QÜESTIIÓ, Vol 20, 2 pp. 273-291, 1996

ESTUDI DELS CODIS OBTINGUTS PER CODIFICACIÓ AUTOMÀTICA ENFRONT DELS CODIS PRECODIFICATS. APLICACIÓ PER AL CPH/91 DE LA VARIABLE ACTIVITAT

MIQUEL DELGADO ALZAMORA*
JOSEP ANTON SÁNCHEZ CEPEDA*

Institut d'Estadística de Catalunya

Al cens de població i habitatge de 1991 es va donar la situació de que les preguntes d'ocupació i activitat es responien mitjançant una descripció —literal— i un codi precodificat. A la fase de tractament de la informació censal es realitza el procés de codificació automàtica d'aquests literals amb la qual cosa s'aconsegueix tenir per cadascuna d'aquestes preguntes dues respostes codificades.

Aquesta situació ens porta a l'estudi de la doble resposta amb la finalitat de constatar si els enquestats codifiquen bé les seves preguntes o pel contrari tenen confussions.

L'objectiu de l'estudi és avaluar la coincidència i estabilitat de les respostes obtingudes amb cada sistema i establir quins problemes es produeixen amb cada tipus de classificació, problemes derivats alguns de l'excessiva generalització a les categories del precodificat, de l'existència d'ambigüitat o d'absència del literal on el precodificat ens pot servir d'ajut per classificar-lo. Per fer-ho definirem uns paràmetres d'estudi, els valors dels quals ens determinaran l'agrupació i/o separació de codis de les variables.

Analysis of automatic coding values versus respondent's values. Application for Activity and Occupation.

Keywords: Cens de població i habitatge, pregunta codificada, codificació automàtica.

*Miquel Delgado Alzamora, Josep Anton Sánchez Cepeda. Institut d'Estadística de Catalunya. Departament d'Economia i Finances. Generalitat de Catalunya. Via Laietana, 58. 08003 Barcelona.

- Article rebut el novembre de 1995.
- Acceptat el juny de 1996.

1. INTRODUCCIÓ

En el Cens de Població i Habitatge de 1991 es va donar la situació de que les preguntes d'Ocupació i Activitat es podien respondre de dues formes, per una banda es posava una de les categories d'una llista de codis (resposta precodificada) i l'altra alternativa era que la persona responia mitjançant una descripció (literal obert).

L'Institut d'Estadística de Catalunya pren la decisió de gravar les dues respostes a cada variable i realitzar la codificació dels literals oberts, mitjançant el corresponent procés de codificació.

A la variable Activitat i a l'Ocupació es procedeix a la codificació dels literals oberts en un dels codis possibles de les classificacions oficials vigents en el moment de la codificació: La Classificació Nacional d'Activitats Econòmiques de 1974 (CNAE-74) i la Classificació Nacional d'Ocupacions (CNO-79), ambdues a un nivell de desagregació de tres dígits.

Aquesta situació ens porta a l'estudi de la doble resposta per a cada pregunta amb la finalitat de constatar si els enquestats codifiquen bé les seves pròpies respostes o pel contrari tenen confusions en alguns casos. Es parteix de la base de que la descripció donada per l'enquestat com a resposta és el valor més fiable; davant la necessitat de codificar-lo es pren com a vàlid el valor resultant de la codificació automàtica enfront del codificat per l'enquestat, per haver estat sotmesa la codificació automàtica a un procés exhaustiu de depuració de les variants.

Tots dos sistemes de classificació tenen avantatges i desavantatges. El precodificat aporta facilitat en el moment de respondre per part de la persona, rapidesa i un menor cost a l'hora de la gravació, però té l'inconvenient de la seva excessiva generalització al categoritzar les variables, amb la conseqüent pèrdua de detall. A més, la taxonomia escollida per classificar cada variable va ser poc afortunada, pel fet d'haver utilitzat unes classificacions a mig camí entre les velles i les noves (aquestes últimes no estaven encara vigents en el moment de l'operació). D'altra banda, les classificacions basades en el codificat aporten un major nivell de detall en la classificació al basar-se en el literal que la pròpia persona descriu, però hi ha problemes si un literal és molt ambigu no podent-se ubicar en una de les categories.

L'objectiu de l'estudi és avaluar la coincidència i l'estabilitat de les respostes obtingudes amb cada sistema i establir quins problemes es produeixen amb cada tipus de classificació, problemes derivats alguns de l'excessiva generalització a les categories del precodificat, de l'existència d'ambigüitat o d'absència del literal on el precodificat ens pot servir d'ajut per classificar-lo. Per fer-ho definirem uns paràmetres d'estudi, els valors dels quals ens determinaran l'agrupació i/o separació de codis de les variables.

2. PARÀMETRES D'ESTUDI

Considerem una variable que té n codis possibles com a resposta. Disposar de dues respostes alternatives per a cada pregunta genera una matriu de dimensions $n \times n$ on a les columnes hi posem els n codis possibles provinents de la precodificació i a les files els n codis provinents de la codificació automàtica. D'aquesta matriu calculem el tant per cent de coincidència entre la resposta segons el precodificat i el total de respostes segons la codificació automàtica. Considerem el següent exemple amb $n = 4$.

