Qüestiió

Quaderns d'Estadística i Investigació Operativa

Any 1996, volum 20, núm. 3
Segona època

Entitats patrocinadores:

Universitat de Barcelona
Universitat Politècnica de Catalunya
Institut d'Estadística de Catalunya

Entitat col·laboradora:

International Biometric Society

Generalitat de Catalunya
Institut d'Estadística de Catalunya

EDITORIAL

Amb la publicació del present Número 3 es completa l'edició del Volum 20 de «Qüestiió» de l'any 1996, on s'hi apleguen vuit originals repartits en les quatre seccions temàtiques de la revista. Com és costum en l'editorial del darrer número de cada Volum, «Qüestiió» fa balanç de l'activitat relativa a la presentació i avaluació d'originals enregistrada a les (quatre) seccions. En aquest sentit, enguany l'evolució dels articles ha estat la següent:

- articles sotmesos (gener 1996 – febrer 1997): 36
- articles en revisió: 26
- articles acceptats en espera de publicació: 9
- articles rebutjats (1996): 5

Comentari de les seccions «Estadística», «Investigació Operativa» i «Biometria»

En relació als originals inclosos a «Estadística», l'article «El contraste de estabilidad predictiva como instrumento para discriminar entre modelos», d'A. Aznar i M.I. Ayuda, versa sobre la selecció entre models lineals, considerant un contrast d'estabilitat predictiva; a «Factor analysis and information criteria», de M. Costa, s'investiga la determinació del nombre de factors en anàlisi factorial exploratori, utilitzant diversos criteris que es comparen mitjançant una simulació; a continuació, «Estimación del parámetro *b* de Richter a partir de las medidas imprecisas de intensidad epicentral», de S. Forcada i J.J. Egozcue, és una contribució a l'estimació a partir de mesures sísmiques, de naturalesa imprecisa, mitjançant la immersió a un espai de Hilbert i que es compara amb altres mètodes només vàlids en el cas de dades precises; finalment, «La combinació de tècniques de geometria diferencial amb anàlisi multivariant clàssica: una aplicació a la caracterització de les comarques catalanes», de S. Vives i À. Villarroya, aplica tècniques clàssiques, com l'anàlisi de correspondències, juntament amb noves eines basades en la geometria diferencial, per tal de representar variables i poblacions, amb una aplicació a la classificació de les comarques de Catalunya.

En relació a «Investigació Operativa», l'article «Una variante del algoritmo de Ahuja-Orlin para los problemas de flujo máximo: experiencias computacionales y comparaciones», de A.A. Sedeño i C. González, obté una variant que té la mateixa complexitat computacional teòrica, però que estalvia temps de processament informàtic i, per tant, té avantatges sobre l'algorisme descrit. Per últim, la secció de «Biometria» conté l'article «Sensibilidad del análisis de la estructura propia en la detección de colinealidad», de R. Estarelles i J. Prieto, el qual constitueix una contribució a la diagnosi de la colinealitat utilitzant índexs basats en l'estructura pròpia de la matriu de predictors, que es verifica amb dades psicomètriques reals.

Carles Cuadras, Editor Executiu

Comentari de la secció «Estadística Oficial» i altres apartats

La secció d'Estadística Oficial aborda monogràficament algunes tècniques per a la preservació del secret estadístic, actualitzant el tractament iniciat en el Volum 17 (1993) i precedint altres contribucions metodològiques que es preveuen publicar en el Volum 21 (1997). En aquest número s'hi reflecteix l'aportació de dos especialistes a l'entorn de l'Institut d'Estadística de Catalunya, que ofereixen una aproximació complementària. El treball «Privacy homomorphisms for statistical confidentiality», de J. Domingo, demostra que la combinació de tècniques criptogràfiques amb homomorfismes «de privacitat», juntament amb mètodes més clàssics (introducció de perturbacions aleatòries, supressió i remostratge) permetria la computació d'estadístiques exactes sobre macro/microdades sotmeses a processos d'encriptació. En segon lloc, l'article d'A. Garín, «Càlcul del risc de revelació estadística en la difusió de microdades ...», descriu les experiències recents en la mesura de la probabilitat d'identificació d'individus d'una mostra anonimitzada del Cens de Població 1991 amb procediments de submostratge (és a dir, mètodes basats en la «reducció» de dades més que en la «modificació» de les mateixes), per tal d'estimar el percentatge de «poblacions úniques», component bàsic del risc de revelació.

En darrer lloc, l'apartat dedicat a «Cursos i congressos d'estadística» referencia, per ordre cronològic, la impartició de tres cursos corresponents a la 3a. edició de la «Applied Statistics Week» (Barcelona, 27 juny - 3 juliol 1997), dedicada al disseny i l'anàlisi de dades d'enquesta, a càrrec de la Universitat Pompeu Fabra i amb la cooperació de l'Institut d'Estadística de Catalunya. D'altra banda, s'anuncia el «IV International Meeting of Multidimensional Data Analysis» (NGUS'97), del New Spad Group Users (Bilbao, 10-12 setembre 1997), que organitza la Universitat del País Basc amb la col·laboració d'EUSTAT, CISIA i altres organismes. En tercer lloc, s'avança la convocatòria de la «VI Conferencia Española de Biometría» (Còrdova, 21-24 setembre 1997), organitzada pel Departament d'Estadística de la Universitat de Còrdova i sota la tutela de la «Región Española de la International Biometric Society».

Enric Ripoll, Editor Executiu

Estadística

QÜESTIIÓ, Vol 20, 3 pp. 385-407, 1996

EL CONTRASTE DE ESTABILIDAD PREDICTIVA COMO INSTRUMENTO PARA DISCRIMINAR ENTRE MODELOS

ANTONIO AZNAR GRASA

Mª ISABEL AYUDA BOSQUE*

Departamento de Análisis Económico

En este trabajo se considera el contraste de estabilidad predictiva como un contraste de especificación en un marco en el que se trata de elegir entre dos modelos. Se propone una modificación del estadístico habitual consistente en sustituir la varianza del modelo que aparece en el denominador, por la varianza estimada obtenida a partir del modelo más amplio de los que se comparan en el caso de modelos anidados o VAR, o por la varianza de un modelo general que anide los dos modelos en el caso de no anidados.

The predictive stability test to choose between econometric models

Keywords: Estabilidad predictiva, selección, modelos econométricos.

* Antonio Aznar Grasa y Mª Isabel Ayuda Bosque. Departamento de Análisis Económico. Facultad de CC. EE. y EE. Doctor Cerrada 1. 50005 ZARAGOZA. Fax: 761770.

- Article rebut el juny de 1995.

- Acceptat el novembre de 1995.

1. INTRODUCCIÓN

El papel del contraste de estabilidad predictiva en el proceso de evaluación de un modelo econométrico ha sido reiteradamente destacado en la literatura. En muchos trabajos y libros de texto pueden encontrarse aproximaciones a cómo definir los estadísticos y cómo interpretar los resultados obtenidos. Ver, por ejemplo, Chow (1960), Fisher (1970) y Dhrymes *et al.* (1972).

Normalmente, la hipótesis nula que se considera es que ciertos parámetros del modelo son los mismos en dos períodos muestrales no solapados y diferentes.

En este trabajo, siguiendo la sugerencia contenida en Pesaran *et al.* (1985) se va a considerar el contraste de estabilidad predictiva como un contraste general de especificación, en un marco, en el que se trata de elegir entre dos modelos. Se propone modificar el estadístico que sirve de base para el contraste, tal como se define habitualmente, variando el estimador de la varianza del modelo que aparece en el denominador de dicho estadístico. Para una amplia gama de modelos que incluye anidados, no anidados y modelos VAR se demuestra que el cambio propuesto supone una mejora sustancial en el proceso de evaluación de los modelos que se comparan.

El trabajo se organiza como sigue. En la Sección 2 se formula el problema y se introducen los elementos básicos para su tratamiento. Los modelos anidados son considerados en la Sección 3, los no anidados en la Sección 4 y la Sección 5 está dedicada a los modelos VAR. En la sección 6 se lleva a cabo una serie de experimentos de Monte Carlo con el objeto de comparar el comportamiento del estadístico de estabilidad predictiva, propuesto en este trabajo para los modelos considerados en las secciones anteriores, con el estadístico de estabilidad predictiva definido de forma habitual. El trabajo termina con unas conclusiones.

2. ELEMENTOS ESENCIALES DEL PROBLEMA. MODELOS E HIPÓTESIS

En general, el contraste de estabilidad predictiva se ha planteado considerando un modelo aislado en dos períodos muestrales diferentes.

Limitándonos a un caso muy simple, consideremos el siguiente Modelo Lineal General:

$$M_1 : y = X\beta + u_1 \tag{1}$$

en donde y es un vector $T \times 1$ de observaciones de la variable endógena; X es una matriz $T \times k$ de las observaciones de k variables explicativas; β es un vector de k parámetros y u_1 es un vector de T perturbaciones aleatorias, todas ellas distribuidas según una distribución normal con media cero y varianza σ_1^2.

Para un período muestral diferente, de T_1 observaciones, un modelo con las mismas variables explicativas que el escrito en (1) toma la forma siguiente:

$$y_p = X_p\beta_p + u_{1p} \tag{2}$$

en donde ahora y_p es un vector con T_1 elementos, X_p es una matriz $T_1 \times k$, β_p es de orden $k \times 1$ y u_{1p} es un vector de perturbaciones aleatorias con media cero y varianza σ^2_{1p}.

La hipótesis nula del contraste de estabilidad predictiva es la siguiente:

$$H_0 : \beta = \beta_p \quad \text{y} \quad \sigma_1^2 = \sigma^2_{1p} \tag{3}$$

Para obtener un estadístico que permita contrastar esta hipótesis nula se parte de la siguiente expresión:

$$e_p = y_p - X_p\hat{\beta} \tag{4}$$

en donde:

$$\hat{\beta} = (X'X)^{-1}X'y \tag{5}$$

Bajo la hipótesis nula escrita en (3) se obtiene que:

$$e_p \sim \mathrm{N}[0, \sigma_1^2(I_{T_1} + X_p(X'X)^{-1}X_p')] \tag{6}$$

Por lo tanto,

$$K = \frac{e_p'[I_{T_1} + X_p(X'X)^{-1}X_p']^{-1}e_p}{\sigma_1^2} \sim \chi^2_{T_1} \tag{7}$$

Para hacer operativo este contraste, hay que estimar σ_1^2 de forma tal que la expresión que aparezca en el denominador sea independiente del numerador. La práctica habitual ha consistido en utilizar el estimador MCO de σ_1^2 obtenido a partir de (1) que podemos escribir como:

$$\hat{\sigma}_1^2 = \frac{\hat{u}_1'\hat{u}_1}{T-k} \tag{8}$$

en donde:

$$\hat{u}_1 = y - X\hat{\beta} \tag{9}$$

El estadístico resultante puede escribirse como:

$$F = \frac{e_p'[I_{T_1} + X_p(X'X)^{-1}X_p']^{-1}e_p/T_1}{\hat{\sigma}_1^2} \tag{10}$$

que sigue una distribución F con T_1 y $T-k$ grados de libertad.

Pero, como ya hemos indicado, nosotros en este trabajo vamos a utilizar este contraste, no para evaluar un modelo aisladamente, sino como un instrumento para comparar dos modelos. Por esta razón, como alternativa al modelo escrito en (1) consideraremos el siguiente:

$$M_2 : y = Z\gamma + u_2 \tag{11}$$

en donde y sigue siendo un vector con T elementos, Z es una matriz $T \times q$ de observaciones de q variables explicativas, γ es un vector de q parámetros y u_2 es un vector con T variables aleatorias con media cero y todas ellas con varianza igual a σ_2^2.

Supondremos que el Proceso Generador de Datos (PGD) puede ser tanto M_1 como M_2; si el PGD es M1 entonces se tiene que:

$$E\hat{\sigma}_1^2 = \sigma_1^2 \tag{12}$$

por lo que el contraste basado en (10) en principio debería funcionar bien.

Pero si el PGD es M_2 entonces se tiene que:

$$E\hat{\sigma}_1^2 = \sigma_2^2 + \frac{\gamma Z' M_x Z\gamma}{T-k} \tag{13}$$

en donde:

$$M_x = I - X(X'X)^{-1}X'$$

A partir de aquí se ve que si el PGD es M_2 el resultado escrito en (7) ya no es válido ya que la esperanza de e_p es diferente de cero y su varianza viene dada por:

$$\text{Var}\,(e_p) = \sigma_2^2[I_{T_1} + X_p(X'X)^{-1}X_p'] \tag{14}$$

En este caso, la utilización de (10) puede llevar a resultados contradictorios ya que, como puede verse en (13), la varianza que aparece en el denominador del estadístico tenderá a ser grande en relación a la verdadera varianza en términos de esperanzas. La implicación de este resultado es que aunque el modelo esté mal especificado y se generen errores de predicción grandes, se puede concluir aceptándolo como bien

especificado porque se compara con una estimación de la varianza que sobrestima el verdadero valor de dicha varianza.

El objetivo de este trabajo, como ya hemos indicado, es el de proponer un contraste de estabilidad predictiva que evite este problema de sobrestimación de la varianza.

Para lograr este objetivo consideraremos un tercer modelo que escribiremos, para el período muestral con T observaciones, como:

$$y = W\delta + v_1 \tag{15}$$

en donde W es una matriz $T \times s$ de observaciones de s variables explicativas, δ es un vector de s parámetros y v_1 es un vector de T perturbaciones aleatorias con media cero y varianza $\sigma^2_{v_1}$.

A partir de este modelo definiremos un estimador de la varianza del PGD, sea éste M_1 ó M_2. El estimador lo escribiremos como:

$$\hat{\sigma}^2 = \frac{y' M_w y}{T - s} \tag{16}$$

en donde:

$$M_w = I - W(W'W)^{-1}W'$$

A continuación, en lugar del estadístico escrito en (10), para llevar a cabo el contraste de estabilidad predictiva, en este trabajo se propone utilizar el siguiente estadístico:

$$F' = \frac{e'_p[I_{T_1} + X_p(X'X)^{-1}X'_p]^{-1}e_p}{T_1 \cdot \hat{\sigma}^2} \tag{17}$$

El modelo escrito en (15) debe garantizar que el estimador $\hat{\sigma}^2$ es independiente del numerador de (17) y que es un estimador insesgado de la varianza del verdadero PGD, sea éste M_1 ó M_2.

