d'acord a una tipologia determinada. En aquest cas s'aplica una tipologia estàndard[1].

Les tècniques de segmentació i d'associació s'han aplicat als individus segons el tipus de llar al que pertanyen. L'interès per aquesta aproximació té la seva justificació en el fet de que molts dels comportaments demogràfics i una part important dels socials i econòmics estan sota la influència d'aquestes formes d'agrupació. D'altra banda, l'Institut d'Estadística de Catalunya, per primera vegada, ha obtingut informació detallada de les estructures familiars catalanes per a qualsevol àmbit territorial a partir de les dades del Cens de població i habitatge de 1991.

Per tal de situar en la seva perspectiva els resultats aquí presentats, convindrà fer un breu recordatori de l'estructura de les llars catalanes l'any 1991. El Cens de població va registrar aquest any 1.933.144 llars familiars on vivien 5.982.674[2] persones. Les característiques més notables d'aquestes llars (quadre 1) són, per una banda, l'elevada nuclearitat (73%) i, d'altra, un significat grau de complexitat familiar, reflectit en la presència de llars extenses i múltiples (11%). Tot i així, s'ha d'assenyalar que el pes de les llars unipersonals (14%), cada vegada més important, és un indici dels canvis que tenen lloc dins la composició familiar. El model dominant de família nuclear és el resultat d'un elevat percentatge de parelles amb fills (47%) i sense fills (19%).

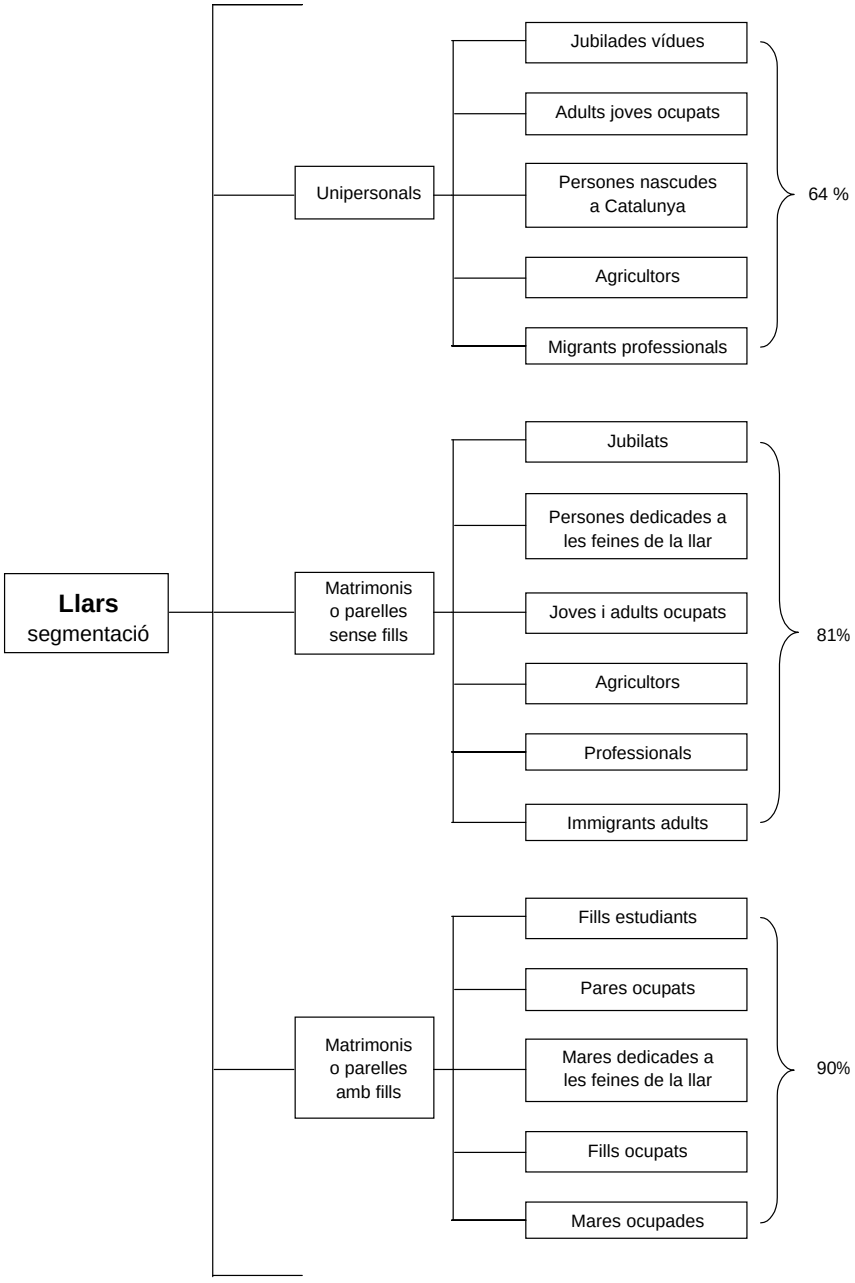
5. SEGMENTACIÓ: PRINCIPALS RESULTATS

Del conjunt d'estructures familiars es presentaran els resultats més rellevants obtinguts de l'aplicació de la tècnica de segmentació a les llars formades per una sola persona −les unipersonals−, i de les nuclears simples formades per matrimonis o parelles sense fills i matrimonis o parelles amb fills. Els tres tipus representen el 79% de les llars i concentren el 75% de la població de Catalunya a l'any 1991.

[1] La tipologia de llars que s'ha utilitzat té com a base, fonamentalment, la identificació i quantificació dels nuclis familiars i la presència o absència d'altres persones. Els nuclis poden estar composats per un matrimoni o parella sense fills, un matrimoni o parella amb fills, i nuclis formats per un sol progenitor, denominats també monoparentals. Així es distingeixen cinc tipus de llars: les unipersonals, formades per persones que viuen soles, les sense nucli, constituïdes per dues o més persones emparentades o no però que no formen un nucli familiar; les nuclears simples formades per nuclis en absència d'altres persones; les nuclears extenses compostes per nuclis amb la presència d'altres persones i per últim, les múltiples constituïdes per dos o més nuclis familiars.

[2] Aquesta dada no inclou les persones residents en llars de col·lectius.