A. Matriu de codi absoluts

	Precodificats				
Codificació	$P1$	$P2$	$P3$	$P4$	total
$C1$	5000	1000	50	50	6100
$C2$	900	1000	50	50	2000
$C3$	250	200	1000	50	1500
$C4$	50	50	50	10000	10150

B. Matriu de percentatges

	Precodificats			
Codificació	$P1$	$P2$	$P3$	$P4$
$C1$	81.9	16.4	0.8	0.8
$C2$	45.0	50.0	2.5	2.5
$C3$	16.6	13.3	66.6	3.3
$C4$	0.5	0.5	0.5	98.5

La diagonal correspondrà, per a cada codi, amb el tant per cent de casos ben precodificats.

1. Paràmetre «diag»: Per a cada codi el tant per cent de casos que estan ben codificats. En el nostre exemple quedaria, per a cada codi:

codi i	*diag (i)*
$C1$	81.9
$C2$	50.0
$C3$	66.6
$C4$	98.5

És interessant, tanmateix, saber per a cadascun dels codis (els anomenarem receptors) com es distribueixen la resta de codis que havien estat prèviament precodificats (els anomenarem donants). Es pot donar la situació en què els casos erròniament precodificats, als que anomenarem residus, es distribueixin entre diferents codis o es concentrin en uns pocs. També mesurarem si els residus es distribueixen de forma homogènia o si alguns destaquen per sobre d'altres.

2. Paràmetre «acum»: Per a cada codi, el nombre de codis (inclòs aquest mateix) necessaris per explicar el 95% dels casos. Mesura si els casos precodificats erròniament per a cada codi es concentren en pocs o molts codis i, per tant, si el codi en qüestió es confon amb pocs o molts.

codi i	*acum (i)*
$C1$	2
$C2$	2
$C3$	3
$C4$	1

3. Distribució «MIJO»: Per a cada codi es calcula el pes del residu sobre la diagonal sempre que el pes sobrepassi un llindar de .05. Mesura els codis que recullen un pes important sobre el codi de la diagonal i la seva distribució.

codi i	*mijo1(i)*	*mijo2(i)*	*mijo3(i)*
$C1$	2	01	01
$C2$	9	05	05
$C3$	25	02	05
$C4$	005	005	005

4. Distribució «Donant»: Es composa dels codis donants de la distribució anterior.

codi i	*donant1(i)*	*donant2(i)*	*donant3(i)*
$C1$	2		
$C2$	1		
$C3$	1	2	
$C4$			

5. Paràmetre «nmijo»: De la distribució Donant, per a cada codi quants residus superen el llindar.

codi i	*acum (i)*
$C1$	1
$C2$	1
$C3$	2
$C4$	0

Categoritzarem el paràmetre «diag» perquè sigui més orientatiu. Les categories que hem realitzat són:

1. Molt atractora o receptora: > 95
2. Atractora: de .85 a .95
3. Bastant atractora: de .7 a .85
4. Poc atractora: de .5 a .7
5. Molt poc atractora: $< .5$

El mètode per a l'obtenció dels paràmetres està programat en SAS i produeix com a output:

- La matriu de precodificació-codificació automàtica en codis absoluts
- La matriu de precodificació-codificació automàtica en percentatges
- Per a cada codi, els percentatges i donants ordenats en ordre decreixent i el paràmetre «diag».
- Les distribucions Mijo i donant amb el paràmetre Nmijo.
- Els paràmetres Diag, pcdonant i acum sense parametritzar i parametritzats.

Amb la distribució donant es genera un graf per visualitzar millor els resultats.

3. ESTUDI DE LA VARIABLE ACTIVITAT

Apliquem el mètode a la codificació de l'«Activitat de l'Establiment» del CPH/91. El nombre de persones que responen que són ocupats, desocupats amb treball anterior o jubilats —que són els únics que han de respondre a la pregunta en qüestió— puja a 3.150.056. D'aquests no hi consta el literal per 165.762, no hi consta el precodificat per 241.378, —en 152.798 no hi consta cap dels dos—, no s'ha pogut codificar el literal per a 399.802 i s'han codificat 2.495.912. Aquests últims són els que prenem per al present estudi. (Els resultats es troben a l'annex 2.)

Els valors de la diagonal principal de la matriu de percentatges oscil·len entre el 95,87% del codi 01 al 42,35 del codi 07, essent la mitjana general del 74,40% amb desviació del 17,35.

Comencem observant l'output dels paràmetres acum, receptor i donant. Podem observar el següent:

- Hi ha 3 codis amb receptor = 5 i acum = 1, és a dir, codis que tenien més d'un 95% dels casos ben precodificats i a més no actuen com a donants d'altres codis. Això vol dir que no recullen casos precodificats erròniament que corresponen a altres codis. Aquests codis —01,06 i 23— no presentaran problemes de confusió.

- En el cas extrem es troben codis amb receptor = 1, és a dir, codis que tenien ben precodificats menys del 50% dels casos. Podem distingir, altrament, dos grups:

    1. Els codis 03 i 07 tenen acum 3 i 4 respectivament, significant que els codis erròniament precodificats provenen de pocs codis.
    2. El codi 15 té acum igual a 11 i, per tant, els seus casos erronis es troben dispersos entre molts codis.

- El paràmetre receptor presenta la següent distribució:

  Val 5 en 3 casos, 4 en 7, 3 en 9, 2 en 7 i 1 en 3; per tant els codis de la variable Activitat que estaven ben precodificats s'igualen amb els que no ho estaven.

- Destaquen com a codis molt dispersos el 19 i el 24 que tenen acum 16 i 14 respectivament, els quals resulten poc atractors.