Veamos en las secciones siguientes la forma que toma este modelo según consideremos modelos anidados, no anidados o modelos VAR.

3. MODELOS ANIDADOS

En esta sección suponemos que M_1 está anidado en M_2 de forma que se puede escribir:

$$Z = [X, X_1] \tag{18}$$

en donde X_1 es una matriz de $T \times (q-k)$ observaciones de $(q-k)$ variables, siendo en este caso $s = q$.

Haciendo $\gamma' = [\beta', \gamma_1']$, el modelo M_2 escrito en (11) puede escribirse como:

$$y = X\beta + X_1\gamma_1 + v_1 \tag{19}$$

Este es el modelo que, para modelos anidados, se propone para estimar la varianza del PGD desconocido. Es decir, en esta sección suponemos que el modelo (15) es el modelo escrito en (19), por lo que: $W = [X, X_1]$.

Suponiendo homogeneidad de parámetros entre los dos períodos muestrales, el modelo escrito en (19) para el período con T_1 observaciones puede escribirse como:

$$y_p = X_p\beta + X_{1p}\gamma_1 + v_p \tag{20}$$

Si el PGD es M_1, a partir de (4) podemos escribir e_p como:

$$\begin{aligned} e_p &= X_p\beta + u_{1p} - X_p(X'X)^{-1}X'y = u_{1p} - X_p(X'X)^{-1}X'u_1 = \\ &= [I_{T_1}, -X_p(X'X)^{-1}X']u_1^* = Au_1^* \end{aligned} \tag{21}$$

en donde:

$$u_1^* = \begin{pmatrix} u_{1p} \\ u_1 \end{pmatrix}$$

Por otro lado, también se tiene que:

$$M_w y = M_w u_1 = [0 \quad M_w]u_1^* = Bu_1^* \tag{22}$$

A partir de esta expresión se obtiene que:

$$Ey'M_w y = \sigma_1^2(T - s) \tag{23}$$

por lo que el estimador escrito en (16) es un estimador insesgado de σ_1^2. Por otra parte, como puede verse en Pesaran *et al.* (1985) se cumple que:

$$(I_{T_1} + X_p(X'X)^{-1}X_p')^{-1} = (AA')^{-1}$$

por lo que el numerador de (7) puede escribirse como:

$$e_p'[I_{T_1} + X_p(X'X)^{-1}X_p']^{-1}e_p = u_1^{*'}A'(AA')^{-1}Au_1^* \tag{24}$$

Por último, por cumplirse que $BA' = 0$ se tiene que el estimador escrito en (16) es independiente del numerador de (17).

Si el PGD es M_2 entonces el vector e_p puede escribirse como:

$$\begin{aligned} e_p &= X_p\beta + X_{1p}\gamma_1 + v_p - X_p\hat{\beta} = \\ &= X_p\beta + X_{1p}\gamma_1 + v_p - X_p\beta - X_p(X'X)^{-1}X'X_1\gamma_1 - X_p(X'X)^{-1}X'v_1 = \\ &= X_{1p}\gamma_1 + v_p - X_p(X'X)^{-1}X'X_1\gamma_1 - X_p(X'X)^{-1}X'v_1 = \\ &= X_{1p}\gamma_1 - X_p(X'X)^{-1}X'X_1\gamma_1 + Av^* \end{aligned} \tag{25}$$

en donde:

$$v^* = \begin{pmatrix} v_p \\ v_1 \end{pmatrix}$$

A partir de esta expresión se ve que la esperanza de (25) es diferente de cero y que $e_p - Ee_p = Av^*$.

Por otra parte, se tiene que:

$$M_w y = M_w v_1 = Bv^* \tag{26}$$

Por lo que, también en este caso, (16) será un estimador insesgado de la varianza del PGD y si se emplea este estimador en (10) en lugar de $\hat{\sigma}_1^2$, las expresiones en el numerador y denominador seran independientes.

4. MODELOS NO ANIDADOS

En este caso, el modelo escrito en (15) que se propone para estimar la varianza del PGD toma la forma siguiente:

$$y = X\beta + Z\gamma + v_1 \tag{27}$$

en donde las matrices X y Z son las introducidas en los modelos escritos en (1) y (11), respectivamente. Por lo tanto:

$$W = [X, Z]$$

con $s = k + q$.

Si el PGD es M_1 entonces se tiene que:

$$e_p = X_p\beta + u_{1p} - X_p\beta - X_p(X'X)^{-1}X'u_1 = Au_1^* \tag{28}$$

Por otra parte, también se tiene:

$$M_w y = M_w u_1 = Bu_1^* \tag{29}$$

en donde:

$$B = [0, M_w]$$

Estos resultados muestran que el estimador de la varianza del PGD que aparece en el denominador de (17) es insesgado y que el numerador y denominador de dicha expresión son independientes entre sí.

Si el PGD es M_2 entonces el vector de errores de predicción de M_1 puede escribirse como:

$$\begin{aligned} e_p &= Z_p\gamma + u_{2p} - X_p(X'X)^{-1}X'Z\gamma - X_p(X'X)^{-1}X'u_2 = \\ &= Z_p\gamma - X_p(X'X)^{-1}X'Z\gamma + Au_2^* \end{aligned} \tag{30}$$

A partir de esta expresión se ve que $Ee_p \neq 0$ y que:

$$e_p - Ee_p = Au_2^* \tag{31}$$

Por otra parte, también se tiene que:

$$M_w y = M_w u_2 = Bu_2^* \tag{32}$$

por lo que $\hat{\sigma}^2$ será un estimador insesgado de σ_2^2 y (31) y (32) serán independientes.

Resultados análogos se obtienen cuando el modelo analizado es el M_2.

5. MODELOS VAR

En principio, la selección entre varios modelos VAR con diferentes órdenes no es más que un caso particular de elegir entre modelos anidados. Pero este tipo de modelos presenta unas peculiaridades que demandan un tratamiento diferenciado.

A lo largo de toda esta sección adoptaremos la notación utilizada por Lütkepohl (1991).

El modelo VAR de orden p lo escribiremos así:

$$y_t = \nu + A_1 y_{t-1} + \cdots + A_p y_{t-p} + u_t \tag{33}$$

en donde y_t es el vector $k \times 1$ de la observación $t-$ésima de las k variables, ν es un vector de $k \times 1$ constantes, las A_i son matrices $k \times k$ de coeficientes y u_t es un vector de k perturbaciones aleatorias con $Eu_t = 0$, $E(u_t\, u_t') = \Sigma_u$ y $E(u_t\, u_s') = 0$ para $s \neq t$.

Para contrastar la estabilidad predictiva, Lütkepohl (1989 y 1991) propone utilizar el siguiente estadístico:

$$\overline{\lambda}_{T_1} = \frac{T\sum_{i=1}^{T_1} \hat{u}'_{T+i}\,\hat{\Sigma}_u^{-1}\,\hat{u}_{T+i}}{(T+kp+1)kT_1} \tag{34}$$

en donde:

$$\hat{\Sigma}_u = \frac{1}{T-kp-1}\,\Sigma\,\hat{u}\hat{u}'_t \tag{35}$$

siendo:

$$\hat{u}_t = y_t - \hat{\nu} - \hat{A}_1 y_{t-1} - \cdots - \hat{A}_p y_{t-p}$$

en donde $\hat{\nu}, \hat{A}_1, \ldots, \hat{A}_p$ son los estimadores máximo-verosímiles de los parámetros correspondientes.

Lütkepohl propone determinar la región crítica comparando el valor del estadístico escrito en (34) con el valor crítico correspondiente a la distribución F con $k \cdot T_1$ grados de libertad en el numerador y $T-kp-1$ grados de libertad en el denominador, después de adoptar un nivel de significación.

Nosotros, en este trabajo, siguiendo con la línea comentada en las secciones anteriores proponemos estimar (35) utilizando el modelo VAR de mayor orden entre todos los que se comparan.

6. ESTUDIO DE MONTE CARLO

En este apartado se llevan a cabo una serie de experimentos de Monte Carlo con el objeto de analizar el comportamiento del estadístico de estabilidad predictiva, propuesto en este trabajo, como un contraste general de especificación en un marco en el que se trata de elegir entre dos modelos, ya sean anidados, no anidados o modelos VAR. El análisis se realiza comparando los resultados que se obtienen con el estadístico propuesto en este trabajo y los que se obtendrían en el caso de utilizar el estadístico normalmente utilizado en la literatura.

Siguiendo con el marco adoptado en las anteriores secciones, los experimentos realizados hacen referencia a:

- Modelos anidados
- Modelos no anidados
- Modelos VAR

6.1. Modelos anidados

Dentro de este apartado vamos a considerar dos casos; en el primero de ellos el PGD es el modelo amplio, ($k = 3$) mientras que, en el segundo, el PGD es el modelo más restringido ($k = 2$).

Los modelos que se han utilizado de forma sucesiva como PGD han sido los siguientes:

$$M_1: \quad y_t = 1 + x_{1t} + \beta_2 x_{2t} + u_{1t}$$

$$M_2: \quad y_t = 1 + \beta_1 x_{1t} + u_{2t}$$

donde β_i toma distintos valores: $\beta_i = (0.1, 0.5, 1)$, $(i = 1, 2)$.

La variable x_1 ha sido generada según una distribución normal independiente con media cero y varianza $V(x_1)$ donde ésta toma los siguientes valores:

$$V(x_1) = (0.01, 0.1, 0.5, 1)$$

Los valores de la perturbación aleatoria u_1 han sido generados según una distribución normal con media cero y varianza:

$$\sigma_{u_1}^2 = \frac{\left[\sum_{i=1}^{k} \beta_i^2 \operatorname{Var}(x_i)\right][1 - R^2]}{R^2} \tag{36}$$

donde k es el número de regresores del modelo y R^2 es el coeficiente de determinación del PGD, que lo variamos para analizar la influencia de éste en el estadístico de estabilidad predictiva:

$$R^2 = (0.7, 0.9, 0.99)$$

La variable x_{2t} se ha creado a partir de x_{1t}:

$$x_{2t} = \mu_1 x_{1t} + v_{1t} \tag{37}$$

donde: $v_{1t} \sim \mathrm{N}(0, 1)$

y

$$\mu_1 = \frac{\rho}{[1 - \rho^2]^{1/2}[\operatorname{Var} x_{1t}]^{1/2}} \tag{38}$$

para asegurar que la correlación entre x_{1t} y x_{2t} es ρ. Los valores de ρ considerados han sido:

$$\rho^2 = (0.4, 0.6, 0.9)$$

Los modelos que se comparan son:

$$H_1: \quad y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + u_{1t}$$
$$H_2: \quad y_t = \gamma_0 + \gamma_1 x_{1t} + u_{2t}$$

Para llevar a cabo esta comparación se calcula el contraste de estabilidad predictiva escrito en (17) de tres formas diferentes:

a) Sustituyendo la varianza del denominador por la supuesta en el PGD en el ejercicio de simulación.

b) Estimando la varianza a partir de cada modelo.

c) Estimando la varianza a partir del modelo menos restringido.

Los resultados de b) y c) sólo serán aceptables si tienden a concidir con los obtenidos bajo a).

Los experimentos realizados se han replicado 500 veces para tres tamaños muestrales:

$$T = (50, 100, 500) \quad \text{y} \quad T_1 = 20$$

Se han hecho pruebas con valores distintos de T_1 y los resultados eran similares.

CASO 1:

El PGD ha sido M_1. En la Tabla 1 presentamos algunos de los resultados obtenidos para las distintas combinaciones de los tamaños muestrales, varianza de x_{1t}, valores de β_2, del coeficiente de determinación del PGD (R_2) y del coeficiente de correlación entre x_{1t} y x_{2t}, (ρ). La interpretación de las últimas cinco columnas es la siguiente:

Ai: número de veces que se acepta la hipótesis de estabilidad predictiva del modelo bajo H_i utilizando la varianza real ($i = 1,2$).

Aij: número de veces que se acepta la hipótesis de estabilidad predictiva del modelo en H_i estimando la varianza con el modelo H_j ($i, j = 1,2$).

Como puede observarse en la Tabla 1, estimando σ^2 con el modelo amplio (H_1) (columnas A11 y A21) obtenemos resultados similares a los que se obtienen utilizando la varianza real (columnas A1 y A2), es decir, hay coincidencia entre c) y a). Por el contrario, utilizando la varianza estimada del modelo restringido (H_2) (columna A22) se tiende a aceptar la hipótesis de estabilidad del propio modelo (H_2), cuando con la varianza real (columna A2) se rechazaría. Puede verse también que cuando β_2 es pequeño, por ser los dos modelos parecidos, se tiende a aceptar la estabilidad predictiva en los dos modelos.

Tabla 1

Modelos anidados. PGD: $y_t = 1 + x_{1t} + \beta_2 x_{2t} + u_{1t}$

T	$v(x_1)$	β_2	R^2	ρ	A1	A2	A11	A21	A22
50	0.01	0.1	0.90	0.6	476	3	474	18	500
50	0.01	0.1	0.99	0.6	480	0	471	0	500
50	0.01	1	0.90	0.6	479	0	484	0	500
50	0.01	1	0.99	0.6	480	0	477	0	500
50	1	0.1	0.90	0.6	480	462	485	474	479
50	1	0.1	0.99	0.6	473	86	474	208	441
50	1	1	0.90	0.6	467	0	471	6	424
50	1	1	0.99	0.6	481	0	474	0	459
100	0.01	0.1	0.90	0.6	474	0	477	0	500
100	0.01	0.1	0.99	0.6	478	0	480	0	500
100	0.01	1	0.90	0.6	479	0	472	0	500
100	0.01	1	0.99	0.6	483	0	478	0	500
100	1	0.1	0.90	0.6	470	464	471	458	471
100	1	0.1	0.99	0.6	476	254	469	290	496
100	1	1	0.90	0.6	478	17	478	37	499
100	1	1	0.99	0.6	468	0	461	0	500
500	0.01	0.1	0.90	0.6	468	0	468	0	250
500	0.01	0.1	0.99	0.6	474	0	473	0	24
500	0.01	1	0.90	0.6	479	0	478	0	174
500	0.01	1	0.99	0.6	470	0	472	0	3
500	1	0.1	0.90	0.6	475	456	468	455	477
500	1	0.1	0.99	0.6	474	165	475	167	497
500	1	1	0.90	0.6	480	1	476	2	499
500	1	1	0.99	0.6	475	0	473	0	500

Tabla 2

Modelos anidados. PGD: $y_t = 1 + \beta_1 x_{1t} + u_{2t}$

T	$\nu(x_1)$	β_1	R^2	ρ	A1	A2	A22	A21	A11
50	0.01	0.1	0.90	0.6	477	473	479	481	480
50	0.01	0.1	0.99	0.6	477	476	480	477	477
50	0.01	1	0.90	0.6	484	484	479	481	478
50	0.01	1	0.99	0.6	471	471	476	478	479
50	1	0.1	0.90	0.6	478	480	476	474	474
50	1	0.1	0.99	0.6	476	476	479	477	476
50	1	1	0.90	0.6	474	474	477	475	474
50	1	1	0.99	0.6	475	474	466	468	467
100	0.01	0.1	0.90	0.6	475	475	475	475	474
100	0.01	0.1	0.99	0.6	469	470	473	473	473
100	0.01	1	0.90	0.6	477	477	477	479	477
100	0.01	1	0.99	0.6	469	469	472	470	470
100	1	0.1	0.90	0.6	469	469	476	474	477
100	1	0.1	0.99	0.6	477	474	478	476	478
100	1	1	0.90	0.6	468	468	476	476	477
100	1	1	0.99	0.6	482	480	474	475	472
500	0.01	0.1	0.90	0.6	474	474	475	474	472
500	0.01	0.1	0.99	0.6	471	471	471	472	472
500	0.01	1	0.90	0.6	480	478	480	481	483
500	0.01	1	0.99	0.6	481	480	474	474	473
500	1	0.1	0.90	0.6	475	475	476	476	477
500	1	0.1	0.99	0.6	478	478	477	478	478
500	1	1	0.90	0.6	476	475	478	479	478
500	1	1	0.99	0.6	472	470	475	475	472

CASO 2

En este caso el PGD es M_2, con $\beta_1 = (0.1, 0.5, 1)$. Los resultados obtenidos aparecen en la Tabla 2.