Quadre 2. Resultats de la segmentació.



En una primera visió de conjunt (quadre 2) pot apreciar-se, pel que fa a les llars unipersonals, com els cinc segments reunits al quadre agrupen el 64% d'aquest tipus de llar; tot i que, com s'examinarà més endavant, el pes de cada segment és desigual amb una major freqüència de llars formades per dones jubilades vídues.

En el cas dels matrimonis sense fills, un nombre de segments molt proper a l'anterior −sis− agrupen un percentatge de llars més elevat, del 81%. També es constatarà la desigual distribució interna on els tres primers d'aquests segments comprenen la major part d'aquell valor, el 74%.

Finalment, respecte els matrimonis o parelles amb fills, un total de cinc segments tornen a reunir un percentatge molt elevat de llars d'aquest tipus, el 90%. S'aprecia, a l'igual que en el cas dels matrimonis sense fills, la presència d'un agrupament de població més homogeni, format ara per fills d'estudiants, pares ocupats, mares dedicades a les feines de la llar, fills ocupats i mares ocupades.

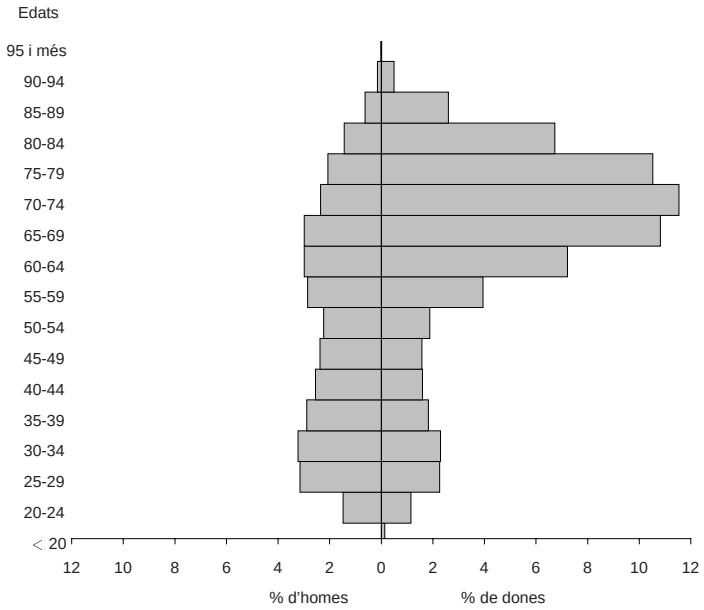
5.1. Llars unipersonals

L'estructura per edat i sexe de la població que viu a les llars unipersonals presenta una distribució dels seus efectius molt irregular (gràfic 1). Les diferències per edat i sexe són el resultat de diferents factors, entre els quals es destaquen, les distintes edats en què homes i dones comencen una relació de parella, l'efecte de les separacions i divorcis i les diferències de mortalitat entre homes i dones.

En aquesta estructura queden representats, com a mínim, tres moments del cicle de vida de les famílies. La solitud a edats joves, quan generalment aquesta és una elecció i en molts casos no resultarà definitiva, marca el moment de l'emancipació familiar. A les edats adultes, especialment pels homes, és el resultat d'una ruptura familiar i a les edats més elevades quan és fruit de la imposició vital, que es manifesta en la contracció del nucli familiar per mort d'algun del dos cònjuges, de manera més freqüents, dels marits.

Aquesta composició per edats condicionarà de manera significativa la determinació dels segments. A partir de l'aplicació d'aquest tipus de tècnica es constata (quadre 3) que la piràmide d'edats ha quedat «fragmentada» en diferents grups, el primer dels quals representa el 43% de la població, mentre que el següent més allunyat del primer agrupa a un 12% i els restants són petits segments amb reduït pes dins del conjunt de llars.

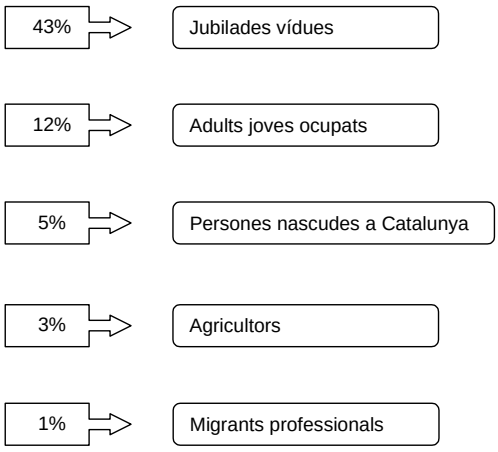
El primer segment representa fonamentalment la fase de contracció del nucli familiar. Es composa de dones vídues pensionistes que majoritàriament no tenen estudis o el nivell d'instrucció correspon a estudis primaris. El coneixement del català guarda estreta



Font: Institut d'Estadística de Catalunya (1994). *Cens de població 1991. Vol. 17.*

Gràfic 1. Estructura de la població de les llars unipersonals. Catalunya 1991.

Quadre 3. Llars unipersonals.



relació amb la naturalesa de la població. Aproximadament la meitat van néixer a Catalunya i les immigrants van arribar a aquesta Comunitat Autònoma abans del 1960 o entre 1960 i 1970. En conjunt la població d'aquest segment, no ha canviat de municipi de residència als últims deu anys. Aquestes dones viuen en habitatges de propietat, de construcció antiga, i resideixen principalment a la província de Barcelona.

El segon grup representa el 12% de les llars unipersonals. Està format de manera equilibrada per homes i dones, amb edats compreses entre els 30 i 50 anys, predominen les persones solteres, i en segon lloc els separats o divorciats. Presenten un nivell d'instrucció alt (secundaris i titulació superior), treballen al sector serveis, com professionals o tècnics i són assalariats amb caràcter fix. Van néixer a Catalunya, resideixen principalment a la província de Barcelona i viuen en habitatges de lloguer. El 10% ha canviat de municipi de residència els últims deu anys.

El tercer segment, amb un 5% del total de llars unipersonals es composa per homes i dones que van néixer a Catalunya. L'estat civil predominant és el de solter o vidu, fonamentalment són pensionistes, amb estudis primaris o secundaris, amb alt coneixement de la llengua catalana, i resideixen a la província de Barcelona en habitatges de lloguer.