Passem ara a les distribucions Mijo i donants en les quals poden observar:

- Destaquen els codis 03 i 07, que tenen un mijol superior a 1, el que significa que hi ha un altre codi que recull més casos precodificats que ell mateix. Això

és indicatiu de gran confusió entre parelles de codis: el 03 amb el 04 i el 07 amb el 17. El següent mijol notable —.6— correspon amb el codi 09 que s'emparella amb el 03.

- Hi ha 12 codis en què cap residu sobrepassa l'.1, però si es baixa el límit al .05 es mantenen 9 d'ells. És a dir, per aquests codis no hi ha cap codi donant suficientment significatiu.

- Per alguns codis hi ha un residu notablement més fort que els altres —cas del codi 03, 07, 09 i 13— però per a altres la importància dels residus està més repartida entre ells sense que en destaqui cap.

  El següent pas és estudiar el graf resultant d'unir les variables que apareixen en la distribució Mijo com a receptors amb les que apareixen en la donant com a tals, el que ens donarà una idea de les confusions que es produeixen entre els diferents codis. Hi ha 9 codis —1, 2, 6, 8, 18, 21, 22, 23 i 25— que queden aïllats: tots tenen un codi alt del paràmetre receptor i baix del paràmetre donant com era d'esperar. La resta es reparteix en 3 grups inconnexes, 2 d'ells petits i un altre més díficil d'interpretar.

  Es pot donar una explicació dels grups resultants:

  1. El grup format pels codis 05-20-19 on s'ha confós el que és, primerament el comerç al detall amb el comerç a l'engròs i seguidament tot el que és comerç amb la indústria alimentària i del tabac.
  2. El format pels codis 24-26-27-28-29, les relacions que s'estableixen entre una sèrie de serveis que s'ha prestat a confusió, essent els «Altres serveis» un sac per algun d'ells. Apareix entre aquests codis l'Administració Pública i la confusió entre el que és i no és la seva pròpia activitat.
  3. El format pels altres codis que no queden aïllats i dels quals es poden veure 3 subgrups més diferenciats:
     3.1. El format pels codis 03-04-09-10 on es confon que són tot tipus d'indústries extractives amb el refinament i tractament dels productes que s'extreuen d'elles.
     3.2. El format pels codis 07-15-17-16 on es confon el que és la indústria de la fusta i el suro amb la fabricació de mobles, producte de matèries plàstiques i també amb el que és la fusteria, fontaneria, etc.
     3.3. El format pels codis 11-12-13-14. Les relacions que es formen aquí giren al voltant dels codis de l'apartat d'indústries metàl·liques i les seves transformacions, amb contínues confusions entre uns i altres, ja que en la majoria de casos per a una bona codificació es necessitarien sistemes de codificació amb un nivell alt de detall.

L'explicació que s'obté per a cada un dels codis a partir de tota la informació disponible és la següent:

Codi 01. *Agricultura, ramaderia, caça i silvicultura (producció i serveis annexes)*

Resulta un codi molt atractor —diag(1) = 95.87 acum(1) = 1—. No apareix en la distribució Mijo ni en la distribució donant, això vol dir que no rep significativament de cap altre codi —el més important és el 13 del que rep un .9%— i no dóna a altres codis de forma significativa —el codi 02 rep d'ell un 4.51%—. Com a conclusió podem afirmar que aquesta variable ha estat ben precodificada.

Codi 02. *Pesca i piscicultura*

Resulta un codi atractor —diag(2) = 89.24 i acum(2) = 3—. No apareix a la distribució Mijo ni a la distribució donant, és a dir, no rep significativament de cap codi —el més important és el 01 d'on rep un 4.51%— i no dóna significativament a altres codis —el codi 20 rep d'ell un .39%—. Com a conclusió és pot afirmar, com abans, que aquesta variable també ha estat ben precodificada per la gent.

Codi 03. *Extracció de combustibles sòlids, petroli, gas natural i minerals radioactius*

Resulta un codi poc atractor —diag(3) = 44.89 i acum(3) = 3—. Aquest codi d'acum ens informa que el que resta per arribar al 95% es distribueix entre només 2 codis més. A la seva distribució Mijo destaca, sobretot, el codi 04 de qual rep un 45.52% i el 10 del que rep un 4.9%; a la vegada actua com a donant del codi 9 d'una forma important.

Es pot concloure que, en gran mesura, aquest codi ha estat confós principalment amb el 04 i en un grau menor amb el 09.

Codi 04. *Resta d'indústries extractives: ferro i minerals metàl·lics no energètics*

Aquest és un codi força atractor —diag(4) = 78.74 i acum(4) = 6—. Amb això sabem que amb 5 codis més arribem al 95%, és a dir, estan bastant repartits els codis del quals rep, sobressortint el 10 de qui rep un 9.3% perquè els altres ja deixen de ser significatius. Important és el fet de ser donant del codi 03 de forma que sobrepassa, inclús, el codi de diag(3). Aquest codi, a més de ser confós amb el 10 —fabricació de productes químics— provoca confusió amb el 03 per ser donant molt important d'ell.

Codi 05. *Indústries de productes alimentaris, begudes i tabac*

Codi força atractor —diag(5) = 78.2 i acum(5) = 5—. No dóna de forma significativa a cap codi, però rep un 11% del 20 —Comerç al detall—, per tant, s'ha confós amb el 20.

Codi 06. *Indústries tèxtil, cuir, sabateria i confeccions tèxtils*

Resulta molt atractor —diag(6) = 95.3 i acum(6) = 1—. No apareix a la distribució Mijo ni a la distribució donant, això vol dir que no rep significativament de cap altre codi ni dóna a altres codis.