A partir de esta Tabla se puede concluir que cuando genera el modelo restringido, el número de veces que se acepta la hipótesis de estabilidad predictiva para cada modelo, cuando se utiliza la verdadera varianza del PGD, es muy similar al que resulta cuando se utiliza el estimador de dicha varianza definido a partir del modelo amplio, aunque para este caso también son similares los resultados obtenidos a partir de la estimación en el modelo restringido.

En definitiva, como en la práctica real no sabemos cuál es el proceso generador de datos, parece lógico utilizar como estimador de σ^2, el estimador proporcionado por el modelo amplio, ya que de las dos tablas se deduce que utilizando esta varianza estimada se obtienen resultados similares a los que se obtendrían si conociésemos la varianza real.

Se observa también que si para el modelo restringido se utiliza la varianza estimada por el propio modelo, se tiende a aceptar la hipótesis de estabilidad predictiva en la mayor parte de los casos cuando dicho modelo no es el PGD.

6.2. Modelos no anidados

Con el objeto de analizar el comportamiento del nuevo estadístico de estabilidad predictiva en modelos no anidados, se han realizado los experimentos considerando casos en que el PGD tiene un número mayor, igual o menor de parámetros que el modelo alternativo.

Las variables empleadas han sido generadas de forma similar a las utilizadas en el caso de modelos anidados. Para este tipo de modelos, hemos distinguido los tres casos mencionados y debido a que las conclusiones son similares para los tres, aquí presentamos sólamente el caso 1 en el que los dos modelos tienen el mismo número de regresores.

En este caso suponemos que el PGD y el modelo alternativo tienen el mismo número de variables explicativas. El PGD que suponemos es:

$$\text{PGD:} \quad M1: \quad y_t = 1 + \beta_1 x_t + u_{1t}$$

y los modelos que se comparan son:

$$H_1: \quad y_t = \beta_0 + \beta_1 x_t + u_{1t}$$
$$H_2: \quad y_t = \gamma_0 + \gamma_1 z_t + u_{2t}$$

donde:

$$x_t \sim \mathrm{N}(0, v(x))$$

y

$$z_t = \mu x_t + v_{1t} \qquad \text{con} \quad v_{1t} \sim \mathrm{N}(0,1) \tag{39}$$

donde: $$\mu = \frac{\rho}{[1-\rho^2]^{1/2}[\mathrm{Var}\, x_t]^{1/2}}$$

ρ mide la correlación entre x_t y z_t.

En este caso, consideramos una nueva opción que sustituye a la c) anterior consistente en sustituir la varianza por una estimación obtenida a partir de un modelo amplio del que son casos particulares los dos que se comparan. Este modelo amplio lo podemos escribir como:

$$y_t = \alpha_0 + \alpha_1 x_t + \alpha_2 z_t + v_t$$

Como en el caso de modelos anidados, entre las opciones b) y c) será aceptable aquélla cuyos resultados se aproximen a los de a).

Los experimentos se han realizado para distinas combinaciones de la varianza de $x_t, v(x)$, de valores de β_1, del coeficiente de determinación del PGD (R^2) y para distintas correlaciones entre x_t y z_t (ρ). En la Tabla 3 aparecen algunos de los resultados obtenidos, donde las 6 últimas columnas se interpretan de la siguiente forma:

Ai: número de aceptaciones de la hipótesis de estabilidad predictiva en el modelo H_i para $i = 1,2$, utilizando la varianza real σ^2.

Aij: número de aceptaciones de la hipótesis de estabilidad predictiva en el modelo H_i utilizando la varianza estimada por el modelo H_j.

AiA: número de aceptaciones de la hipótesis de estabilidad predictiva utilizando, para estimar la varianza, el modelo amplio.

La conclusión que se deriva de esta tabla es clara: hay coincidencia de resultados utilizando la varianza real y estimando dicha varianza con el modelo amplio; también coinciden los datos si se estima la varianza con el modelo que es el PGD pero no, si se estima utilizando el modelo que no ha generado los datos.

Debido a que en la realidad no se conoce el PGD, sería aconsejable utilizar la varianza estimada del modelo que anide a los dos modelos no anidados, ya que utilizando para cada modelo la varianza estimada por el propio modelo se tiende a aceptar en un porcentaje alto de veces la hipótesis de estabilidad predictiva para el modelo H_2, cuando utilizando la varianza real no se aceptaría casi nunca.

Tabla 3

Modelos no anidados. PGD: $y_t = 1 + \beta_1 x_t + u_{1t}$

T	$v(x)$	ρ	R^2	β_1	A1	A2	A11	A22	A1A	A2A
50	0.01	0.4	0.90	1	478	0	474	500	475	0
50	0.01	0.4	0.99	1	471	0	471	500	472	0
50	0.01	0.9	0.90	1	476	0	480	500	482	0
50	0.01	0.9	0.99	1	468	0	481	500	481	0
50	1	0.4	0.90	1	476	0	473	500	475	0
50	1	0.4	0.99	1	476	0	474	500	475	0
50	1	0.9	0.90	1	461	229	477	481	477	325
50	1	0.9	0.99	1	472	0	471	500	473	0
100	0.01	0.4	0.90	1	479	0	471	458	471	0
100	0.01	0.4	0.99	1	481	0	479	500	480	0
100	0.01	0.9	0.90	1	477	0	476	491	477	0
100	0.01	0.9	0.99	1	478	0	476	500	477	0
100	1	0.4	0.90	1	470	0	468	500	469	0
100	1	0.4	0.99	1	477	0	467	500	467	0
100	1	0.9	0.90	1	478	174	484	487	483	237
100	1	0.9	0.99	1	475	0	475	500	473	0
500	0.01	0.4	0.90	1	471	0	472	429	472	0
500	0.01	0.4	0.99	1	470	0	468	494	468	0
500	0.01	0.9	0.90	1	471	0	474	490	474	0
500	0.01	0.9	0.99	1	476	0	475	500	474	0
500	1	0.4	0.90	1	467	0	473	499	473	0
500	1	0.4	0.99	1	479	0	481	500	482	0
500	1	0.9	0.90	1	472	99	476	462	476	96
500	1	0.9	0.99	1	469	0	472	495	472	0

6.3. Modelos VAR

Dentro de este tipo de modelos vamos a considerar dos casos atendiendo a los órdenes del PGD y del modelo de la hipótesis alternativa, aunque en este trabajo sólo presentamos los resultados que se refieren al caso en el que el PGD es de orden menor que el del modelo alternativo.

Comenzaremos considerando el caso en el que el PGD es un VAR(2) siendo el modelo alternativo, un VAR(1). En concreto, el PGD para dos variables, que adoptamos es el siguiente:

$$M_1: \quad (\text{PGD}) y_t = \begin{bmatrix} 0.02 \\ 0.03 \end{bmatrix} + \begin{bmatrix} 0.7 & 0 \\ 0.4 & 0.4 \end{bmatrix} y_{t-1} + \begin{bmatrix} a & 0 \\ c & b \end{bmatrix} y_{t-2} + u_t$$

donde:

$$y_t = \begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix}; \quad u_t = \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix}$$

con $u_t \sim \text{N}[0, \Sigma]$

y $$\Sigma = \begin{bmatrix} v(u_1) & 0 \\ 0 & v(u_2) \end{bmatrix}$$

Los modelos que se comparan son:

$$H_1: \quad y_t = \begin{bmatrix} \beta_{10} \\ \beta_{20} \end{bmatrix} + \begin{bmatrix} \beta_{111} & \beta_{121} \\ \beta_{211} & \beta_{221} \end{bmatrix} y_{t-1} + \begin{bmatrix} \beta_{112} & \beta_{122} \\ \beta_{212} & \beta_{222} \end{bmatrix} y_{t-2} + u_t$$

$$H_2: \quad y_t = \begin{bmatrix} \gamma_{10} \\ \gamma_{20} \end{bmatrix} + \begin{bmatrix} \gamma_{111} & \gamma_{121} \\ \gamma_{211} & \gamma_{221} \end{bmatrix} y_{t-1} + v_t$$

y los coeficientes a, b, c toman cada uno de ellos valores distintos dentro de los límites que hacen que el proceso sea estacionario.

$$\begin{aligned} a &= (-0.8, 0.2) \\ b &= (-0.8, 0.5) \\ c &= (0.5, 1) \end{aligned}$$

En los experimentos realizados también se han considerado varianzas de la perturbación del PGD, $v(u_{1t})$ y $v(u_{2t})$ distintas:

$$v(u_{it}) = (0.01, 5) \quad (i = 1, 2)$$

En la Tabla 4 se indica el número de veces que se acepta la hipótesis de estabilidad predictiva para cada modelo en las 500 réplicas realizadas por cada experimento. En

esta tabla, al igual que en las anteriores, mostramos algunos de los resultados que se han obtenido para todas las combinaciones de tamaño muestral, $T = (50, 100, 500)$, varianzas de las perturbaciones $v(u_{it})$ $i = 1, 2$ y los distintos valores de los coeficientes del PGD, a, b y c.

Los resultados para los distintos valores de c son similares, por lo que aquí presentamos únicamente los resultados obtenidos para $c = 1$. Como los resultados para $T = 50$ y $T = 100$ son similares, presentamos únicamente los de $T = 100$ y $T = 500$ ya que la presentación de todos los resultados aportaría poca información adicional.

La interpretación de las cinco últimas columnas es la siguiente:

Ai: número de veces que se acepta la hipótesis de estabilidad predictiva en el modelo H_i utilizando la matriz de varianzas y covarianzas conocida del PGD $(i = 1, 2)$.

Ai1: número de aceptaciones de la hipótesis de estabilidad predictiva en el modelo H_i $(i = 1, 2)$, cuando se estima la matriz de varianzas y covarianzas con el modelo VAR(2).

A22: número de veces que se acepta la hipótesis de estabilidad predictiva del modelo H_2 cuando se estima la matriz de varianzas y covarianzas con el modelo H_2, VAR(1).

Como puede verse en la Tabla 4, la conclusión a la que se llega es que utilizando la estimación de la matriz de varianzas del modelo amplio, VAR(2), se tiende a aceptar la hipótesis de estabilidad predictiva, para cada modelo, en un número de veces similar al que se aceptaría utilizando la varianza real. En cambio, utilizando la matriz de varianzas y covarianzas estimada por el propio modelo, para el modelo VAR(1) se tiende a aceptar la hipótesis de estabilidad predictiva en una proporción de veces mayor de la que se aceptaría utilizando la varianza real.

El segundo de los casos que hemos analizado es el de un VAR (2) frente a un VAR (4) y debido a que la conclusión es la misma que en el primer caso, no presentamos los resultados para no ampliar el trabajo.

CONCLUSIONES

En este trabajo se ha demostrado que una versión corregida del contraste de estabilidad predictiva puede ser un instrumento muy útil en un proceso para elegir entre dos modelos econométricos.

Tabla 4

Modelos VAR. PGD: $y_t = \begin{bmatrix} 0.02 \\ 0.03 \end{bmatrix} + \begin{bmatrix} 0.7 & 0 \\ 0.4 & 0.4 \end{bmatrix} y_{t-1} + \begin{bmatrix} a & 0 \\ c & b \end{bmatrix} y_{t-2} + u_t$

T	$v(u_{1t})$	$v(u_{2t})$	a	b	c	A1	A2	A11	A21	A22
100	0.01	0.01	−0.8	−0.8	1	467	0	468	0	116
100	0.01	0.01	−0.8	0.5	1	461	12	458	15	224
100	0.01	0.01	0.2	−0.8	1	469	6	463	12	283
100	0.01	0.01	0.2	0.5	1	450	58	446	69	211
100	0.01	5	−0.8	−0.8	1	469	3	468	4	219
100	0.01	5	−0.8	0.5	1	463	20	459	34	306
100	0.01	5	0.2	−0.8	1	455	41	452	59	311
100	0.01	5	0.2	0.5	1	442	251	451	264	352
100	5	0.01	−0.8	−0.8	1	457	0	461	0	89
100	5	0.01	−0.8	0.5	1	461	0	450	0	72
100	5	0.01	0.2	−0.8	1	459	0	456	0	258
100	5	0.01	0.2	0.5	1	444	0	450	0	71
500	5	5	−0.8	−0.8	1	466	0	465	0	99
500	5	5	−0.8	0.5	1	467	11	468	13	219
500	5	5	0.2	−0.8	1	473	7	470	11	272
500	5	5	0.2	0.5	1	459	52	462	64	233
500	0.01	0.01	−0.8	−0.8	1	480	0	482	0	101
500	0.01	0.01	−0.8	0.5	1	483	9	480	9	205
500	0.01	0.01	0.2	−0.8	1	471	13	473	13	280
500	0.01	0.01	0.2	0.5	1	477	55	478	52	203
500	0.01	5	−0.8	−0.8	1	477	1	481	1	205
500	0.01	5	−0.8	0.5	1	474	36	473	39	303
500	0.01	5	0.2	−0.8	1	465	45	468	48	258
500	0.01	5	0.2	0.5	1	472	273	471	273	381
500	5	0.01	−0.8	−0.8	1	473	0	477	0	102
500	5	0.01	−0.8	0.5	1	480	0	479	0	49
500	5	0.01	0.2	−0.8	1	478	0	477	0	245
500	5	0.01	0.2	0.5	1	464	0	460	0	71
500	5	5	−0.8	−0.8	1	472	0	477	0	90
500	5	5	−0.8	0.5	1	476	10	474	11	201
500	5	5	0.2	−0.8	1	477	14	473	15	273
500	5	5	0.2	0.5	1	473	50	476	55	195

La corrección hace referencia al estimador de la varianza de las perturbaciones que aparece en el denominador del estadístico que se usa para llevar a cabo el contraste de estabilidad predictiva.