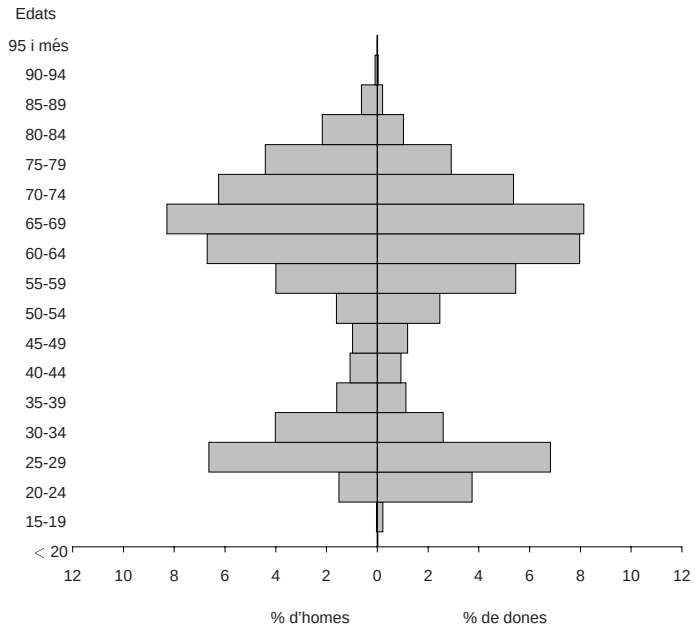
Dels restants segments, s'han de destacar al menys dos: el primer representa als agricultors, que és un grup que tornarà a aparèixer en els següents tipus de llars i el segon, els migrants professionals.

El segment dels agricultors es composa per un 3% de les persones de les llars unipersonals. Són homes i dones de 65 anys i més, vidus o solters que van néixer a Tarragona, Lleida i Girona, tenen estudis primaris, declaren un coneixement del català alt, resideixen en habitatges de propietat i no han canviat de municipi als últims deu anys.

El segment que representa als migrants professionals, constitueixen l'1% del total de llars unipersonals. Són homes solters, amb titulació superior i alta mobilitat als últims deu anys. Tenen edats entre els 30 i 50 anys, estan ocupats, són professionals i tècnics, el 71% van néixer a la resta d'Espanya i el 3% són de nacionalitat estrangera. Resideixen principalment a la província de Barcelona, viuen en habitatges de lloguer, han arribat a Catalunya en l'última dècada i el seu coneixement de català és alt.

5.2. Matrimonis o parelles sense fills

Mostren una estructura per edats molt envellida com a conseqüència de la falta d'infants. La piràmide (gràfic 2) presenta dos sortints molt marcats: un pels voltants dels 25-29 anys i un altre entre els 65-69 anys. Aquests valors modals representen dos moments del cicle de vida familiar. El primer, correspon a les edats joves i assenyalaria la fase de



Font: Institut d'Estadística de Catalunya (1994). *Cens de població 1991*.

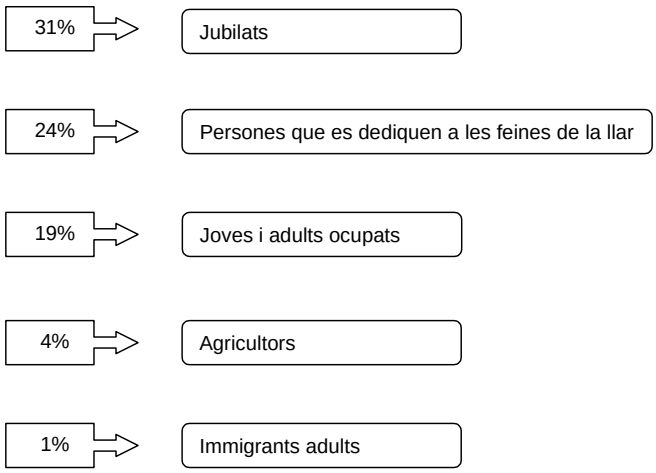
Gràfic 2. Estructura de la població de les llars de parelles sense fills. Catalunya 1991.

formació de la família i el segon, a les edats grans, la fase de contracció, és a dir, quan els fills han marxat de la casa dels pares. La variació d'aquests grups estarà en funció de diferents factors entre els quals s'han de destacar el calendari de la nupcialitat (o de la formació de la parella), de la fecunditat, de l'edat d'emancipació dels fills i del descens de la mortalitat.

Dels resultats obtinguts s'ha de destacar que els tres primers segments representen el 74% del total de la població dels matrimonis o parelles sense fills. El primer el composen els homes jubilats, el segon les dones dedicades a les feines de la llar i el tercer, les persones joves i adultes ocupades. Entre els restants segments, de menor importància relativa destaquen els agricultors, els professionals i els immigrants adults (quadre 4).

El primer dels segments esmentats inclou el 31% de la població d'aquest tipus de llar. Conté una alta representació d'homes casats, pensionistes, per tant persones de 65 anys i més, el 54% dels quals van néixer a Catalunya. Els que treballen o han treballat anteriorment, més de la meitat ho fan com treballadors industrials. El nivell d'estudis és baix, no han migrat als últims deu anys, viuen en habitatges de propietat, i resideixen majoritàriament a la província de Barcelona.

Quadre 4. Matrimonis o parelles sense fills.



El segon segment mencionat representa el 24% de la població i està compost fonamentalment per dones casades que treballen en les feines de la llar. Tenen edats de 65 anys i més o entre 51 i 64 anys, predominen les persones sense estudis o estudis primaris amb baix coneixement de la llengua catalana. No han canviat de municipi de residència els últims deu anys i majoritàriament viuen a la província de Barcelona en habitatges de propietat.

El tercer dels principals segments, amb un 19% de la població, es composa bàsicament per joves de menys de 30 anys, i persones adultes-joves d'entre 31 i 50 anys, dels quals més de la meitat són homes. Van néixer a la província de Barcelona, declaren un nivell molt alt de català i presenten una titulació més elevada que els casos anteriors: secundaris i superiors. Majoritàriament estan ocupats al sector serveis, es desplacen a treballar o a estudiar en cotxe o taxi i la seva situació professional correspon a una persona que treballa amb un contracte fix. No han deixat mai de residir a Catalunya i viuen a la província de Barcelona en habitatges de compra, que han estat construïts entre 1960-1986. Un elevat percentatge han canviat de municipi als últims deu anys i hi ha un 8% de persones solteres que formen parelles de fet.