Podem concloure que aquest codi ha estat ben precodificat per la gent.

Codi 07. *Indústria de la fusta i el suro*

Resulta un codi molt poc atractor —diag(6) = 42.35 i acum(6) = 4—. Aquest codi d'acum ens informa que el que resta per arribar al 95% es distribueix entre només 3 codis més. A la seva distribució Mijo destaca sobretot el codi 17 —Construcció— del que rep un 7.12%, a la vegada actua com a donant del codi 15 d'una forma significativa. Entre els codis 15 i 07 hi ha una confusió recíproca, actuant mútuament com a donants i com a receptors.

Codi 08. *Indústries de paper, arts gràfiques, edició i reproducció de suports i gravats*

És un codi atractor —diag(8) = 90.1 i acum(8) = 5—. No apareix a la distribució Mijo ni a la distribució donant, no rep significativament de cap altre codi i no dóna a altres codis de forma significativa. Els casos precodificats erròniament d'aquest codi són el 10% però es distribueixen entre els altres de forma homogènia.

Codi 09. *Coqueries*

Codi poc atractor —diag(9) = 50.89 i acum(9) = 6—. No és un codi significatiu per a cap altra codi però rep del 3 -30.97% —i del 10 -5.3%—. Poc més de la meitat dels codis han estat ben precodificats i ha estat confós amb els codis 03 i 10.

Codi 10. *Fabricació de productes químics, fibres artificials i sintètiques, productes minerals no metàl·lics*

Codi força atractor —diag(10) = 77.05 i acum(10) = 8—. Aquest codi és molt donant ja que apareix com a tal per a 4 codis —03, 04, 09 i 15—. Possiblement aquest codi té un camp de significat ampli que porta a la confusió amb diferents codis. D'altra banda, rep del 17 un 9.25% i els altres 7 codis que ens indica acum estan molt repartits. Podem considerar aquest codi com molt poc definit i molt ampli.

Codi 11. *Producció de metalls*

Codi força atractor —diag(11) = 72.28 i acum(11) = 6—. Rep del codi 12 un 17.18% però dóna un 14.19% dels casos a aquest mateix codi, indicant-nos que hi ha una mútua confusió entre ambdós codis.

Codi 12. *Fabricació de productes metàl·lics, construcció de màquines, equip i material mecànic*

Codi poc atractor —diag(12) = 62.2 i acum(12) = 11—, trobant-se els casos erròniament precodificats molt repartits, a excepció dels provinents de l'11. Apareix com a donant dels codis 11, 13, i 14, sent tots ells de fabricació d'algun material o equip, fet que pot portar a confusió.

Codi 13. *Fabricació d'equip i material elèctric, electrònic i òptic*

Codi poc atractor —diag(13) = 50.64 i acum(13) = 12—, necessita de molts donants per recollir el 95% dels casos. No actua com a donant però sí com a receptor dels codis 17 —19.09%— i 12 —8.9%—. Codi poc definit que ha estat precodificat amb molts altres codis. La confusió amb el 17 —Construcció— és de difícil explicació.

Codi 14. *Fabricació de material de transport*

Codi bastant atractor —diag(14) = 74.05 i acum = 7—. No actua com a donant però sí rep del 12 un 9.4%, presentant una lleugera confusió amb ell.

Codi 15. *Fabricació de productes de suro i matèries plàstiques. Altres indústries manufactureres*

Codi molt poc atractor —diag(15) = 49.46 i acum(15) = 11—. Necessita, també, de molts donants per recollir el 95% dels casos. Actua com a donant del 07 —ja s'explica anteriorment— i és receptor dels codis 10, amb un 10.93%, del 07 amb un 9.19% i del 17 amb un 5.62%. Codi molt repartit entre altres codis. Existeix una confusió mútua entre el 15 i el 17. —els dos treballen amb la fusta— i confusió amb el 10 i el 17 significativa.

Codi 16. *Producció, transport i distribució d'energia elèctrica, gas i aigua*

Codi bastant atractor —diag(16) = 70.97% i acum(16) = 8—. No actua com a donant de cap altre codi però rep del 17 un 11.9%. La confusió pot originar-se per la definició dels codis 16 i 17.

Codi 17. *Construcció*

Codi atractor —diag(17) = 92.28 i acum(17) = 5—. No és receptor però sí, en canvi, és sumament donant, significant que a ell han anat a parar força casos d'altres codis; per a 4 codis dels 5 que dóna —07, 10, 13 i 16— és el principal codi donant i, a més, dóna de forma significativa al 15. Notem que la definició del codi pot confondre a la gent d'altres activitats en ser molt àmplia.

Codi 18. *Venda, manteniment i reparació de vehicles a motor. Gasolineres*

Codi bastant atractor —diag((18) = 77.29 i acum(18) = 7)—. No és receptor ni donant de cap codi. Es considera un codi ben precodificat.

Codi 19. *Comerç a l'engròs i intermediaris*

Codi poc atractor —diag(19) = 59.19 i acum(19) = 16—. Destaca el seu codi d'acum que ens indica que s'havia precodificat amb molts altres codis, del qual sobressurt el 20 amb un 8.6%, a més, actua a la vegada com a donant del mateix 20. La confusió mútua del 19 i el 20 és clara en ser tots dos diferents tipus de comerç.

Codi 20. *Comerç al detall i reparacions d'efectes personals i béns domèstics*

Codi poc atractor —diag(20) = 68.82 i acum(20) = 12—, trobant-se els casos erròniament precodificats molt repartits, amb l'excepció dels que provenen del 19. És donant del 05 i del 19.