La bondad de la corrección que se propone queda demostrada en el trabajo por la coincidencia que se produce entre los resultados obtenidos con el estadístico corregido, con los que se obtendrían conociendo la verdadera varianza de la perturbación del proceso. Esta coincidencia no tiene lugar si se utiliza el procedimiento habitual que aparece en los textos.

La demostración del resultado se hace mediante procedimientos analíticos y mediante un ejercicio de simulación para tres tipos de modelos diferentes: anidados, no anidados y VAR. En todos los casos la evidencia es clara a favor de la propuesta que se hace en el trabajo.

REFERENCIAS BIBLIOGRÁFICAS

[1] **Chow, G.C.** (1960). «Tests of equality between sets of coefficients in two linear regressions». *Econometrica*, **28**, 591–603.

[2] **Dhrymes, P.J., E.P. Howrey, S.H. Hymans, H.T. Shapiro** and *V. Zarnowitz* (1972). «Criteria for evaluation of econometric models». *Annals of Economic and social measurement*, **1**, 291–324.

[3] **Fisher, F.M.** (1970). «Tests of equality between sets of coefficients in two linear regressions: an expository note». *Econometrica*, **38**, 361–366.

[4] **Lütkepohl, H.** (1989). «Prediction tests for structural stability of multiple time series». *Journal of Business & Economic Statistics*, **7**, 129–135.

[5] **Lütkepohl, H.** (1991). *Introduction to multiple time series analysis*. Berlin. Springer-Verlag.

[6] **Pesaran, M.H., R.P. Smith** and **J.S. Teo** (1985). «Testing for structural stability and predictive failure: a review». *The Manchester School*, **53**, 280–295.

ENGLISH SUMMARY

THE PREDICTIVE STABILITY TEST TO CHOOSE BETWEEN ECONOMETRIC MODELS

ANTONIO AZNAR GRASA

Mª ISABEL AYUDA BOSQUE*

Departamento de Análisis Económico

In this paper we consider the predictive stability test as a specification test in a framework in which we try to choose between two models. We propose a modification in the habitual test that consists of replace the variance of the model, in the denominator, by the estimated variance of the less restricted model, in nested models or VAR models, or by the estimated variance of the one general model that nested the two models in the case of non-nested models.

Keywords: Predictive stability, Selection, Econometric models.

* Antonio Aznar Grasa y Mª Isabel Ayuda Bosque. Departamento de Análisis Económico. Facultad de CC. EE. y EE. Doctor Cerrada 1. 50005 ZARAGOZA. Fax: 761770.

- Received june 1995.

- Accepted november 1995.

Testing for structural stability is a standard exercise in econometric analysis and has become an established part of the econometrics tool box, see for example, Chow (1960), Fisher (1970), Dhrymes *et al.* (1972) and Pesaran *et al.* (1985).

Most of these tests can also be regarded as a measure of how a model estimated from data from period 1 adequately predicts the values of the dependent variable in period 2, conditional upon the observed values of the regressors. In this sense, they are often called tests of predictive failure.

These tests are based on the values taken by a quadratic form of the vector of errors of prediction, weighted by the elements of the inverse of the variance-covariance matrix of that vector. In order to have an operative version of this quadratic form, this matrix has to be estimated. The standard practise has been to estimate it by using the model that is tested.

In this paper and within a framework in which two models are compared, we propose to estimate that matrix by using a general model, in three different situations, depending on the models that we compare are nested, non-nested or VAR models. The proposed test is:

$$F = e_p'[I_{T_1} + X_p(X'X)^{-1}X_p']^{-1}e_p/T_1 \cdot \sigma^{+2}$$

where σ^{+2} is the estimated variance in a general model and T_1 is the number of extrasample observations.

After two sections where we introduce the paper and we present the standard models and the test, respectively, we derive the form of this test for nested models (Section 3) where σ^{+2} is the estimated variance of the less restricted model. In section 4 we derive the form of this test when we compare two non-nested models; in this case, σ^{+2} is the estimated variance derived from a general model which nests the two models. In section 5 we deal with VAR models. For each of the three type of models we show that this corrected stability test follows an F distribution. The evidence presented in the last section of the paper from the Monte-Carlo experiment confirms the analytical results derived in sections 3-5 and proves the usefulness of the new proposal.

In this experiment we calculate the number the times, in 500 replications, that each of three alternative predictive stability tests choose one of the two compared models. The three criteria differ on how the variance of the predictive stability test expresion is estimated, that is:

a) Substituting the variance by the known variance of the data generating process assumed in the Monte Carlo experiments.

b) Estimating the variance with each model.

c) Estimating the variance with the less restrictive model, in the case of nested models, or a general model that nests the two compared models, in the case of non-nested models.

All the comparisons show that, using the general model to estimate the variance, the results obtained do not differ from those one would obtain in the hypothetical case where the variance is known.

QÜESTIIÓ, Vol 20, 3 pp. 409-425, 1996

FACTOR ANALYSIS AND INFORMATION CRITERIA*

MICHELE COSTA•

University of Bologna

In this paper the research of the true number of latent factors in exploratory factor analysis model is studied through a comparison between the log likelihood ratio test statistics, the information criteria of Akaike, Schwarz and Hannan-Quinn and a procedure of cross-validation. In a simulation study the a priori knowledge of the exact factor structure is used to evaluate the goodness of the different methods.

Keywords: Factor analysis, number of factors, financial market, simulation.

* Support for this research was provided by MURST grant 40% «Scelte discrete, variabili latenti e modelli economici stocastici».

I'm grateful to Sandra Marciatori for computing assistance.

• Michele Costa. Department of Statistics. University of Bologna. Via Belle Arti, 41. 40126 Bologna.

- Article rebut el setembre de 1995.

- Acceptat l'abril de 1996.

1. INTRODUCTION

Factor analysis is closely related to unobservability problems, and especially to the problem of variables that «do not correspond directly to anything that is likely to be measured» (Griliches, 1977). Indeed the factor analysis model specifies a set of linear relations in which p observable variables are determined by k unobservable factors and p error terms.

The determination of the «true» number of factors is the first problem to be solved in the selection of the «true» factor model

$$X = f\,\Lambda + U$$

where $X_{N\times p}$ is the matrix of the observable data, $U_{N\times p}$ is the matrix of the errors, $\Lambda_{k\times p}$ is the matrix of the factor loadings, $f_{N\times k}$ is the matrix (with $k < p$) of the factors and N is the number of observations of the series used.

The identification of a stable factor structure is traditionally done by means of the likelihood ratio test statistics and, more recently, through other methods, as information criteria and cross-validation.

The purpose of this paper is to study using Monte Carlo methods the contribution of all these methods to the determination of the «true» number of factors.

2. THE SELECTION OF THE MODEL

2.1. The likelihood ratio test statistics

The possibility to test the number of factors is one of the main reasons for the success of the procedure of maximum likelihood (Lawley and Maxwell, 1971; Kim and Mueller, 1983).

The usual statistics (Anderson, 1984; Kendall and Stuart, 1973, vol. II, chapter 24) used to test the number of factors is the log likelihood ratio test statistics

$$LR = N^*(\log|\Lambda'\Lambda + \Psi| - \log|\Sigma|)$$

where Σ is the sample covariance matrix, Ψ is the diagonal covariance matrix of the error U and the number of observations is corrected by Bartlett's formula $N^* = N - (2p + 4k + 11)/6$.

Conway and Reinganum (1988) indicate cross-validation as an alternative solution for the determination of the number of factors. Cross-validation can be considered

as a two stage procedure. In the first stage maximum likelihood estimates of the parameters are calculated in a sample of p variables X. In the second stage the estimates obtained are not compared with the respective sample variance matrix Σ, but with a Σ^* of another sample of the same variables X, in order to isolate the stable factor structure from the random components.

The log likelihood ratio test statistics

$$LR = N^*(\log|\Lambda'\Lambda + \Psi| - \log|\Sigma|)$$

is therefore (Conway and Reinganum, 1988) modified into

$$CV = N^*(\log|\Lambda'\Lambda + \Psi| - \log|\Sigma^*|) + N^*(tr((\Lambda'\Lambda + \Psi)^{-1}\Sigma^*) - p)$$

2.2. Information criteria

Akaike's information criterion (Akaike 1979 and 1987) is probably the most relevant and famous as for the comparison and selection between different models and is constructed on log likelihood

$$AIC = -2\ \log\max L + 2h$$

where L denotes the likelihood function of the factor model and h is the number of the model's free parameters.

The first term can be interpreted as a goodness-of-fit measure, while the second gives a growing penalty to increasing numbers of parameters, according to the parsimony principle.

In the choice of the model a minimisation rule is used to select the model with the minimum Akaike information criterion value.

Following the modification of FPE (Final Prediction Error) proposed by Bhansali and Downham (1977), Smith and Spiegelhalter (1980) suggested to modify the AIC by transforming the second term into a generic αh:

$$AIC_\alpha = -2\ \log\max L + \alpha h$$

Still in the context of likelihood based procedures, Schwarz (1978) proposed the alternative information criterion

$$SCH = -\log\max L + \frac{1}{2}h\ \log N$$

that, unlike AIC, considers the number N of the observations and is therefore less favourable to factors inclusion.

Hannan and Quinn (1979) suggested another information criterion, based, as usual, on the minimisation of $-\log\max L + hC$

$$HQ = -2\log\max L + 2\,h\,c\,\log\log N \qquad c > 1$$

3. A SIMULATION

The purpose of this paper is to illustrate some results obtained on simulated data, for which the factor structure is perfectly known. The different methods, illustrated in the previous paragraphs, are applied to the simulated data and the indications of the number of factors are compared with the true value k, which is a priori known.

The following model is used to obtain the simulated matrices X^*

$$X^* = f^*\Lambda^* + U^*$$

where

- f^* is the $N \times k$ matrix of the factors, obtained by random extractions from a multivariate normal distribution with covariance matrix the identity matrix
- U^* is the $N \times p$ matrix of error terms, randomly extracted from a multivariate normal distribution with covariance matrix the identity matrix too
- Λ^* is the $k \times p$ matrix of factor loadings, obtained from a factor analysis of a sample of p assets returns randomly extracted from a set of 100 assets returns daily quoted at Milan stock exchange from 1986 to 1989.

The various methods illustrated above are thus applied to samples of simulated matrices X^* (with dimension $p = 20$, $p = 30$ and $p = 40$) to analyze the influence of variations in the number of the original variables.

For each value of p different numbers N of the observations analyzed were considered, in order to study how variations of N can influence the number of factors detected. Specifically the cases $N = 100$, $N = 200$, $N = 1000$, $N = 5000$ were considered.

Finally, in the simulations three different factor structures were analyzed in order to evaluate the chosen criteria for different values of k, specifically the cases $k = 1$, $k = 5$, $k = 10$.

In order to generate the k independent factors f^* a matrix of dimension 5000×10, corresponding to the maximum value of N and k, was randomly extracted from a multivariate normal distribution with covariance matrix the identity matrix. For other values of N and k appropriate submatrices were extracted from this matrix: for example, for the case of $k = 1$ and $N = 200$ the relative submatrix $f^*_{200\times 1}$ contains the first 200 rows of the first column of the $f^*_{5000\times 10}$.

To obtain the factor loadings Λ^* three samples of 20 assets returns, three samples of 30 and three of 40 were randomly and independently extracted from a set of 100. Then a factor analysis was performed with $k = 1$ to obtain $\Lambda^*_{20\times 1}$, $\Lambda^*_{30\times 1}$, $\Lambda^*_{40\times 1}$, with $k = 5$ to obtain $\Lambda^*_{20\times 5}$, $\Lambda^*_{30\times 5}$, $\Lambda^*_{40\times 5}$, with $k = 10$ to obtain $\Lambda^*_{20\times 10}$, $\Lambda^*_{30\times 10}$ and $\Lambda^*_{40\times 10}$.

The factor loadings Λ^* and the factors f^* are assumed as fixed. Having thus obtained the term $f^* \Lambda^*$, the simulated matrices X^* are obtained by p random extractions of the error terms vector U^*.

The factor structure is so a priori known as k are the columns of Λ^*, and the variability of the X^* is entirely attributable to the different determinations of the vector U^*: it's also possible to compare the indications given by the different criteria with the true and known k.

Summarizing, for $k = 1$ three matrix Λ^* were randomly and independently calculated, one of dimension 1×20 from a sample of 20 assets for the case $p = 20$, one of dimension 1×30 from a sample of 30 assets for the case $p = 30$ and the last of dimension 1×40 from a sample of 40 assets for the case $p = 40$.

The ensuing three models are the following:

$$
\begin{aligned}
p = 20 \qquad & X^*_{N\times 20} = f^*_{N\times 1}\, \Lambda^*_{1\times 20} + U^*_{N\times 20} \\
p = 30 \qquad & X^*_{N\times 30} = f^*_{N\times 1}\, \Lambda^*_{1\times 30} + U^*_{N\times 30} \\
p = 40 \qquad & X^*_{N\times 40} = f^*_{N\times 1}\, \Lambda^*_{1\times 40} + U^*_{N\times 40}
\end{aligned}
$$

For each model 100 extractions of U^* are considered, thus obtaining by 100 replications as many simulated matrices X^* for each value of N.

Therefore for $k = 1$, $p = 20$ and $N = 200$, 100 samples of 20 variables X^* are considered and so for each combinations of k, p and N. The same as for $k = 1$ is repeated for $k = 5$ and for $k = 10$.