En aquest tipus de llar, també destaquen com segments diferenciats els corresponents als agricultors, als professionals i als immigrants adults.

Els agricultors representen el 4% de la població. La variable que els determina en primer lloc és el sector d'activitat i la professió o ocupació. Treballen en l'agricultura i la seva situació professional és d'empresaris o treballadors per comte propi que no contracten personal. Són majoritàriament homes, casats, el 80% són persones majors de 65 anys, que van néixer a Catalunya, actualment viuen a Tarragona, Lleida i Girona,

en habitatges de propietat. Predominen les persones amb estudis primaris encara que és significatiu el percentatge de les sense estudis, el coneixement de català és molt alt i no han canviat de municipi als últims deu anys.

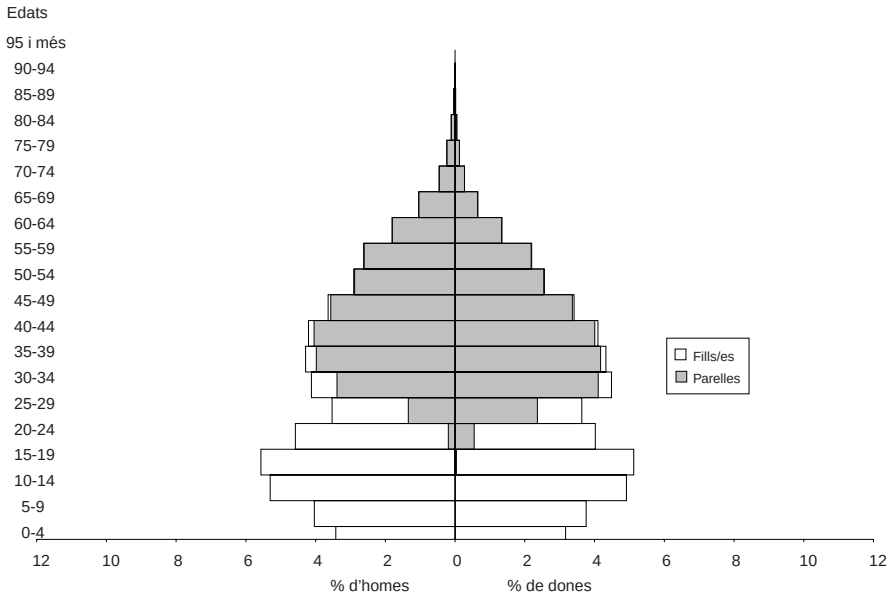
El segment que representa als professionals, inclou al 2% de la població. Ho formen adults joves (entre 30 i 50), dels quals més de la meitat són dones. El 9% de les persones són estrangeres i una alta proporció són parelles de fet (27%). El 92% tenen estudis superiors, són professionals o tècnics que treballen al sector serveis amb caràcter fix i utilitzen el cotxe o taxi per anar a treballar. La meitat han canviat de municipi de residència els últims deu anys, han viscut fora de Catalunya, tenen un coneixement molt alt de català i aproximadament la meitat van néixer a aquesta Comunitat Autònoma. Viuen en la província de Barcelona, en habitatges de recent construcció, de lloguer o en propietat amb pagament pendent amb una superfície d'entre 90 i 110 metres quadrats.

Finalment els immigrants adults representen l'1% del conjunt de matrimonis o parelles sense fills. Presenten una alta mobilitat, ja que el 72% han canviat de municipi als últims deu anys. Predominen els homes, pensionistes amb un nivell d'instrucció correspon a titulació secundària i superior. Aproximadament, la meitat del segment el composen persones de nacionalitat estrangera, dels quals més de la quinta part són de la Unió Europea. Han arribat a Catalunya entre 1981 i 1991, resideixen fonamentalment a la província de Girona i viuen en habitatge de propietat. Les persones que treballen ho fan al sector serveis, amb caràcter fix.

5.3. Matrimonis o parelles amb fills

Representen el grup més nombrós de tota la població amb un 59% dels seus habitants. La forma de la piràmide (gràfic 3) guarda relació amb el tipus d'estructura de la llar, és a dir, el fet que siguin parelles que conviuen amb els seus fills fa que el pes de les persones grans resulti molt baix. S'aprecia, també, una reducció de la piràmide al voltant dels 30 anys, edat que correspon a l'emancipació dels fills, o a la formació de noves parelles. De fet la piràmide pot dividir-se en dues parts: una primera comprèn des de la base fins als 25 anys i són majoria els fills i la segona, a partir d'aquesta edat, ocupada fonamentalment pels pares.

Aquest tipus de llars és el més homogeni dels investigats, ja que cinc segments reuneixen el 90% de la població, ells són: els fills estudiants, els pares ocupats, les mares dedicades a les feines de la llar, els fills ocupats i les mares ocupades (quadre 5).



Font: Institut d'Estadística de Catalunya (1994). *Cens de població 1991*.

Gràfic 3. Estructura de la població de les llars de parelles amb fills. Catalunya 1991.

Quadre 5. Matrimonis o parelles amb fills.

35%	Fills estudiants
21%	Pares ocupats
15%	Mares dedicades a les feines de la llar
11%	Fills ocupats
8%	Mares ocupades

El primer segment, representa el 35% de la població que viu a les llars de matrimonis o parelles amb fills i està compost pels fills estudiants. La variable més significativa en la determinació del segment és l'edat: el 70% són menors de 16 anys i el 30% restant són joves d'edats compreses entre els 16 i 30 anys. Són fills, homes i dones, estudiants, solters, que es caracteritzen en què majoritàriament van néixer a Catalunya (76% a la província de Barcelona), no han migrat d'aquesta Comunitat Autònoma, viuen en llars de 4 o més persones, presenten un coneixement de català molt alt, viuen en habitatges en propietat, són de nacionalitat espanyola, el 75% resideixen a la província de Barcelona.