Codi 21. *Hotels, restaurants i bars*

Codi atractor —diag(21) = 93.54 i acum(21) = 2—. No apareix a la distribució Mijo ni a la distribució donant, és a dir, no rep significativament de cap altre codi i no dóna a altres codis de forma significativa. Es pot concloure que aquesta variable ha estat ben precodificada per la gent.

Codi 22. *Transport i activitats annexes. Comunicacions*

Codi atractor —diag(22) = 82.45 i acum(22) = 7—, els casos erròniament precodificats estan molt repartits però cap d'ells és significatiu. No apareix a la distribució Mijo ni a la distribució donant, és a dir, no dóna a un altre codi de forma significativa ni rep de cap codi significativament. Com a conclusió es pot afirmar que aquest codi ha estat ben precodificat per la gent.

Codi 23. *Institucions financeres i assegurances*

Codi molt atractor —diag(23) = 95.47 i acum = 1—. No apareix a la distribució Mijo ni en la donant, per tant, ni rep ni dóna a cap altra codi de forma significativa. Es pot concloure que ha estat ben precodificat per la gent.

Codi 24. *Activitats immobiliàries i de lloguer de béns. Serveis prestats a les empreses*

Codi poc atractor —diag(24) = 52.59 i acum(24) = 14—. Destaca el codi d'acum que ens indica que s'havia precodificat com molts altres codis, sobressortint significativament el 27 amb un 11.55% i el 29 amb un 8.65%. No actua com a donant de cap codi.

Codi 25. *Educació*

Codi atractor amb diag(25) = 93.19 i acum(25) = 2. No apareix a la distribució Mijo ni a la donant, és a dir, no dóna a un altre codi de forma significativa ni rep de cap codi significativament. Com a conclusió es pot afirmar que aquest codi ha estat ben precodificat per la gent.

Codi 26. *Sanitat, serveis veterinaris i assistència social*

Codi atractor amb diag(26) = 87.64 i acum(26) = 3. No actua com a receptor però sí és donant del codi 27, ja que inclou a la seguretat social.

Codi 27. *Administracions públiques, Defensa i Seguretat Social*

Codi atractor amb diag(27) = 83.22 i acum(27) = 4. És receptor del codi 26 i donant del 24.

Codi 28. *Servei domèstic*

Codi atractor amb diag(28) = 93.26 i acum(28) = 2. No és receptor de cap codi però sí donant del 29 amb un 20.27% de forma significativa.

Codi 29. *Altres serveis recreatius, culturals i esportius. Representacions diplomàtiques*

Codi poc atractor amb diag(29) = 56.65 i acum(29) = 12 i els seus codis erronis es reparteixen entre molts codis, fet esperat ja que es tracta d'una activitat definida com «altres». Actua com a donant del codi 24 amb un 8.65% i del 28 amb un 20.27%.

Annex 1

Correspondència entre la classificació censal[1] i la utilitzada a la codificació automàtica per la variable Activitat[2]

CNAE	CPH/91	CNAE	CPH/91	CNAE	CPH/91	CNAE	CPH/91
011	1	341	13	474	8	812	23
012	1	342	13	475	8	813	23
013	1	343	13	481	15	814	23
014	1	344	13	482	15	819	23
015	1	345	12	491	15	821	23
016	1	346	13	492	15	822	23
019	1	347	13	493	15	823	23
021	1	351	13	494	15	831	23
022	1	352	13	495	15	832	23
023	1	353	13	501	17	833	24
024	1	354	13	502	17	834	24
029	1	355	13	503	17	841	24
030	1	361	14	504	17	842	24
040	1	362	14	611	19	843	24
051	1	363	14	612	19	844	24
052	1	371	14	613	19	845	24

[1]La classificació censal de la variable activitat de l'establiment és la que es troba en el mateix qüestionari del Cens de Població i Habitatge la qual conté els 29 codis possibles amb els quals els enquestats precodificaven l'activitat en la que treballaven.

[2]La classificació utilitzada per codificar automàticament els literals d'activitat és la CNAE a 3 dígits.

CNAE	CPH/91	CNAE	CPH/91	CNAE	CPH/91	CNAE	CPH/91
061	2	372	14	614	19	846	24
062	2	381	14	615	19	849	24
111	3	382	14	616	19	851	24
112	3	383	14	619	19	852	24
113	3	389	14	621	19	853	24
114	9	391	13	629	19	854	24
121	3	392	13	631	19	855	24
122	3	393	13	632	19	856	24
123	3	399	13	633	19	859	24
124	3	411	5	634	19	861	24
130	9	412	5	635	19	869	24
140	3	413	5	636	19	911	27
151	16	414	5	637	19	912	27
152	16	415	5	638	19	913	27
153	16	416	5	639	19	914	27
160	16	417	5	641	20	915	27
211	4	418	5	642	20	916	27
212	4	419	5	643	20	917	27
221	11	420	5	644	20	921	29
222	11	421	5	645	18	922	29
223	11	422	5	646	18	931	25
224	11	423	5	647	20	932	25
231	4	424	5	648	20	933	25
232	4	425	5	651	21	934	25
233	4	426	5	652	21	935	25
234	4	427	5	653	21	936	25
239	4	428	5	654	21	937	25
241	10	429	5	661	21	941	26
242	10	431	6	662	21	942	26
243	10	432	6	663	21	943	26
244	10	433	6	669	21	944	26