To summarize the results and to compare the different methods two quantities were calculated: the root mean square error and the bias.

The root mean square error (RMSE) is

$$S = \left(\frac{1}{100} \sum_{i=1}^{100} (k_i^* - k)^2 \right)^{\frac{1}{2}}$$

k^* is the number of factors indicated by the generic method, k is the true number of factors underlying the simulated matrices X^* and i indicates the generic $i-$th sample.

Obviously S is calculated for each method and in general method A is better than method B if $S_A < S_B$, as S measures the distance between the true k and the empirical k^* and so the smaller S the better approximation of k one obtains through k^*.

The bias

$$D = \frac{\sum_{i=1}^{100} k_i^*}{100} - k$$

indicates the direction of the RMSE and is negative when the method underestimates the true number of factors and positive when k is overestimated.

The bias is calculated in order to complete the informations about the distribution of the k^* around k, indeed the RMSE indicates only the distance between k^* and k; information on the sign of this deviations is shown in the bias.

In what follows the results related to the simulation are illustrated. In order to make the exposition easier the different methods are assembled by «family»: first the Akaike's information criterion and his variants, second the information criterion of Hannan and Quinn in four different forms. Finally a conclusive table contains the best of *AIC*'s, the best of HQ's and the other methods.

Starting with the formulations of Smith and Spiegelhalter, the results obtained by transforming the *AIC* in:

$$AIC3 = -2 \log \max L + 3h$$

and

$$AIC4 = -2 \log \max L + 4h$$

are not particularly good, because $AIC3$ and $AIC4$ generally converge to the true value k more slowly than *AIC*.

The following tables show S_{AIC}, S_{AIC3}, S_{AIC4}, D_{AIC}, D_{AIC3} and D_{AIC4} for the different cases considered.

In this and the next tables the values below 0,05 are set to 0.

When $k = 1$, $AIC3$ and $AIC4$ are slightly better than AIC, even if AIC doesn't strongly depart from the true k. Beside, the dimension p of the simulated matrices doesn't seem to influence the results.

Table 1

Values of S (RMSE) and D (Bias) for AIC, AIC3 and AIC4 when k=1

$k=1$		AIC			$AIC3$			$AIC4$		
N		$p=20$	$p=30$	$p=40$	$p=20$	$p=30$	$p=40$	$p=20$	$p=30$	$p=40$
100	S	0,3	0,3	0,3	0	0	0	0	0	0
	D	0,1	0,1	0,1	0	0	0	0	0	0
200	S	0,4	0,3	0,2	0	0	0	0	0	0
	D	0,1	0,1	0	0	0	0	0	0	0
1000	S	0,5	0,5	0,5	0	0	0	0	0	0
	D	0,2	0,2	0,2	0	0	0	0	0	0
5000	S	0,4	0,5	0,5	0	0	0	0	0	0
	D	0,1	0,2	0,2	0	0	0	0	0	0

Table 2

Values of S (RMSE) and D (Bias) for AIC, AIC3 and AIC4 when k=5

$k=1$		AIC			$AIC3$			$AIC4$		
N		$p=20$	$p=30$	$p=40$	$p=20$	$p=30$	$p=40$	$p=20$	$p=30$	$p=40$
100	S	1,0	0,4	1,0	2,6	1,6	2,9	3,3	2,8	4,0
	D	−0,7	0	−0,5	−2,5	−1,3	−2,8	−3,3	−2,8	−4,0
200	S	0,5	0,6	0,4	1,1	0,1	0,2	1,9	0,8	0,5
	D	−0,1	0,2	0,1	−0,9	0	0	−1,8	−0,5	−0,3
1000	S	0,4	0,3	0,3	0	0	0	0	0	0
	D	0,2	0,1	0,1	0	0	0	0	0	0
5000	S	0,6	0,4	0,4	0	0	0	0	0	0
	D	0,2	0,1	0,1	0	0	0	0	0	0

When $k = 5$, $AIC3$ and $AIC4$ are initially much worse than AIC but, by increasing N,AIC shows a tendency to overestimate the number of factors and, on the contrary, $AIC3$ and $AIC4$ converge to the true value $k = 5$. Furthermore, getting from 20 to 40 variables, the true value of k is more easily detected.

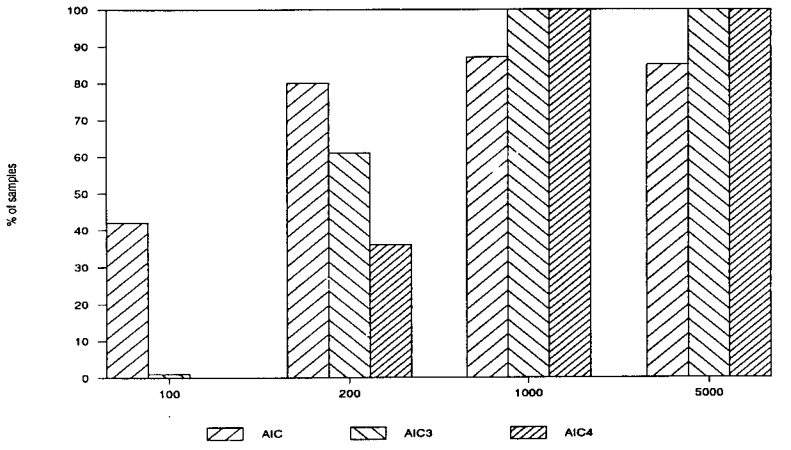


Figure 1. ***Number of samples (p = 30) in which AIC, AIC3 and AIC4 indicate k* = 5 when k=5.***

Table 3

Values of S (RMSE) and D (Bias) for AIC, AIC3 and AIC4 when k=10

$k = 10$		AIC			$AIC3$			$AIC4$		
N		$p = 20$	$p = 30$	$p = 40$	$p = 20$	$p = 30$	$p = 40$	$p = 20$	$p = 30$	$p = 40$
100	S	2,8	2,6	1,8	5,8	6,3	5,5	7,8	8,1	7,2
	D	−2,5	−2,3	−1,4	−5,7	−6,2	−5,4	−7,8	−8,1	−7,2
200	S	1,1	0,7	0,4	2,3	2,1	1,7	3,9	4,0	3,9
	D	−0.8	−0,2	0	−2,2	−1,9	−1,5	−3,7	−3,9	−3,7
1000	S	0,3	0,3	0,4	0,1	0	0	0,1	0	0
	D	0,1	0,1	0,2	0	0	0	0	0	0
5000	S	0,3	0,5	0,4	0	0,2	0	0	0	0
	D	0,1	0,2	0,2	0	0	0	0	0	0

When $k = 10$ the situation of $k = 5$ is repeated and AIC seems to be generally better than $AIC3$ and $AIC4$ which strongly underestimate the number of factors.

As for variations of α in AIC, variations of c in Hannan-Quinn criterion bring to different methods

$$\begin{aligned} HQ1 &= -2\log\max L + 2h\log\log N \\ HQ2 &= -2\log\max L + 2h\,2\log\log N \\ HQ3 &= -2\log\max L + 2h\,3\log\log N \\ HQ4 &= -2\log\max L + 2h\,4\log\log N \end{aligned}$$

and the relative results are illustrated in the following tables.

Table 4

Values of S (RMSE) and D (Bias) for HQ1, HQ2, HQ3 and HQ4 when k=1

$k=1$		$HQ1$			$HQ2$			$HQ3$			$HQ4$		
N		$p=20$	$p=30$	$p=40$	$p=20$	$p=30$	$p=40$	$p=20$	$p=30$	$p=40$	$p=20$	$p=30$	$p=40$
100	S	1,8	2,2	2,7	0	0	0	0	0	0	0	0	0
	D	1,3	1,6	2,1	0	0	0	0	0	0	0	0	0
200	S	1,2	0,9	1,1	0	0	0	0	0	0	0	0	0
	D	0,8	0,6	0,7	0	0	0	0	0	0	0	0	0
1000	S	0,5	0,6	0,6	0	0	0	0	0	0	0	0	0
	D	0,2	0,3	0,3	0	0	0	0	0	0	0	0	0
5000	S	0,3	0,3	0,2	0	0	0	0	0	0	0	0	0
	D	0,1	0,1	0,1	0	0	0	0	0	0	0	0	0

When $k=1$ only $HQ1$ shows difficulties in detecting the only factor and $HQ2$, $HQ3$ and $HQ4$, even with only 100 observations, perform adequately.

Table 5

Values of S (RMSE) and D (Bias) for HQ1, HQ2, HQ3 and HQ4 when k=5

$k=5$		$HQ1$			$HQ2$			$HQ3$			$HQ4$		
N		$p=20$	$p=30$	$p=40$	$p=20$	$p=30$	$p=40$	$p=20$	$p=30$	$p=40$	$p=20$	$p=30$	$p=40$
100	S	1,3	2,8	5,7	2,6	1,7	2,3	3,6	3,0	4,0	3,9	3,3	4,0
	D	0,7	1,6	5,7	−2,5	−1,4	−2,2	−3,6	−3,0	−4,0	−3,9	−3,3	−4,0
200	S	0,7	1,4	1,3	1,3	0,3	0,3	2,6	2,3	1,1	3,3	3,0	1,9
	D	0,3	0,9	1,0	−1,2	−0,1	−0,1	−2,5	−2,1	−1,0	−3,3	−3,0	−1,9
1000	S	0,5	0,4	0,5	0	0	0	0	0	0	0,2	0	0
	D	0,2	0,2	0,2	0	0	0	0	0	0	0	0	0
5000	S	0,4	0,2	0,2	0	0	0	0	0	0	0	0	0
	D	0,1	0,1	0	0	0	0	0	0	0	0	0	0

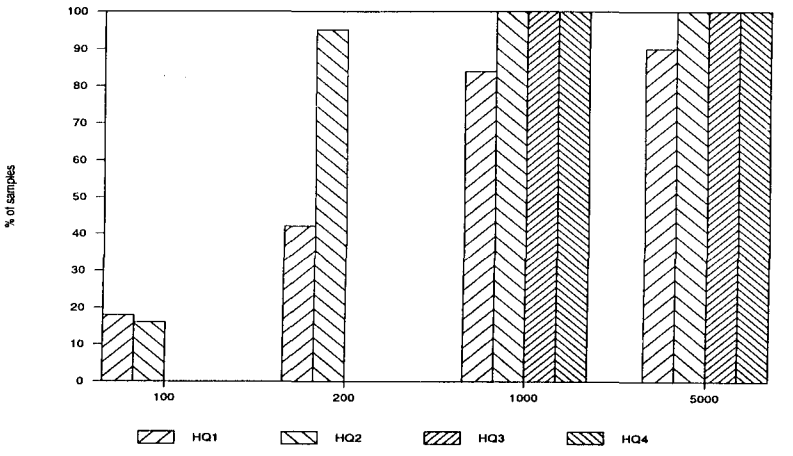


Figure 2. ***Number of samples (p = 30) in which HQ1, HQ2, HQ3 and HQ4 indicate $k^* = 5$ when k=5.***

When $k = 5$ the indications of $HQ2, HQ3$ and $HQ4$ are more differentiate and $HQ2$ seems to converge to the true value k more quickly than the other types of Hannan-Quinn criterion. When a larger number of factor is present in the model, $HQ1$'s goodness improves sensibly.

Table 6

Values of S (RMSE) and D (Bias) for HQ1, HQ2, HQ3 and HQ4 when k=10

$k = 10$		$HQ1$			$HQ2$			$HQ3$			$HQ4$		
N		$p = 20$	$p = 30$	$p = 40$	$p = 20$	$p = 30$	$p = 40$	$p = 20$	$p = 30$	$p = 40$	$p = 20$	$p = 30$	$p = 40$
100	S	1,6	0,9	0,9	6,1	6,4	5,6	8,4	8,7	7,9	9,0	9,0	8,8
	D	−1,1	0,2	0,7	−5,9	−6,4	−5,6	−8,4	−8,7	−7,9	−9,0	−9,0	−8,7
200	S	0,9	0,7	0,7	2,7	2,5	2,4	6,1	5,9	5,7	7,9	8,6	7,3
	D	−0,4	0,4	0,5	−2,6	−2,4	−2,3	−6,0	−5,9	−5,7	−7,9	−8,5	−7,2
1000	S	0,3	0,4	0,4	0	0	0	0,3	0	0	0.9	0,3	0
	D	0,1	0,1	0,2	0	0	0	−0,1	0	0	−0,7	−0,1	0
5000	S	0,2	0,4	0,2	0	0	0	0	0	0	0	0	0
	D	0,1	0,2	0,1	0	0	0	0	0	0	0	0	0

When $k = 10$ the situation of $k = 5$ is confirmed for $HQ2, HQ3$ and $HQ4$; yet $HQ1$ seems to be better than $HQ2$.

The results related to Akaike's, Hannan-Quinn's ($c = 2$), Schwarz's information criteria, cross-validation and log likelihood ratio test are reported in the following tables.

Table 7

Values of S (RMSE) and D (Bias) for AIC, HQ2, SCH, CROSS and LR when k=1

$k=1$		*AIC*			*HQ2*			*SCH*			*CROSS*			*LR*		
N		$p=20$	$p=30$	$p=40$	$p=20$	$p=30$	$p=40$	$p=20$	$p=30$	$p=40$	$p=20$	$p=30$	$p=40$	$p=20$	$p=30$	$p=40$
100	S	0,3	0,3	0,3	0	0	0	0	0	0	0	0	0	0,4	0,3	0,8
	D	0,1	0,1	0,1	0	0	0	0	0	0	0	0	0	0,1	0,1	0,3
200	S	0,4	0,3	0,2	0	0	0	0	0	0	0,1	0	0	0,4	0,3	0,4
	D	0,1	0,1	0	0	0	0	0	0	0	0	0	0	0,1	0,1	0,1
1000	S	0,5	0,5	0,5	0	0	0	0	0	0	0	0	0	0,4	0,3	0,3
	D	0,2	0,2	0,2	0	0	0	0	0	0	0	0	0	0,1	0,1	0,1
5000	S	0,4	0,5	0,5	0	0	0	0	0	0	0	0	0	0,3	0,2	0,1
	D	0,1	0,2	0,2	0	0	0	0	0	0	0	0	0	0,1	0	0

It's interesting to note how with $k = 1$ Schwarz's information criterion, cross-validation and log likelihood ratio test statistics, as *AIC* and *HQ2*, can detect the presence of the only factor with a satisfactory performance.