S'ha de destacar que si bé és un segment representatiu dels infants i adolescents, hi ha un 30% de joves entre 16 i 30 anys que tenen les mateixes característiques que els infants. També s'ha d'indicar que és l'únic segment on el sexe és la variable menys important en la determinació del mateix.

El segon segment correspon als pares ocupats, amb un 21% de la població d'aquest tipus de llar. Són homes, pares, majoritàriament d'edats compreses entre els 31 i 50 anys, casats, ocupats, que treballen amb caràcter fix, al sector d'indústria-energia, com treballadors industrials. Aproximadament la meitat van néixer i no han deixat de residir a Catalunya, sols el 8% van canviar de municipi als últims deu anys. El nivell d'instrucció predominant és el d'estudis primaris, seguit pel secundari, el coneixement de català és alt, viuen principalment en llars de 4 persones, en habitatges de propietat, construïdes entre 1960-1986 i resideixen a la província de Barcelona.

El tercer segment, el formen les mares dedicades a les feines de la llar amb un 15% de la població de les llars de matrimonis o parelles amb fills. Són dones casades, principalment d'edats compreses entre els 31 i 50 anys i en menor proporció d'entre 51 i 64 anys. Treballen a la llar en feines no remunerades, presenten un nivell d'estudis primaris o sense estudis i un alt coneixement de la llengua catalana. Més de la meitat d'aquestes persones no van néixer a Catalunya i les que són immigrants van arribar a aquesta Comunitat Autònoma entre 1961-1970. Viuen en llars de 4 o més persones, en habitatges de propietat, construïdes entre 1960-1986, a la província de Barcelona. Sols un 10% han canviat de municipi de residència els últims deu anys.

El quart segment, amb un 11% de la població de les llars de matrimonis o parelles amb fills representa als fills ocupats. Aquests fills, tenen edats compreses entre els 16 i 30 anys, l'estat civil és el de solters i el 60% són homes. Respecte a l'activitat econòmica, són assalariats que treballen amb caràcter eventual (52%) o fix (39%), principalment al sector serveis. El nivell d'estudi predominant és el de secundari, declaren un coneixement molt alt de català i un 16% estan cursant estudis. Fonamentalment són catalans que no han canviat de municipi de residència en els últims deu anys, viuen en llars de 4 o més persones, en habitatges de propietat i el 82% resideixen a la província de Barcelona.

El cinquè segment, està integrat per mares ocupades i representen el 8% de la població de les llars de matrimonis o parelles amb fills. Són dones, mares, casades, de 31 a 50 anys, que estan ocupades, treballen al sector serveis, com assalariades amb caràcter fix, la distribució de les professions és molt homogènia encara que predominen les administratives. El nivell d'estudis amb major pes és el secundari, el coneixement de català és alt, viuen en llars de 3-4 persones. El 63% van néixer a Catalunya, la mobilitat dels últims deu anys és baixa. Resideixen a la província de Barcelona en habitatges de propietat.

6. ASSOCIACIONS: PRINCIPALS RESULTATS

Aquesta tècnica permet conèixer l'associació entre diferents elements. D'una banda, s'obté la intensitat de la relació i d'altra, la freqüència amb que es detecta aquesta regla d'associació en el conjunt que s'analitza. En el cas de les dades censals, la recerca d'associacions s'ha de fer en cada llar per cada variable per separat. Això permet obtenir les associacions entre les diferents categories de la variable escollida.

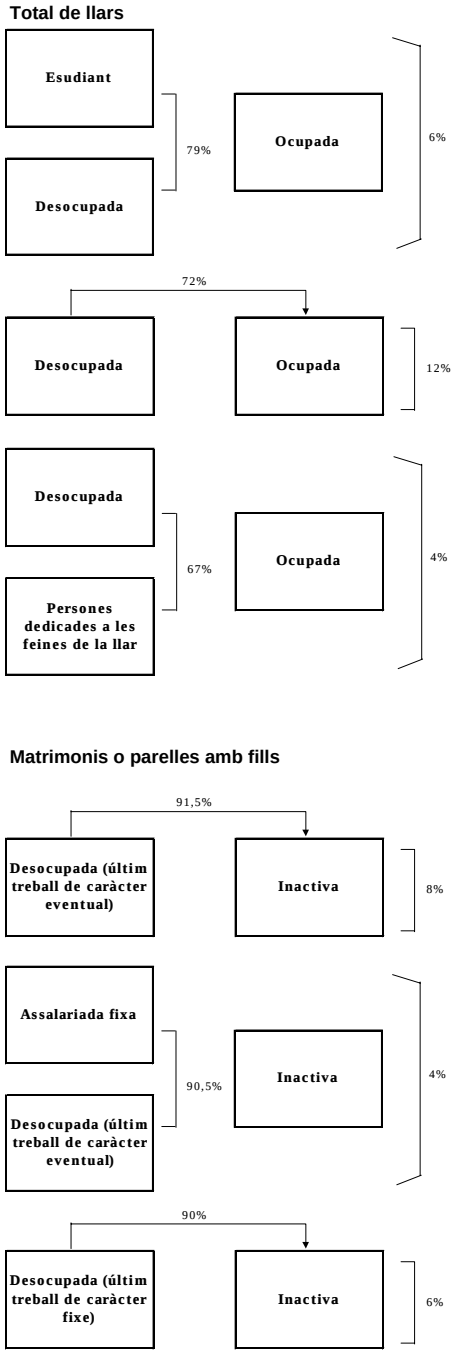
Dins les limitacions que en l'aplicació d'aquesta tècnica convé assenyalar és que l'associació només pot fer-se per cada variable per separat i no detecta els casos en què en una mateixa llar es repeteixen algunes categories de la variable. Aquestes restriccions es podrien solucionar creant unes noves que resultaran de la combinació de diferents variables, o que permetin registrar el nombre de vegades que alguna categoria de la variable en qüestió es repeteix.

Les dades censals presenten la particularitat que moltes de les associacions possibles o són conegudes o tenen un escàs valor analític. Per exemple si s'analitza la variable estat civil, per una banda es podria esperar en llars on hi ha fills, associacions del tipus casats (els pares)-solters (els fills), i d'altra banda, no enregistraria els casos on a les llars hi ha dues o més persones del mateix estat. Tampoc resultarien de major interès associacions del tipus casat/casada-vidu/vídua, atès que ja acostumen a estar prou tractades en els estudis sobre pautes familiars de convivència i/o corresidència.