CNAE	CPH/91	CNAE	CPH/91	CNAE	CPH/91	CNAE	CPH/91
245	10	434	6	671	20	945	26
246	10	435	6	672	18	946	26
247	10	436	6	679	20	951	26
249	10	437	6	711	22	952	29
251	10	439	6	712	22	953	29
252	10	441	6	721	22	954	29
253	10	442	6	722	22	955	29
254	10	451	6	723	22	959	29
255	10	452	6	724	16	961	29
311	11	453	6	729	22	962	29
312	12	454	6	731	22	963	29
313	12	455	6	732	22	964	29
314	12	456	6	733	22	965	29
315	12	461	7	741	22	966	29
316	12	462	7	742	22	967	29
319	12	463	7	751	22	968	29
321	12	464	7	752	22	969	29
322	12	465	7	753	22	971	29
323	12	466	7	754	22	972	29
324	12	467	7	755	22	973	29
325	12	468	15	756	22	979	29
326	12	471	8	761	22	980	28
329	12	472	8	762	22	990	29
330	13	473	8	811	23		

Annex 2

Resultats de la variable activitat

1. Matriu de correspondències entre els codis precodificats i els codificats. % horitzontals.

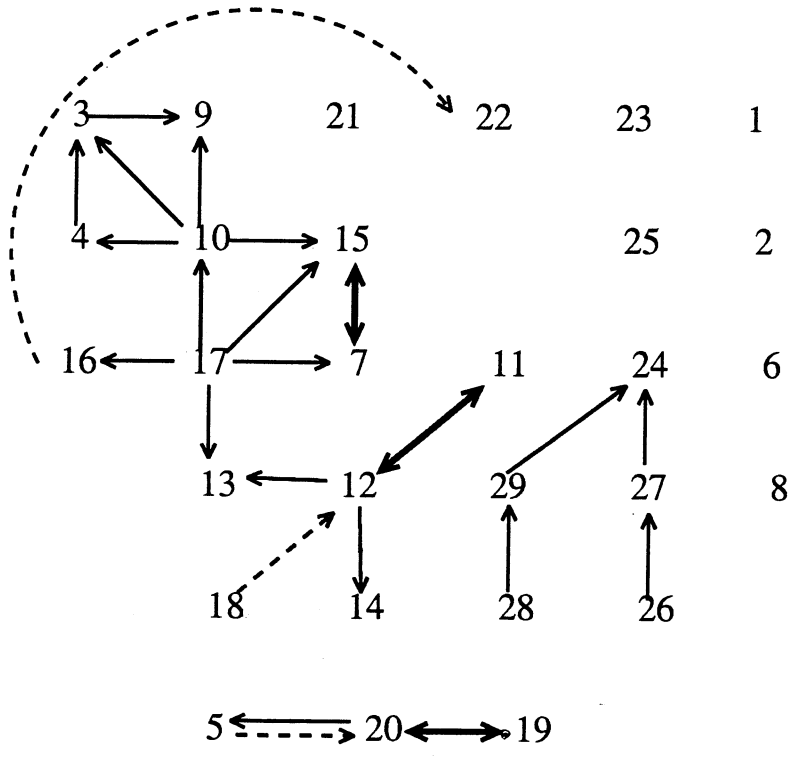
	Precodificat																												
Codificat	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
1	95,87	0,16	0,04	0,03	0,47	0,04	0,21	0,01	0,03	0,08	0,07	0,12	0,91	0,01	0,03	0,04	0,15	0,02	0,21	0,22	0,17	0,04	0,01	0,04	0,01	0,03	0,32	0,04	0,62
2	4,52	89,24	0,13	0,01	1,43	0	0,13	0,01	0,07	0,01	0	0,12	0,78	0	0,08	0,01	0,03	0,04	0,29	2,58	0,03	0,12	0,03	0,00	0,00	0,01	0,22	0,00	0,13
3	0,12	0,03	44,89	45,52	0,17	0,21	0,14	0,02	0,75	4,9	1,17	0,24	0,14	0,09	0,09	0,49	0,33	0,07	0,19	0,16	0,00	0,07	0,02	0,07	0,02	0,03	0,03	0,00	0,03
4	0,26	0,12	1,27	78,74	1,01	0,4	1,3	0	0,26	9,37	0,89	0,35	0,03	0,17	0,23	0,14	4,18	0,03	0,14	0,00	0,00	0,87	0,00	0,00	0,00	0,00	0,09	0,00	0,14
5	2,31	0,14	0,11	0,05	78,2	0,51	0,1	0,15	0,29	1,2	0,02	0,15	0,07	0,04	0,26	0,09	0,38	0,16	2,68	11,04	0,94	0,48	0,01	0,02	0,07	0,08	0,06	0,03	0,36
6	0,03	0,03	0,03	0,04	0,18	95,3	0,16	0,15	0,04	0,44	0,03	0,12	0,01	0,03	0,28	0,14	0,79	0,06	0,31	1,31	0,01	0,02	0,00	0,01	0,01	0,04	0,01	0,03	0,38
7	0,24	0,07	0,03	0,05	0,11	0,27	42,35	0,17	0,04	0,47	0,12	3,39	0,11	0,1	7,12	0,20	43,51	0,13	0,30	0,78	0,01	0,13	0,01	0,07	0,03	0,00	0,05	0,02	0,10
8	0,06	0,08	0,11	0,03	0,17	0,56	0,69	90,1	0,11	0,43	0,1	0,39	0,15	0,04	0,58	0,03	0,73	0,20	0,45	0,42	0,01	0,54	0,02	0,50	0,07	0,01	0,16	0,07	3,17
9	0,22	0,09	30,97	0,58	2,23	0,22	0,18	0,13	50,89	5,3	0,22	0,22	0,04	0,09	0,22	2,32	0,40	3,70	0,85	0,09	0,00	0,85	0,00	0,00	0,00	0,00	0,13	0,00	0,04
10	0,26	0,08	0,15	0,58	0,31	1,71	0,28	0,45	0,28	77,05	0,66	0,91	0,14	0,13	2,18	0,23	9,25	0,26	1,10	1,78	0,02	0,14	0,01	0,04	0,03	1,03	0,12	0,06	0,72
11	0,11	0,02	0,03	2,44	0,02	0,22	0,18	0,15	0,06	1,65	72,38	17,18	0,86	0,98	0,95	0,97	0,94	0,25	0,32	0,04	0,02	0,05	0,03	0,01	0,01	0,02	0,06	0,02	0,02