Table 8

Values of S (RMSE) and D (Bias) for AIC, HQ2, SCH, CROSS and LR when k=5

$k=5$		*AIC*			*HQ2*			*SCH*			*CROSS*			*LR*		
N		$p=20$	$p=30$	$p=40$	$p=20$	$p=30$	$p=40$	$p=20$	$p=30$	$p=40$	$p=20$	$p=30$	$p=40$	$p=20$	$p=30$	$p=40$
100	S	1,0	0,4	1,0	2,6	1,7	2,3	3,6	3,0	4,0	1,8	0,6	1,2	1,6	1,1	2,3
	D	−0,7	0	−0,5	−2,5	−1,4	−2,2	−3,6	−3,0	−4,0	−1,4	−0,2	−0,8	−1,4	−0,8	0,7
200	S	0,5	0,6	0,4	1,3	0,3	0,3	2,8	2,6	1,3	0,7	0	0,1	1,0	0,7	0,5
	D	−0,1	0,2	0,1	−1,2	−0,1	−0,1	−2,8	−2,6	−1,2	−0,4	0	0	−0,6	0	0
1000	S	0,4	0,3	0,3	0	0	0	0,1	0	0	0,1	0	0	0,4	0,3	0,3
	D	0,2	0,1	0,1	0	0	0	0	0	0	0	0	0	0,1	0,1	0,1
5000	S	0,6	0,4	0,4	0	0	0	0	0,1	0	0	0,1	0	1,2	0,2	0,3
	D	0,2	0,1	0,1	0	0	0	0	0	0	0	0	0	0,4	0	0,1

When $k = 5$, *AIC* and cross-validation seem to be the best methods and they converge to the true value more quickly than the other ones.

Schwarz's criterion shows an evident tendency to underestimate the true number of factors.

Table 9

Values of S (RMSE) and D (Bias) for AIC, HQ2, SCH, CROSS and LR when k=10

$k=10$		*AIC*			*HQ2*			*SCH*			*CROSS*			*LR*		
N		$p=20$	$p=30$	$p=40$	$p=20$	$p=30$	$p=40$	$p=20$	$p=30$	$p=40$	$p=20$	$p=30$	$p=40$	$p=20$	$p=30$	$p=40$
100	*S*	2,8	2,6	1,8	6,1	6,4	5,6	8,5	8,7	7,9	3,6	4,1	3,5	3,2	3,1	2,6
	D	−2,5	−2,3	−1,4	−5,9	−6,4	−5,6	−8,4	−8,7	−7,9	−3,0	−3,7	−3,1	−3,0	−2,9	−2,4
200	S	1,1	0,7	0,4	2,7	2,5	2,4	6,7	6,5	5,9	1,1	1,1	0,6	1,4	1,2	1,1
	D	−0,8	−0,2	0	−2,6	−2,4	−2,3	−6,6	−6,4	−5,9	−0,7	−0,6	−0,2	−1,1	−0,9	−0,9
1000	S	0,3	0,3	0,4	0	0	0	0,7	0	0	0,3	0,1	0	0,5	0,2	0,2
	D	0,1	0,1	0,2	0	0	0	−0,4	0	0	0,1	0	0	0,2	0	0,1
5000	S	0,3	0,5	0,4	0	0	0	0	0	0	0,2	0,4	0	1,0	0,2	0,2
	D	0,1	0,2	0,2	0	0	0	0	0	0	0,1	0,1	0	1,0	0	0

When $k = 10$ the situation for $k = 5$ is repeated again and *AIC* and cross−validation are the best methods.

4. THE NUMBER OF FACTORS IN THE FINANCIAL MARKET

The choice of the number of factors represents a crucial point in the theory of financial markets and especially in two of the most important assets returns models.

On one side the Capital Asset Pricing Model (CAPM) of Sharpe (1964) and Lintner (1965) assumes that only one factor can explain the assets returns; on the other the Arbitrage Pricing Theory (APT) of Ross (1976) states that k factors underlie the market.

Following the CAPM the return of the i−th asset is characterized by

$$E(r_i) = r_0 + (E(r_m) - r_0)\,\beta_i$$

where $\begin{cases} r_0 \text{ is the risk free rate;} \\ r_m \text{ is the return of the market portfolio;} \\ \beta_i = \operatorname{cov}(r_i, r_m)/\operatorname{var}(r_i). \end{cases}$

The resulting market model is

$$r_{it} = \alpha_i + \beta_i\, r_{mt} + \varepsilon_{it}$$

where $\begin{cases} \alpha_i = (1 - \beta_i) r_0; \\ \varepsilon_{it} \text{ is an error term.} \end{cases}$

The APT assumes that the generating model of the $i-$th asset is

$$E(r_i) = r_0 + \sum_{j=1}^{k} \Lambda_{ij}\, y_j$$

where y_j is the premium for risk associated with the factor j and the coefficients Λ_{ij} are estimated from the model

$$r_{it} = E(r_i) + \sum_{j=1}^{k} \Lambda_{ij}\, f_{jt} + u_{it}$$

where $\begin{cases} f_{jt} \text{ is the value at time } t \text{ of the latent factor } j; \\ u_{it} \text{ is an error term.} \end{cases}$

In order to discriminate between CAPM and APT it is necessary to determinate the number of factors; and this is the aim of this paper.

5. CONCLUSIONS

In this paper a simulation study is performed to compare different methods for choosing the number k of factors in a factor model. The definitions of considered methods are given in the next table 10, in which the last column contains the average percentage of successful indications, while in the second the average RMSE is reported

$$\overline{S} = \frac{1}{36} \sum_{k,N,p}^{36} \left(\frac{1}{100} \sum_{i=1}^{100} (k_i^* - k)^2 \right)^{\frac{1}{2}}$$

with $k = 1,5,10$; $N = 100,200,1000,5000$; $p = 20,30,40$.

Cross-validation indicates the true value in 76,9% of cases and, with the AIC (70,5%), it seems to be the most accurate method. Cross-validation and AIC have also the minimum $\overline{S}$ value. On the contrary, modifications of AIC don't improve the results ($\overline{S}_{AIC} < \overline{S}_{AIC3}$, ($\overline{S}_{AIC} < \overline{S}_{AIC4}$) and the percentage of success gets from 70,5 of AIC to 69,3 of AIC3 and to 66,3 of AIC4. In a similar way, modifications of HQ don't seem to produce better indications ($\overline{S}_{HQ2} < \overline{S}_{HQ1}$, $\overline{S}_{HQ2} < \overline{S}_{HQ3}$, $\overline{S}_{HQ2} < \overline{S}_{HQ4}$) and the percentage of success gets from 67,3 of HQ2 to 63,9 of HQ3 and to 61,5 of HQ1 and HQ4. Values of 3 or 4 for α in AIC and for c in HQ bring to a strong underestimate of the true value of k. Schwarz's information criterion, with a 62,7% of successful indications, also underestimates sensibly the number of factors, particularly when the number N of the observations is very large. When N increases, Schwarz's criterion do not overestimate the number of factors, as other criteria do. The usual test for the number of factors, the log likelihood ratio test gives 66,8% correct results.

Table 10

Methods for the determination of k and values of the medium RMSE

Method	$\overline{S}$	*% of success*
$CV = N^*(\log\|\Lambda'\Lambda+\Psi\| - \log\|\Sigma^*\|) + N^*(tr((\Lambda'\Lambda+\Psi)^{-1}\Sigma^*) - p)$	0,55	76, 9
$AIC = -2\log\max L + 2h$	0,63	70,5
$AIC3 = -2\log\max L + 3h$	0,90	69,3
$HQ2 = -2\log\max L + 2h2\log\log N$	0,95	67,3
$LR = N^*(\log\|\Lambda'\Lambda+\Psi\| - \log\|\Sigma\|)$	0,81	66,8
$AIC4 = -2\log\max L + 4h$	1,34	66,3
$HQ3 = -2\log\max L + 2h3\log\log N$	1,64	63,9
$SCH = -\log\max L + 0,5h\log N$	1,73	62,7
$HQ1 = -2\log\max L + 2h\log\log N$	0,98	61,5
$HQ4 = -2\log\max L + 2h4\log\log N$	1,98	61,5

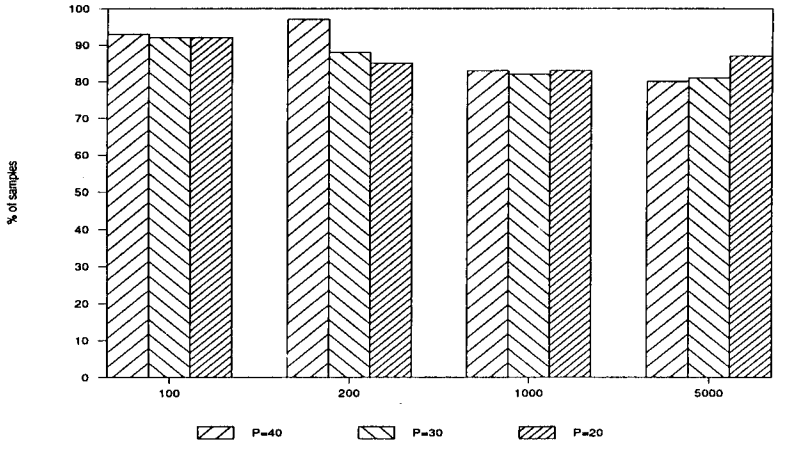


Figure 3. ***Number of samples in which AIC indicates k* = 1 when k=1.***

A further consideration is that the goodness of the different methods is a function of the number k of «true» factors underlying the simulated matrices.

Indeed in the case of $k = 1$ all methods analyzed indicate the right value $k = 1$ with the exceptions of AIC, HQ1 and LR. However for AIC and LR the distances from the exact value are quite small. In the previous picture the results related to AIC are illustrated: as N increases AIC gets a little worse.

With $k = 1$, even though the number of observations is only 100, there are already generally good indications and the dimension p seems to be not particularly relevant. This result shows how, when the true model contains only one factor, information criteria and cross validation can detect it with a good precision.

In the case $k = 5$ the situation is more complex and the number N of the observations is particularly relevant: asymptotically, indeed, all methods converge to the true value $k = 5$. However, it is important to emphasize that AIC and cross-validation converge more quickly.

The situation for $k = 10$ is similar to the one for $k = 5$: AIC and cross-validation show the best performance.

From the sign of the bias, reported in the next table 11, one can observe how the minus prevails thus meaning a stronger tendency to underestimate rather than to overestimate the true number of factors.

Table 11

Percentage of cases in which bias is negative, null or positive

	–	0	+
AIC	22	8	70
*AIC*3	28	72	–
*AIC*4	33	67	–
*HQ*1	6	–	94
*HQ*2	33	67	–
*HQ*3	36	64	–
*HQ*4	39	61	–
SCH	36	64	–
CV	28	64	8
LR	25	22	53

Concluding one can affirm that when only one factor constitutes the factor model a small number of observations is sufficient to detect it. When, on the contrary, more factors underlie the observed variables, cross-validation and AIC seem to be the more appropriate indicators.

6. REFERENCES

[1] **H. Akaike** (1979). «A bayesian analysis of the minimum AIC procedure». *Annals of the Institute of Statistical Mathematics A*, **30**, 9–14.

[2] **H. Akaike** (1987). «Factor analysis and AIC». *Psychometrika*, **52, 3**, 317–332.

[3] **T.W. Anderson** (1984). *An Introduction to Multivariate Statistical Analysis.* New York, Wiley.

[4] **M.S. Bartlett** (1950). «Tests of Significance in Factor Analysis». *British Journal of Mathematical and Statistical Psychology*, **3**, 77–85.

[5] **P. Bekker** (1989). «Identification in restricted factor models and the evaluation of rank conditions». *Journal of Econometrics*, **41**, 5–16.

[6] **R.J. Bhansali, D.Y. Downham** (1977). «Some properties of the order of an autoregressive model selected by a generalization of Akaike's EPF criterion». *Biometrika*, **64, 3**, 547–551.

[7] **H. Bozdogan, D.E. Ramirez** (1987). *An expert model selection approach to determine the best pattern structure in factor analysis models, in Multivariate statistical modeling and data analysis.* D. Reidel Publishing Company, 35–60.

[8] **D.E. Conway, M.R. Reinganum** (1988). «Stable Factors in Security Returns: Identification Using Cross-Validation». *Journal of Business & Economic Statistics*, **6**, 1–15.

[9] **Z. Griliches** (1977). «Errors in variables and other unobservables». In Aigner D., Goldberger A. (1977), *Latent variables in socio-economic models.* North-Holland.

[10] **E.J. Hannan, B.G. Quinn** (1979). «The determination of the order of an autoregression». *Journal of the Royal Statistical Society B*, **41, 2**, 190–195.

[11] **M.G. Kendall, A. Stuart** (1973). *The Advanced Theory of Statistics.* Griffin, London.

[12] **J. Kim, C.W. Mueller** (1983). *Factor Analysis.* London, Sage.

[13] **D.N. Lawley, A.E. Maxwell** (1971). *Factor Analysis as a Statistical Method.* London, Butterworths.

[14] **J. Lintner** (1965). «The Valuation of Risky Assets and the Selection of Risk Investments in Stock Portfolios and Capital Budgets». *Review of Economics and Statistics*, **47**, 13–37.

[15] **S. Ross** (1976). «The Arbitrage Theory of Capital Asset Pricing». *Journal of Economic Theory*, **13**, 341–360.

[16] **Y. Sakamoto, M. Ishiguro, I. Kitagawa** (1986). *Akaike information criterion statistics*. D. Reidel Publishing Company.

[17] **G. Schwarz** (1978). «Estimating the Dimension of a Model». *The Annals of Statistics*, **6**, 461–464.

[18] **W.F. Sharpe** (1964). «Capital Asset Prices: a Theory of Market Equilibrium under Conditions of Risk». *Journal of Finance*, **19**, 425–442.

[19] **A.F.M. Smith, D.J. Spiegelhalter** (1980). «Bayes factors and choice criteria for linear models». *Journal of Royal Statistical Society B*, **42, 2**, 213–220.

[20] **M. Stone** (1977). «An Asymptotic Equivalence of Choice of Model by Cross-validation and Akaike's Criterion». *Journal of Royal Statistical Society B*, **39, 1**, 44–47.

[21] **M. Stone** (1979). «Comments on Model Selection Criteria of Akaike and Schwarz». *Journal of Royal Statistical Society B*, **41, 2**, 276–278.