En aquesta ocasió, per a l'aplicació d'aquesta tècnica es va seleccionar la variable «Relació amb l'activitat econòmica». L'estudi de l'activitat econòmica de les persones a nivell de les llars, ha estat sovint assenyalada com una de les millors vies per conèixer la incidència de l'atur a les famílies, avaluar el grau de dependència familiar i poder analitzar com el grup familiar col·labora en la reducció dels efectes de l'atur. També s'ha creat una variable que combina la relació amb l'activitat econòmica amb la situació professional.

Al moment d'aplicar aquesta segona tècnica no es va poder seguir el mateix enfocament utilitzat per la segmentació. Així, resulta del tot impossible el seu ús per a les llars

Quadre 6. Resultats de les associacions.



unipersonals i té molt poc sentit aplicar-la a les formades per matrimonis o parelles sense fills i sense altres persones, pel fet d'estar constituïdes per dues persones. Per aquestes raons s'han buscat associacions per al conjunt de llars de Catalunya que eren 1.933.044 i per al tipus de llars de major importància relativa que són els matrimonis o parelles amb fills que agrupen el 46,5% de les llars.

Al quadre 6 són reunits els resultats principals d'aquesta aplicació per als dos grups de llars esmentats. Així, pel que fa al total de llars catalanes es presenten les associacions amb una major incidència, en ordre decreixent. En aquestes tres s'observa l'existència d'una persona desocupada relacionada amb una altra en diverses situacions respecte l'activitat, tot indicant també el percentatge de llars on aquesta associació té lloc.

En un 6% de les llars catalanes, la presència d'un desocupat s'acompanya el 79% dels casos d'un estudiant i ambdues situacions són relacionades amb una persona ocupada. Aquest darrer status associat ara a l'individu desocupat assoleix un valor del 72% i afecta a un 12% de les llars. Convé assenyalar com aquest resultat, tot i que acotat al moment censal de 1991, il·lustra l'abast d'una tipologia social relativa als efectes de l'atur en les famílies. En aquest mateix sentit pot apreciar-se com en un 4% de les llars, un 67% dels casos la persona desocupada s'associa a una dedicada a les feines de la llar, sempre amb la presència d'un membre empleat. Aquestes dues situacions descrites, poden considerar-se en concordança amb allò observat, a partir d'altres fonts estadístiques, pels estudis sobre la composició de les famílies, segons la relació dels seus membres amb l'activitat. En particular, el fet que a moltes de les llars amb alguns membres desocupats com a mínim un s'ha mantingut empleat en les etapes de major desocupació de finals dels anys vuitanta i primers de la dècada dels noranta.

Pel que fa als matrimonis o parelles amb fills, el quadre 6 ofereix unes modalitats d'associació molt intensa, al voltant del 90% dels casos i totes relacionades amb persones declarades inactives, en aquest cas, mestresses de casa o fills estudiants. En un 8% d'aquest grup de llars, l'individu desocupat, però amb un darrer treball eventual, és present en el 91,5% de casos amb un d'inactiu. Un percentatge molt proper presenta en el 6% de les llars, l'associació relativa ara a un últim treball de caràcter fix. Totes dues situacions semblarien detectar grups amb certes dificultats en relació als efectes de l'atur i es diferencien de la tercera obtinguda, on si bé en un nombre menor de llars, el 4%, ara la persona desocupada s'associa a una d'assalariada, amb un treball fix.

7. CONSIDERACIONS FINALS

Fins ara, les aplicacions de Data Mining s'han efectuat principalment amb informació del sector financer, hospitalari, farmacèutic, serveis, entre altres, com també s'ha utilitzat per extreure informació de texts. En canvi, menys freqüent havia estat la seva aplicació al tractament de grans bases de dades sociodemogràfiques com pot ser un cens

de població. Gràcies a la col·laboració de l'Idescat, el Cesca i IBM, s'ha pogut realitzar l'aplicació de Data Mining a l'àmbit de l'estadística oficial. Per les característiques de les dades solament es van aplicar les tècniques de segmentació i d'associació.

Com a primer aspecte d'interès fruit d'aquests procediments pot afirmar-se que, amb la mineria de dades, s'ha tingut l'oportunitat de treballar amb la totalitat de la població (6.000.000 de registres), s'ha fet una lectura integral de la informació i s'han determinat grups de dades homogènies. Així, s'ha pogut resumir i sintetitzar un gran volum d'informació i també quantificar la intensitat de relacions entre les variables.

Convé assenyalar que moltes de les variables que recull el cens de població presenten un comportament condicionat al cicle vital de les persones i/o de les famílies. L'edat i el sexe són atributs fonamentals, i per això, alguns dels resultats obtinguts eren previsibles. Ara bé, l'aplicació de la segmentació ha constatat que de vegades l'edat i/o el sexe són les variables que defineixen el segment, però en altres casos la naturalesa de la població, l'ocupació, el sector d'activitat, etc., són aquelles que han tingut el major pes principal en la determinació del mateix.

L'aproximació clàssica en l'anàlisi de la informació censal acostuma a partir de l'estructura per sexe i edat de la població i, sobre aquesta base, continua amb l'estudi de les característiques demogràfiques, socials i econòmiques associades. En aquesta ocasió, mitjançant la utilització d'aquestes tècniques ha estat possible operar simultàniament amb un major nombre de variables i poder identificar modalitats d'agrupaments més ajustats a la natura heterogènia que caracteritza tota població. Aquesta nova perspectiva cal considerar-la com una anàlisi no convencional però complementària de l'estadística oficial.

L'Institut d'Estadística de Catalunya té previst continuar en aquesta línia de treball amb l'aplicació de la mineria de dades a la informació que proporciona l'Estadística de població de 1996 i les dades del proper cens del 2001, que permetrà aprofundir en l'anàlisi de la dimensió temporal i territorial de les estadístiques censals.