12	0,14	0,05	0,03	1,01	0,16	0,97	0,66	0,57	0,03	0,92	14,19	62,2	3,61	2,82	0,99	1,26	3,46	4,06	0,59	1,23	0,08	0,19	0,02	0,10	0,01	0,02	0,04	0,03	0,55
13	0,09	0,12	0,08	0,07	0,04	0,29	0,11	1,45	0,05	2,24	1,14	8,89	50,64	1,12	1,79	4,01	19,09	0,88	1,31	2,41	0,02	0,39	0,06	0,75	0,05	1,58	0,15	0,02	1,14
14	0,06	0,19	0,02	0,11	0,07	0,47	0,27	0,06	0,03	0,49	1,05	9,42	0,66	74,05	1,44	0,69	0,79	7,29	0,87	0,60	0,02	0,91	0,03	0,07	0,03	0,00	0,05	0,02	0,23
15	0,1	0,08	0,05	0,06	1,01	4,27	9,19	3,17	0,14	10,92	0,68	3,78	1,68	0,73	49,46	0,21	5,62	0,78	1,30	4,08	0,04	0,31	0,10	0,35	0,08	0,03	0,05	0,03	1,70
16	0,14	0,03	2,14	0,21	0,84	0,3	0,1	0,04	1,09	2,37	0,12	0,93	4,78	0,1	0,23	70,97	11,94	0,33	0,40	0,63	0,01	0,67	0,02	0,11	0,00	0,31	0,84	0,04	0,29
17	0,09	0,03	0,04	0,16	0,03	0,09	0,19	0,06	0,02	0,49	0,09	0,75	0,34	0,27	0,93	1,03	92,28	0,16	0,47	0,39	0,03	0,49	0,01	0,33	0,01	0,01	0,47	0,24	0,50
18	0,06	0,02	0,21	0,1	0,07	0,2	0,05	0,17	0,4	0,26	0,52	6,66	0,22	5,49	0,23	1,15	1,20	77,30	1,44	2,72	0,13	0,40	0,04	0,04	0,01	0,01	0,25	0,09	0,57
19	0,91	0,3	0,13	0,4	4,61	2,29	1,32	2,85	0,39	2,11	1,04	2,54	0,95	0,23	0,79	0,72	1,47	1,88	59,19	8,61	0,21	3,96	0,45	0,93	0,02	0,29	0,29	0,04	1,10
20	0,49	0,4	0,03	0,03	4,71	2,59	0,31	0,93	0,92	0,73	0,1	0,93	0,99	0,13	0,81	0,13	0,71	2,17	9,93	68,62	0,72	0,23	0,05	0,22	0,05	1,85	0,09	0,08	1,05
21	0,11	0,07	0,01	0,01	1	0,02	0,02	0,01	0,05	1,39	0,07	0,08	0	0,01	0,02	0,01	0,08	0,03	0,14	0,62	93,54	0,13	0,02	0,16	0,26	0,14	0,11	0,26	1,61
22	0,1	0,31	0,04	0,04	0,5	0,21	0,19	0,12	0,03	0,2	0,09	0,42	0,38	0,45	0,07	5,23	0,62	1,67	1,00	0,43	0,34	82,45	0,15	0,51	0,03	0,08	2,69	0,13	1,53
23	0,03	0,05	0,11	0,01	0,05	0,12	0,29	0,17	0,01	0,03	0,01	0,05	0,07	0,01	0,01	0,01	0,02	0,02	0,04	0,06	0,05	0,14	95,47	1,01	0,12	0,88	0,94	0,15	0,10
24	0,57	0,22	0,06	0,29	0,18	0,25	0,41	1,86	0,07	0,3	0,3	0,64	2,24	0,14	0,1	0,41	3,54	0,80	1,52	2,24	0,98	4,21	3,21	52,59	0,56	0,60	11,55	1,54	8,65
25	0,19	0,06	0,03	0,64	0,09	0,07	0,03	0,04	0	0,17	0,01	0,04	0,02	0,03	0,08	0,02	0,11	0,02	0,03	0,03	0,10	0,08	0,11	0,38	93,19	0,77	2,31	0,29	1,07
26	0,46	0,33	0,02	0,05	0,03	0,11	0,06	0,01	0	0,37	0,03	0,08	0,03	0,01	0,04	0,13	0,06	0,02	0,05	0,16	0,42	0,18	0,14	0,20	3,70	87,64	4,18	0,47	1,02
27	0,16	0,08	0,01	0,04	0,15	0,27	0,47	0,05	0,01	0,05	0,2	0,06	0,03	0,01	0,04	0,10	0,41	0,03	0,12	0,24	0,14	0,89	0,44	0,94	0,61	10,30	83,22	0,18	0,76
28	0,14	0,06	0	0	0	0,09	0	0,11	0,01	0,1	0	0,76	0,01	0,01	0	0,01	0,12	0,33	0,08	0,37	0,37	0,23	0,17	0,21	0,69	0,57	0,16	93,26	2,11
29	0,32	0,15	0,16	0,04	0,08	1,19	0,14	1,58	0,15	0,67	0,08	0,87	0,44	0,05	0,16	0,23	0,82	0,45	0,61	1,92	0,97	1,21	0,30	3,76	1,16	1,82	3,77	20,27	56,65