QÜESTIIÓ, Vol 20, 3 pp. 427-447, 1996

ESTIMACIÓN DEL PARÁMETRO *b* DE RICHTER A PARTIR DE LAS MEDIDAS IMPRECISAS DE INTENSIDAD EPICENTRAL

S. FORCADA*

J.J. EGOZCUE*

Universitat Politècnica de Catalunya

El objetivo de este trabajo es la estimación del parámetro exponencial de Richter que determina la distribución de probabilidad de la variable intensidad de movimientos sísmicos que se producen en una determinada zona. Las medidas de intensidad sísmica son, por propia naturaleza, imprecisas y por ello se propone estimar el parámetro mediante un procedimiento que incorpore la imprecisión de los datos. Se cuantifica la imprecisión asignando a cada medida de intensidad una distribución de probabilidad sobre el rango de intensidades. Cada medida imprecisa de intensidad se puede entonces interpretar como una realización de una variable aleatoria valorada en el espacio de Hilbert $l^2(R)$ cuya esperanza se identifica con la distribución de la variable intensidad. La estimación del parámetro se obtiene entonces a través de la minimización de la distancia cuadrática en $l^2(R)$ entre la distribución teórica de la variable intensidad y la media muestral de las medidas imprecisas de intensidad. Se puede demostrar que el estimador así obtenido proporciona bajo ciertas hipotesis, razonables en la práctica, un estimador del parámetro asintóticamente centrado y normal. Se aplica el método a la obtención de los estimadores del parámetro de Richter para dos catálogos sísmicos concretos, uno real correspondiente a una zona del Pirineo Oriental y otro simulado. Los resultados se comparan con las estimaciones de máxima verosimilitud y Bayes obtenidas al considerar los datos como precisos.

Estimation of Richter *b* parameter based on imprecise measures of epicentral intensity

Keywords: Point estimation, maximum likelihood, imprecise data, random elements, Hilbert spaces, epicentral intensity.

* S. Forcada y J.J. Egozcue. Departament de Matemàtica Aplicada III. Universitat Politècnica de Catalunya. TERRASSA.

- Article rebut el febrer de 1995.

- Acceptat el març de 1996.

1. INTRODUCCIÓN

En la práctica de la ingeniería sísmica es preciso determinar la posibilidad de que una cierta obra civil se vea afectada por una acción sísmica durante determinado período de tiempo. Este tipo de cálculos suele recibir el nombre genérico de «evaluación de la peligrosidad sísmica». Entre los parámetros que determinan la peligrosidad se encuentran los períodos de retorno de distintos tipos de terremotos y la proporción entre distintos tipos de acción.

En las regiones de sismicidad moderada surge una problemática específica relacionada con la calidad de los datos disponibles. En dichas regiones, los períodos de retorno a estimar pueden ser muy largos, frecuentemente por encima de los cien años, y , en consecuencia, para proceder a su estimación es necesario disponer de los datos recogidos a lo largo de varios siglos. Obviamente estos datos no son homogéneos en su calidad y, además, pueden ser incompletos.

Por ejemplo, en el Pirineo Oriental, no se dispone de datos instrumentales hasta 1914 y estos datos no son completos hasta los años sesenta o más tarde. Se hace, por tanto, necesario utilizar la intensidad sísmica como parámetro para caracterizar el «tamañano» de un terremoto. Esta opción permite la extensión de los catálogos sísmicos más allá del período de tiempo en que se dispone de datos instrumentales.

La intensidad sísmica se define a partir de los efectos que un terremoto produce en un lugar. Se trata de los efectos sobre edificios, ruido percibido, vibraciones de ciertos objetos, alteraciones de topografía, e incluso reacciones de la población frente al terremoto. En base a esta información se construye de forma cualitativa una escala discreta de intensidades que va de uno a doce. Nos referiremos aquí a la escala MSK de uso en Europa (ver [29]).

La definición de los distintos grados indica claramente que la intensidad que se asigna a un terremoto depende de muchos factores que varían con los tiempos históricos. Por ejemplo, cultura y dispersión de la población, calidad de los edificios, posición del epicentro respecto de los núcleos de población, etc.

En la actualidad la intensidad se evalúa mediante encuestas sobre la población afectada. El número de personas encuestadas es, aún hoy en día, muy variable (normalmente la extensión de la muestra es muy baja) y la muestra no se puede considerar aleatoria.

Además, el procesado de las encuestas para deducir la intensidad en el epicentro adolece de un enorme grado de subjetividad. La situación es más precaria para datos correspondientes a siglos anteriores al XX. Para esos terremotos la intensidad epicentral debe deducirse de informaciones documentales casi siempre parciales y que no tienen por qué adecuarse a la definición de la escala de intensidades.

En consecuencia, la intensidad sísmica, incluso en tiempos recientes es un dato impreciso. Un terremoto consignado, en un catálogo sísmico, como de intensidad epicentral siete en el siglo XVIII podría haber sido de intensidad ocho si el epicentro no coincidió con un núcleo de población, pero también podría corresponder a intensidad seis si los informes consultados tendieron a exagerar los daños o los efectos sobre la población. El margen de error puede incluso extenderse a otros grados de intensidad. En este planteamiento subyace la idea de que cada terremoto posee una intensidad epicentral «verdadera» pero desconocida.

Otro factor que interviene críticamente en la utilización de catálogos históricos de terremotos es la completitud de los datos. Ciertamente algunos sucesos sísmicos pueden quedar ignorados en los catálogos, ya sea porque no afectaron a poblaciones, porque no se consignaron los efectos o porque los informes documentales se perdieron. Esto involucra especialmente a las intensidades más bajas. En nuestro enfoque del problema no contemplamos la problemática derivada de la completitud de los datos, limitándonos a considerar catálogos supuestamente completos.

El problema de la evaluación de la peligrosidad sísmica teniendo en cuenta la imprecisión de los datos sólo ha sido abordado tímidamente. Ello es debido, en parte, a que los desarrollos más importantes se han realizado en zonas de alto riesgo en las que se dispone de muchos datos instrumentales. Así, los métodos tradicionales (ver [7], [18], [19]) sólo tienen en cuenta incertidumbres en la atenuación sísmica para estudiar la peligrosidad local, y de forma un tanto vaga la imprecisión de la localización del epicentro.

Posteriormente se ha comprendido que la mera estima puntual de probabilidades de ocurrencia no informa suficientemente al usuario de los resultados. Por ello se han introducido modificaciones en los métodos clásicos. Especial mención merecen los métodos bayesianos que permiten obtener fácilmente estimas por intervalo de probabilidad (ver [5], [9], [11], [22]). Pero la introducción de los errores de la intensidad en la estimación de los parámetros de la peligrosidad, desde distintas perspectivas, es más reciente (ver [6], [10], [12], [14]).

La escasez de técnicas estadísticas para analizar datos imprecisos de intensidad es una de las razones por las que las intensidades se anotan, en la mayor parte de catálogos, como datos precisos. A lo sumo, en algunos catálogos se consignan códigos de fiabilidad de las intensidades. Resulta, por tanto, de interés ensayar diferentes metodologías de estimación con datos imprecisos así como la codificación de los mismos.

Existen diversos intentos de utilizar datos intervalares (ver [1], [20]) y parámetros borrosos (ver [15], [26], [31]) en el problema de la peligrosidad sísmica (ver [14]). No obstante, el enfoque intervalar implica la suposición de un comportamiento pro-

babilísticamente uniforme en el intervalo de error, lo que no refleja adecuadamente muchas situaciones reales.

El planteamiento con datos y parámetros borrosos completa, en cierto sentido, las técnicas intervalares. Sin embargo conlleva la interpretación y manejo del concepto de parámetro borroso, que no es práctica usual en la ingeniería sísmica ni en la evaluación de la peligrosidad sísmica.

Las técnicas de estimación bayesiana (ver [3], [25]) han sido utilizadas ampliamente en el análisis de la peligrosidad sísmica (ver [7], [11]). Sin duda, estos métodos bayesianos son una potente herramienta para resolver problemas asociados a la evaluación de la peligrosidad sísmica, con la ventaja de una interpretación clara desde un punto de vista probabilístico. No obstante su utilización, incluso en casos muy sencillos, puede, como veremos, presentar graves dificultades.

En este trabajo abordamos el problema de la estimación del parámetro de Richter cuando la imprecisión en los datos de intensidad se cuantifica mediante distribuciones de probabilidad sobre el rango de intensidades. A partir de aquí la muestra de intensidades imprecisas se entiende como una muestra de una variable aleatoria a valores en un espacio de Hilbert. Ello nos permite la utilización de ciertos argumentos de carácter geométrico que combinados con los de tipo probabilístico, nos llevan a la obtención de un estimador asintóticamente centrado y normal. Los resultados teóricos se aplican a la obtención de los estimadores en dos casos prácticos concretos. Uno corresponde a un catálogo sísmico real y el otro a un catálogo simulado.

2. EL MODELO DE INTENSIDAD

Nos planteamos como objetivo estimar el parámetro exponencial (parámetro de Richter) que determina la frecuencia relativa de las diferentes intensidades, en una determinada región sísmica, a partir del catálogo correspondiente de intensidades.

Consideraremos las intensidades sísmicas $i = i_0, \ldots, i_s$, donde i_0 e i_s son respectivamente las intensidades menor y mayor de todas las consideradas. El valor de i_0 se fija normalmente en función del interés en ingeniería sísmica y frecuentemente es $i_0 = 5$. La intensidad superior i_s suele determinarse a partir de las intensidades máximas registradas en la región y frecuentemente $i_s = 10$, 11, ó 12.

En general se procede de la manera siguiente: dado un terremoto, se le asigna una intensidad epicentral, I, entre los valores i_0 e i_s. I es un experimento multinomial y en consecuencia la distribución de la variable aleatoria $I : \Omega \longrightarrow \mathbb{R}$, queda especificada

por las probabilidades

$$p_i = P(I = i), \qquad i = i_0, \ldots, i_s \qquad \text{con} \qquad \sum_{i=i_0}^{i_s} p_i = 1.$$

Sin embargo, tradicionalmente se suele considerar que estos parámetros están ligados por alguna ley exponencial. Esto proviene del hecho experimental (ver [27]) de que las magnitudes sísmicas a gran escala obedecen a una distribución exponencial, conocida con el nombre de ley de Richter.

Por analogía, se suele considerar que las intensidades epicentrales en una zona también siguen este tipo de distribución en su versión discreta. De las diferentes versiones existentes hemos adoptado aquí el siguiente modelo acumulado.

$$P(I \geq i) = e^{-b(i-i_0)} \qquad i = i_0, \ldots, i_s$$

$$P(I \geq i_0) = 1, \qquad P(I > i_s) = 0$$

que da lugar a la función de probabilidad

$$p_i = P(I = i) = e^{-b(i-i_0)}(1 - e^{-b}), \qquad i = i_0, \ldots, i_{s-1}$$

$$p_s = P(I = i_s) = e^{-b(i_s-i_0)}.$$

El parámetro b determina ahora la distribución de I y lo denominaremos parámetro de Richter.

El modelo de comportamiento probabilístico más utilizado para describir las ocurrencias de terremotos en un período de tiempo es el de Poisson. Se han propuesto diversos modelos alternativos (Ver [24], [28], [30]) que se caracterizan por ser no estacionarios y depender de múltiples parámetros, lo cual los hace poco utilizables en zonas de baja sismicidad donde escasean los datos.

3. ESTIMACIÓN DEL PARÁMETRO *b* CUANDO SE CONSIDERA LOS DATOS PRECISOS

Supongamos que disponemos de un catálogo de datos sísmicos completo a lo largo de un tiempo T en el que se han registrado m sucesos cuyas intensidades, supuestas precisas, son $y_1, \ldots, y_m$. Suponemos además que el comportamiento probabilístico de esta muestra corresponde a la función de probabilidad dada por la Ley de Richter.

Si m_1 es el número de veces que los valores $i < i_s$ aparecen en la muestra, entonces la función de verosimilitud es

$$L(y_1, \ldots, y_m; b) = e^{-b(\sum_{j=1}^{m}(y_j - i_0))}(1 - e^{-b})^{m_1}$$

y consecuentemente el estimador máximo verosímil es

$$\hat{b} = log\left(\frac{m_1}{\sum_{j=1}^{m}(y_j - i_0)} + 1\right)$$

cuya varianza asintótica viene dada por:

$$\hat{V}(b) = \frac{1}{nE_b\left[\left(\frac{\partial}{\partial b} log P_I(b)\right)^2\right]} =$$

$$= \frac{\frac{1}{n}}{\sum_{i=i_0}^{i_s-1}\left(\frac{e^{-b}}{1-e^{-b}} - (i - i_0)\right)^2 (1-e^{-b})e^{-b(i-i_0)} + (i_s - i_0)^2 e^{-b(i_s - i_0)}}$$

El intervalo de confianza asintótico se obtendrá sustituyendo b en esta expresión por el valor estimado $\hat{b}$ y tomando el valor resultante como la varianza del estimador, que a su vez se supone normal. Si escribimos $\hat{V}(\hat{b})$ para indicar la varianza asintótica así obtenida, el intervalo de confianza $1 - \alpha$ será

$$(\hat{b} - z\sqrt{\hat{V}(\hat{b})}, \hat{b} + z\sqrt{\hat{V}(\hat{b})})$$

donde z representa el punto para el que la función de distribución correspondiente a una distribución normal estandarizada toma valor igual a $1 - \frac{\alpha}{2}$. Si adoptamos el enfoque bayesiano para estimar b, nos encontramos con que no se dispone de funciones de densidad a priori conjugadas de fácil manejo.