Annex: breu descripció de les operacions de Mineria de Dades

Segmentació

Es tracta d'agrupar automàticament registres que comparteixen característiques similars. Per exemple ajuda a saber quines agrupacions lògiques existeixen en el cens de Catalunya segons les característiques o el comportament de les persones estudiades.

La segmentació es basa en dues tècniques de les que es detallen les seves característiques fonamentals:

- Segmentació basada en Anàlisi Relacional:

– El número de segments es determina durant l'execució de l'algoritme.
– Processa tant variables quantitatives com qualitatives.
– Maximitza la similaritat entre els membres d'un mateix segment i les diferències entre els membres de segments diferents.
– Segmenta en base a mètriques de similaritat, no de distància.
– És eficient per a la detecció de conjunts molt petits de registres.

- Segmentació Neuronal:

– És necessari predefinir el número de segments que es vol obtenir, així com la seva distribució bidimensional.
– Processa tan variables qualitatives com quantitatives, però funciona millor amb les darreres.
– És eficient quan es vol particionar una població imposant certa relació entre els segments obtinguts.

Classificació

Es tracta de predir un resultat entre diferents possibilitats i d'explicar els factors de classificació. Per exemple saber quines característiques tindrà una llar catalana en la que els seus membres tinguin coneixement del català parlat i escrit.

La classificació es basa en dues tècniques que són els arbres de decisió i les xarxes neuronals i les seves principals característiques són:

- Classificació basada en Arbres de Decisió:

– Model de classificació en forma d'arbre de decisió binària.

– Treballa amb la tecnologia SLIQ (Supervised learning In Quest), processant tant variables quantitatives com qualitatives.
– Incorpora una tècnica de podat basada en el principi de la mínima longitud de descripció (MLD), que proporciona arbres més senzills.
– Es escalable i pot processar conjunts amb independència del número de classes, atributs i registres.

• Classificació Neuronal:

– Basada en xarxes neuronals de propagació cap endarrere.
– Detecta de forma automàtica la topologia més adequada per a cada problema. També permet especificar una concreta.
– Realitza una anàlisi de sensibilitat per a detectar les variables més significatives per a cada tipologia.

Predicció

Es tracta de predir un valor per un registre entre un conjunt continu de possibilitats. Per exemple saber quina probabilitat hi ha que una mesura sanitària concreta sigui necessària per les llars catalanes d'unes característiques determinades.

Les principals característiques de les dues tècniques en què es basa la predicció són les següents:

• Funcions de Base Radial:

– Poden processar variables quantitatives i qualitatives a la vegada.
– Detecta el número de centroides òptim, predefinint el número màxim d'aquests i el número mínim de registres assignats a cada centre.
– Funciona especialment bé quan l'estructura de les dades tendeix a agrupar-se en conjunts, ja que implementa cert tipus de segmentació.

• Predicció Neuronal:

– Basada en xarxes neuronals de propagació cap endarrere.
– Detecta de forma automàtica la topologia més adequada para cada problema, encara que permet especificar una concreta.
– Permet predir dades en forma de sèries temporals.
– Permet implementar regressió logística.

Anàlisi de Relacions

Es tracta relacionar característiques o comportaments que estan relacionats entre si. Per exemple saber quines variables soci-cultural-econòmiques van normalment juntes.

L'anàlisi de relacions es basa en tres tècniques de les que es detallen les seves característiques fonamentals:

- Anàlisi d'Associacions:

– Detecta elements en una transacció que impliquen la presència d'altres elements en la mateixa.
– Determina les afinitats entre elements en forma de regles d'associació XY, facilitant una sèrie de mètriques com son el suport, la confiança, el tipus de la regla, etc.
– Permet incorporar taxonomies de productes, permetent la detecció d'associacions a diferents nivells.

- Patrons seqüencials:

– S'introdueix la variable temps. Es detecten patrons entre transaccions, el que permet determinar elements en una transacció que impliquen la presencia d'altres elements en una altra transacció.

- Anàlisi de similaritat en series temporals:

– Detecta totes les ocurrències de seqüències similars en una col·lecció de sèries temporals.

Com a complement a les operacions/tècniques de mineria de dades el producte «DB2 Intelligent Miner for Data» incorpora les següents funcions de preprocés i funcions estadístiques que van ser de gran utilitat en el desenvolupament del projecte.

- Funcions de preprocés de dades:

– Recollida de dades: extraccions de bases de dades i d'altres fitxers.
– Selecció de dades: filtrat, preparació de mostres, agregació, projecció, agrupació i, resum de dades.
– Transformació: conversió, codificació, discretització de dades.
– Fusió i enllaç de taules.
– Consistència: verificació, filtrat i descart de dades.
– Derivació: càlcul de nous descriptors a partir de les dades disponibles.
– Processat de sentències SQL contra bases de dades.

- Funcions estadístiques:

– Elaboració de descriptives univariants i bivariants; càlcul de percentils, ANOVA, i proves F.
– Anàlisi de components principals i anàlisi de factors.
– Ajust de corbes per a sèries temporals.
– Regressió lineal múltiple i regressió polinòmica.
– Anàlisis d'ocurrències.

BIBLIOGRAFIA

Adriaans, P. & Zantinge, D. (1996). *Data Mining*. Addison-Wesley: Harlow England.

Bigus, J. (1996). *Data Mining with Neural Networks*. McGraw-Hill: New York.

Cabena, Peter, *et al.* (1997). *Discovering Data Mining: from concept to implementation*. Prentice-Hall: New Jersey.

Hinde, A. (1998). *Demographic Methods*. Arnold: London.

Institut d'Estadística de Catalunya (1994). *Cens de població 1991*. Generalitat de Catalunya: Barcelona.

Institut d'Estadística de Catalunya (1997). *Llars i famílies a Catalunya 1991*. Generalitat de Catalunya: Barcelona.

Leridon, H. & Toulemon, L. (1997). *Démographie. Approche statistique et dynamique des populations*. Economica: Paris.

ENGLISH SUMMARY

POPULATION CENSUS: AN INTERPRETATION USING DATA MINING

CRISTINA GUISANDE ALLENDE*
FRANCESC SUBIRADA CURCO**

This article presents the results of a research job carried out jointly by the Institut d'Estadística de Catalunya (Idescat), the Centre de Supercomputació de Catalunya and IBM. They analysed data from the 1991 population and households Census using Data Mining methodology. Segmentation and association techniques were used on individuals according to the type of household they belonged to.