2. Quadre resum

Codi	*Diag*	*Nmijo*	*Mijo*$_1$	*Mijo*$_2$	*Mijo*$_3$	*Donant*$_1$	*Donant*$_2$	*Donant*$_3$	*Acum*	*Receptor*
01	95,87	0	,	,	,	.	.	.	1	5
02	89,24	0	,	,	,	.	.	.	3	4
03	44,89	2	1,01	0,11	,	4	10	.	3	1
04	78,74	1	0,12	,	,	10	.	.	6	3
05	78,20	1	0,14	,	,	20	.	.	5	3
06	95,30	0	,	,	,	.	.	.	1	5
07	42,35	2	1,03	0,17	,	17	15	.	4	1
08	90,10	0	,	,	,	.	.	.	5	4
09	50,89	2	0,61	0,10	,	3	10	.	6	2
10	77,05	1	0,12	,	,	17	.	.	8	3
11	72,38	1	0,24	,	,	12	.	.	6	3
12	62,20	1	0,23	,	,	11	.	.	11	2
13	50,64	2	0,38	0,18	,	17	12	.	12	2
14	74,05	1	0,13	,	,	12	.	.	7	3
15	49,46	3	0,22	0,19	0,11	10	7	17	11	1
16	70,97	1	0,17	,	,	17	.	.	8	3
17	92,28	0	,	,	,	.	.	.	5	4
18	77,30	0	,	,	,	.	.	.	7	3
19	59,19	1	0,15	,	,	20	.	.	16	2
20	68,62	1	0,14	,	,	19	.	.	12	2
21	93,54	0	,	,	,	.	.	.	2	4
22	82,45	0	,	,	,	.	.	.	7	3
23	95,47	0	,	,	,	.	.	.	1	5
24	52,59	2	0,22	0,16	,	27	29	.	14	2
25	93,19	0	,	,	,	.	.	.	2	4
26	87,64	0	,	,	,	.	.	.	3	4
27	83,22	1	0,12	,	,	26	.	.	4	3
28	93,26	0	,	,	,	.	.	.	2	4
29	56,65	1	0,36	,	,	28	.	.	12	2

3. Graf de relacions entre codis[3]

Graf de valors donants i receptors per a la variable activitat



[3]La direcció de la fletxa indica que el codi de sortida dóna al codi de destí de forma significativa.

ENGLISH SUMMARY:

ANALYSIS OF AUTOMATIC CODING VALUES VERSUS RESPONDENT'S VALUES. APPLICATION FOR ACTIVITY AND OCCUPATION

Miquel Delgado Alzamora and Josep Anton Sánchez Cepeda

In the Census of Population 1991, the questions related to «Occupation or profession» and to «Activity of the stablishment or place of work» had to be answered twice: the first question had to be answered with a description and the second one by coding this description according to a list supplied with the questionnaire (this will be called the respondent's code). In the case of «Occupation» the list had 20 items and the list of «Activity» had 29 items. In the Institut d'Estadística de Catalunya both the description and the respondent's code were recorded in order to be compared.

At the stage of information processing a process of automatic coding of the descriptions made by respondents is carried out; for «Occupation» answer, a 3 digit number according to the National Classification of Occupations (CNO) is got, and for «Activity» answer a 3 digit number according to the National Classification of Economic Activities (CNAE) is got. These three digit numbers are translated into their corresponding codes of the 20 and 29 item lists. Those resulting codes will be called automatic codes. So, two coded answers are available for each question: the first one, precoded by the respondant, and the second one, the result of the automatic coding. This situation leads us to the study of the double response in order to check if the respondents make mistakes in coding their own responses or they provide good codings and to detect the confusing items. It is supposed that the description supplied by the respondent is the most reliable; so, the automatic code is taken as the valid value versus the respondent's code. This assumption can be made because in the automatic coding process, the dictionary with the variants has been studied deeply.

The aim of this study is to check the values for each variable in order to discriminate answers leading to no confussion and to group the more mixed up answers. In order to do this, a square matrix A is defined; the element a_{ij} indicates the number of answers with automatic code value «i» and respondent's code value «j». The diagonal gives the number of matches and the rest, the mismatches or residues. Some parameters are defined for each value of each variable; their values will determine which codes are to be grouped or separated. These parameters are:

1. Diag: rate of well coded values; calculated as the weight of the diagonal over the row.

2. Acum: number of necessary values of the row to achieve 95% of success.
3. MIJO: for each value a_{ij}, the weight of the residue over the diagonal a_{ii}.
4. Donant distribution: for each row, values of MIJO distribution bigger than 0.05; they will be the donant values. A graph is made to a better understanding of the values.
5. Nmijo: for each row, the number of residues exceeding the theshold.