Keywords: Data Mining, official statistics, association, segmentation, data analysis, applied statistics, statistical methods

AMS Classification (MSC 2000): 62-07, 62P25, 62H30, 62H20

* Institut d'Estadística de Catalunya. Via Laietana, 58. 08003 Barcelona. E-mail: cguisande@idescat.es
** CEPBA-IBM Research Institute. Jordi Girona, 1-3. 08034 Barcelona. E-mail: frsubirada@es.ibm.com

– Received April 2001.
– Accepted November 2001.

1. INTRODUCTION

The Institut d'Estadística de Catalunya (Idescat), the Centre de Supercomputació de Catalunya (Cesca) and IBM signed a cooperation agreement to carry out a research on data obtained from the 1991 population and households Census using Data Mining. This article aims to present the main results of this experience, where Data Mining was applied on census information.

2. DATA MINING

The following items were used while conducting the research:

1. Softwarewise:

IBM's DB2 Intelligent Miner for Data. With this program, Data Mining operations can be performed on records in order to find out previously unknown information.

2. Hardwarewise:

IBM's System Parallel 2 (SP2). This parallel process computer allows the simultaneous use of several microprocessors to address a single problem. When using Data Mining it can deal with big databases in a reasonable amount of time.

The following operations can be performed: segmentation, classification, forecasting and analysis of relationships. Because of the nature of census data, only two of the aforementioned operations were carried out: segmentation and, as regards analysis of relationships, association.

3. CENSUS DATA

The main source of demographic data is the population census, and that makes it a very valuable tool for social and economic planning. It collects information about all the inhabitants on a given date –cross-section data– and it gives out the so called «picture» of the population at a given moment.

The survey for the 1991 Census included 31 questions relating to individuals and 6 others relating to housing.

The main objects of investigation were:

a) geographical aspects: place of residence.

b) demographic aspects: sex, age, marital status, place of birth, nationality, relationship of the dwellers, fertility, date of marriage, migration, mobility.

c) social or cultural aspects: education, knowledge of Catalan.

d) economic aspects: work, sector, professional status, occupation.

e) housing aspects: date of construction, property, area, number of rooms, fittings and appliances.

4. DATA PROCESSING

A process of selection, testing and analysis of the variables has to take place before using data mining techniques.

Some variables, like household equipment, have been discarded for they don't really discriminate, having always very similar values. Other variables, despite their relative low weight against the whole of the population, turned out to be important for determining specific groups (population niches), for example foreign population, which represented 1% of the total population.

In order to optimise the data processing, some continuous variables were grouped in intervals (age, date of arrival in Catalonia, date of construction of the house), thus turning discrete, while others were turned into categorical variables, for example the change of municipality of residence during the previous ten years as a migration indicator. Other variables were combined, like work and professional status, and yet other new ones were created, like number of individuals in the household.

Segmentation and association techniques have been applied on individuals according to the type of household to which they belong. What makes this approach interesting is that many demographic and some social and economic aspects are quite influenced by some types of household composition. Also, the Institut d'Estadística de Catalunya has, for the first time, obtained detailed information from the 1991 population Census about family structures for any region in the Catalonian territory.

5. SEGMENTATION: MAIN RESULTS

This article presents the most outstanding results obtained from applying segmentation to households formed by a single person and to simple family units formed by couples

with or without children. As it turns out, these three types represented 79% of all the households and 75% of the population of Catalonia back in 1991.

A quick overview –figure 2– reveals, as regards single-person households, that the five segments grouped in the figure represent as much as 64% of these households, although each segment has a different relative weight, the most frequent being households formed by retired widows.

Regarding couples without children, a similar number of of segments –six– represent an even higher percentage of households (81%). They also show an uneven distribution, with the first three segments answering for 74% of the total.

Finally, for couples with children a total of five segments once again stand for 90% of the total. As with couples without children, a more homogeneous population grouping –formed by students' children, working fathers, housewives, working sons or daughters and working mothers– is revealed.

6. ASSOCIATIONS: MAIN RESULTS

The use of this technique reveals associations between different elements. On one hand, we obtain a measure of the intensity of the relation and, on the other hand, the frequency with which this rule of association is detected. When it comes to census data, search for associations must be done separately for each variable in each household. By doing that, we obtain the associations between different categories of the chosen value.

It should be noted that, among other limitations of this technique, association can only be done on each saparate variable and it won't detect those cases where some categories of the variable are repeated in a household. These restrictions could be avoided by creating new ones, either by combining different variables, or by registering the number of times a given category of that variable is repeated.

The variable selected for the application of this technique was «Relationship to economic activity». The study of the individuals' economic activities householdwise has often been pointed out as one of the best markers for revealing the impact of redundancy on families, evaluate the degree of family dependence and analyse the way the family unit contributes to reducing the effects of redundancy. Another variable, which combines relationship to economic activity with professional status, has been created.

When it came to using this second technique, we couldn't use the same approach as we did with segmentation. Moreover, it is absolutely impossible to use it with households formed by a single individual, and makes little sense with couples without children or other dependants, since they're made of only two persons. Therefore, we have looked for associations for the most important type of households, which are those formed by

couples with children, amounting up to 46.5% of the 1,933,044 total households censed in Catalonia.

7. FINAL CONSIDERATIONS

So far, Data Mining had been applied mainly on information from the financial, hospital, pharmaceutical, and service sectors, as well as to extract information from texts. It had seldom been applied to large sociodemographic databases like a population census. Thanks to the cooperation between Idescat, Cesca and IBM, Data Mining has been used on official statistical information. Because of the nature of the data only segmentation and association techniques were used.

It should be noted that using these procedures (data mining) has given us the chance to work with the total population (6,000,000 records), achieve an integral reading of the information, and determine groups of homogenous data. Moreover, it has allowed us to summarize and synthetize a great volume of information, and also to quantify the intensity of the relatioships between variables.

The Institut d'Estadística de Catalunya plans to continue using data mining on information provided by 1996 population Statistics and on data from the 2001 Census, which will help us to further deepen the analisys of census statistics in both dimensions, time and space.