La distribución gama de parámetros β, q parece adecuada para modelar la distribución a priori de b ya que dicho parámetro b es positivo, y a priori debería suponerse unimodal.

$$f_b'(b) = \frac{\beta^q}{\Gamma(q)} b^{q-1} e^{-\beta b}$$

A partir de la función de verosimilitud $L(y_1, \ldots, y_m; b) = (1 - e^{-b})^{m_1} e^{-b\left(\sum_{j=1}^{m}(y_j - i_0)\right)}$ calculamos la función de densidad a posteriori:

$$f_b''(b) = \frac{1}{C} b^{q-1}(1 - e^{-b})^{m_1} e^{-b\left(\beta + \left(\sum_{j=1}^{m}(y_j - i_0)\right)\right)} \qquad 0 < b$$

donde la constante C queda determinada (ver [13]) por:

$$C^{-1} = \int_0^{+\infty} b^{q-1}(1-e^{-b})^{m_1} e^{-b\left(\beta+\left(\sum_{j=1}^{m}(y_j-i_0)\right)\right)} db =$$

$$= \Gamma(q)(-1)^{m_1} \sum_{j=1}^{m}(-1)^k \binom{m_1}{k} \frac{1}{\left(m_1+\beta+\left(\sum_{j=1}^{m}(y_j-i_0)\right)-k\right)^q}$$

Si se toma la esperanza a posteriori de b como estimador bayesiano puntual se tiene, resolviendo integrales similares a la anterior:

$$E''(b) = q\frac{\sum_{k=1}^{m}(-1)^k \binom{m_1}{k} \left(m_1+\beta+\left(\sum_{j=1}^{m}(y_j-i_0)\right)-k\right)^{-(q+1)}}{\sum_{k=1}^{m}(-1)^k \binom{m_1}{k} \left(m_1+\beta+\left(\sum_{j=1}^{m}(y_j-i_0)\right)-k\right)^{-q}}$$

El intervalo de probabilidad corresponderá a los percentiles $\frac{\alpha}{2}$, $1-\frac{\alpha}{2}$ de la distribución a posteriori.

Puede utilizarse una función de densidad a priori impropia no informativa (ver [3], [25]) poniendo en la distribución a posteriori $q=1$, $\beta=0$. Con ello, la moda de la densidad a posteriori coincide con el estimador máximo verosímil.

Observemos que en este caso la densidad a priori no es la conjugada de la distribución de la variable I. La densidad a posteriori no corresponde a una densidad usual y, por tanto, el cálculo de los percentiles para el intervalo de probabilidad es algo más costoso que cuando no es así.

4. DATOS IMPRECISOS DE INTENSIDAD

En los catálogos sísmicos, normalmente, se especifica el momento en que ocurrió el suceso, su posición epicentral, y una estima de su intensidad epicentral. De este primer catálogo se extraen los sucesos que corresponden a la región estudiada, al rango de intensidades de interés y que además, están dentro del intervalo de tiempo en el que se supone que el catálogo está completo para cada intensidad. Después de esta primera depuración de los datos procederemos a la codificación de su imprecisión.

Asignaremos una distribución de probabilidad a cada suceso, de manera que dicha distribución constituirá el dato impreciso de intensidad epicentral. Así, por ejemplo, supongamos que en el catálogo consta que el suceso ω ha sido de intensidad k. Entonces se asocia a ω el dato impreciso $\chi(\omega)$ consistente en una distribución de probabilidad sobre $i_0,\ldots,i_s$, que para cada intensidad $i=i_0,\ldots,i_s$ toma el valor $\chi(\omega)(i)$. Este valor, $\chi(\omega)(i)$, se interpretará como la probabilidad de que el suceso ω, consignado como de intensidad k en el catálogo, haya sido en realidad de intensidad epicentral

i. Si disponemos de m sucesos $\omega_1, \ldots, \omega_m$, con sus respectivas distribuciones de probabilidad $\chi(\omega_j)$ sobre los valores de intensidad, escribiremos, a fin de simplificar la notación, $\chi_1 = \chi(\omega_1), \ldots, \chi_m = \chi(\omega_m)$. Entonces la probabilidad $\chi(\omega_j)(i)$ de un valor i de intensidad según $\chi(\omega_j)$ se escribirá $\chi_j(i)$

De este modo, a cada suceso ω del espacio de probabilidad se asocian: a) una intensidad epicentral «verdadera» $I(\omega)$, que no es directamente observable y que por tanto se desconoce y b) una distribución de probabilidad representada por la función de probabilidad $\chi = \chi(\omega)$. En términos generales el problema que se nos plantea es el de estimar el parámetro asociado a la distribución de probabilidad de una variable I : $\Omega \longrightarrow \mathbb{R}$, cuando la información de que se dispone al realizarla es una distribución de probabilidad. Al realizar m repeticiones independientes de un mismo experimento, la muestra que se obtiene consiste en m distribuciones de probabilidad que generalmente no son iguales, y es a partir de estos datos que queremos estimar el parámetro de la distribución de X. Observemos que si $I(\omega_1) = I(\omega_2)$, ello no implica necesariamente que $\chi_1 = \chi(\omega_1)$ sea igual a $\chi_2 = \chi(\omega_2)$ ya que en general serán diferentes. También puede suceder, tal como lo estamos planteando, que $\chi(\omega_1) = \chi(\omega_2)$ y $I(\omega_1) \neq I(\omega_2)$.

Los catálogos sísmicos tradicionales sólo especifican la intensidad epicentral de cada suceso y, a lo sumo, una indicación más o menos vaga de su precisión. Por tanto, la obtención de un catálogo de datos imprecisos de intensidad comporta una tarea larga de revisión de datos y documentos históricos. A continuación presentamos una evaluación preliminar de los datos contenidos en el catálogo del Pirineo Oriental que más tarde utilizaremos para ilustrar las técnicas de estimación.

La zona del Pirineo Oriental considerada está definida por una poligonal cuyos vértices son los de la tabla 1 y que se señalan en el mapa adjunto de la zona (Fig. 1).

Esta zona ya fue utilizada en [9] en un estudio con datos supuestamente precisos. El catálogo, proviene de Mezcua y Martínez ([21]) y ha sido complementado con otras informaciones provenientes del Servei Geològic de la Generalitat de Catalunya (ver [4]).

Tabla 1

Pirineo Oriental

Long.	0.30	1.80	2.30	3.60	3.00	0.90
Lat.	42.60	42.00	42.10	42.20	43.20	43.00

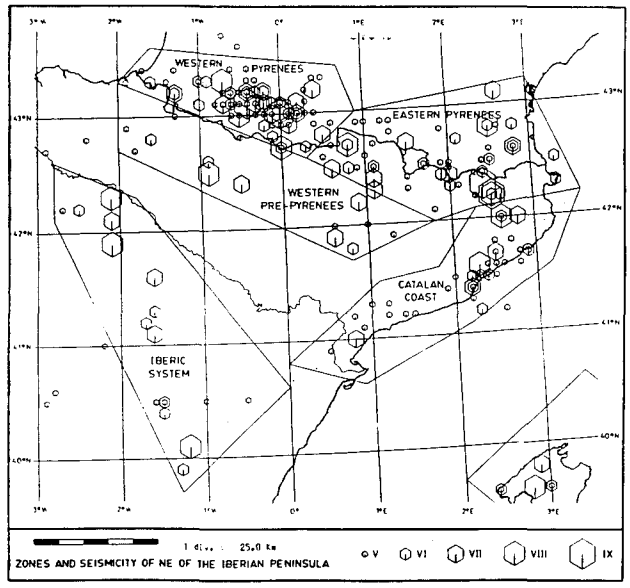


Figura 1

Dependiendo de la fecha y de la intensidad que consta en el catálogo se ha asignado una distribución de probabilidad al dato impreciso. La tabla 2 muestra los diferentes tipos de función de probabilidad (dato impreciso) que se suponen centradas en la intensidad que figura en el catálogo original.

Tabla 2

Período	Intensidad	Catálogo impreciso				
		$i-2$	$i-1$	i	$i+1$	$i+2$
1000-1800	$5 \leq i \leq 7$	0.05	0.10	0.70	0.10	0.05
1000-1800	$i \geq 8$	0.00	0.10	0.75	0.10	0.05
1801-1914	$5 \leq i \leq 6$	0.01	0.12	0.75	0.12	0.00
1801-1914	$i \geq 7$	0.00	0.12	0.80	0.08	0.00
1915-1950	i	0.00	0.10	0.80	0.10	0.00
1951-1980	i	0.00	0.07	0.90	0.03	0.00
1981-1990	i	0.00	0.04	0.95	0.01	0.00
1427-1428	$i = 10$	0.00	0.10	0.60	0.20	0.10

De esta forma disponemos de una muestra imprecisa de la variable intensidad I que notaremos:

$$\chi_1 = \Big(\chi_1(i_0),\dots,\chi_1(i_s)\Big), \chi_2 = \Big(\chi_2(i_0),\dots,\chi_2(i_s)\Big),\dots,\chi_m = \Big(\chi_m(i_0),\dots,\chi_m(i_s)\Big).$$

Sin duda, la asignación sistemática de estas funciones de probabilidad viene afectada por una componente fuertemente subjetiva. Su modificación y mejora corresponde a un largo y costoso estudio de los datos históricos. En este sentido, los estudios históricos iniciados bajo los auspicios del Servei Geològic de la Generalitat de Catalunya pueden aportar en un futuro próximo mejoras importantes. La figura 2 proporciona una representación gráfica del catálogo del Pirineo Oriental, con datos imprecisos. Cada línea vertical corresponde a la probabilidad de que el terremoto ocurrido en la fecha sobre la que se encuentra dicha línea, haya sido de la intensidad correspondiente. En esta representación gráfica, de cara a la información visual que proporciona, hay que tener en cuenta dos factores. En primer lugar, sucede a menudo que para un mismo año se superponen distintos terremotos en las diferentes intensidades. Ello significa que el conjunto de todas las probabilidades sobre un mismo año no constituyen una distribución de probabilidad. En segundo lugar la representación es general para todos los registros y no para cada período de completitud del catálogo.

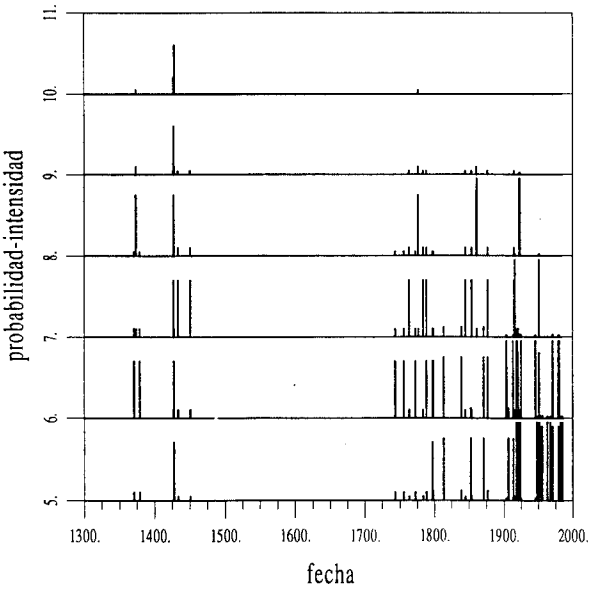


Figura 2

Pirineo Oriental

Desde un punto de vista práctico, una primera aproximación al problema de la estimación de b a partir de estos datos imprecisos, nos lo proporciona la posibilidad de obtener diversas muestras precisas de la variable I a partir de la realización de las distintas distribuciones χ_j de la muestra imprecisa. Cada vez que realizamos un elemento de la muestra imprecisa obtenemos una muestra «precisa» de la variable inicial I que nos da inmediatamente el estimador de máxima verosimilitud $\hat{b}$ del parámetro b. Si repetimos el proceso un determinado número de veces obtendremos un número de estimadores igual al número de repeticiones, y podremos tomar como estimador el promedio de todos ellos.

Desde la óptica bayesiana, podemos plantear a nivel teórico la estimación de b incluyendo la imprecisón de los datos en el cálculo de la densidad a posteriori realizado en el apartado **3**. Si consideramos la muestra de I imprecisa, disponemos de las distribuciones $\chi_1, \ldots, \chi_m$. Llamamos Y_j a la variable aleatoria, que proporciona el valor de la intensidad en la observación j-ésima y que por tanto su distribución de probabilidad es χ_j. Entonces el número de veces que observaremos sucesos para los que $i < i_s$ es una variable aleatoria que indicaremos por M_1. La densidad a posteriori es

$$f''_{b|M_1,Y_1,\ldots,Y_m}(b|m_1,y_1,\ldots,y_m) = \frac{1}{C} b^{q-1}(1-e^{-b})^{M_1} e^{-b\left(\beta+\left(\Sigma_{j=1}^{m}(Y_j-i_0)\right)\right)} \qquad 0 < b$$

y al ser $M_1, Y_1, \ldots, Y_m$ variables aleatorias, la distribución a posteriori de b es en realidad la distribución marginal de b. O sea:

$$\tilde{f}''_b(b) = \sum_{m_1,y_1,\ldots,y_m} f''_b(b) P(M_1 = m_1, Y_1 = y_1, \ldots, Y_m = y_m)$$

Al no poder calcular la distribución conjunta de $M_1, Y_1, \ldots, Y_m$ no se puede obtener $\tilde{f}''_b(b)$.

5. ESTIMACIÓN DEL PARÁMETRO *b* CUANDO SE CONSIDERA LOS DATOS IMPRECISOS

Tenemos la variable aleatoria intensidad $I : \Omega \longrightarrow \mathbb{R}$ cuya distribución de probabilidad corresponde a la ley de Richter del apartado **2** y que indicaremos por P_b. Queremos estimar b a partir de una muestra imprecisa, en el sentido del apartado **4**, de la variable I. Sea $\chi_1, \ldots, \chi_m$ dicha muestra.

Enfocaremos el problema de estimación de b en el contexto de elementos aleatorios en espacios de Hilbert. Para abordar el problema utilizaremos los resultados de [23] sobre elementos aleatorios en espacios métricos separables.

Representaremos cada distribución de probabilidad, $(p_{i_0},\dots,p_{i_s})$ definida sobre el rango de posibles intensidades, $i_0,\dots,i_s$, mediante el elemento $(0,\dots,0,p_{i_0},\dots,p_{i_s},0,\dots)$ del espacio vectorial l^1 de las sucesiones de números reales absolutamente sumables. En estos términos, la ley de Richter corresponde a

$$P_b=\left(0,\dots,0,1-e^{-b},\dots,e^{-kb}(1-e^{-b}),\dots,e^{-(s-1)b}(1-e^{-b}),e^{-sb},0,\dots\right)$$

Al ser l^1 subespacio del espacio de Hilbert l^2 de las sucesiones de números reales de cuadrado sumable, interpretaremos una muestra imprecisa

$$\chi_1=(\chi_1(i_0),\dots,\chi_1(i_s)),\dots,\chi_m=(\chi_m(i_0),\dots,\chi_m(i_s))$$

de la variable I, como una muestra de una variable aleatoria χ a valores en l^2. Esto es, una función medible

$$\chi:(\Omega,\mathcal{A},P)\longrightarrow(l^2,\mathcal{B}_2)$$

con $\mathcal{B}_2$ la σ-álgebra generada por los abiertos de l^2 según la norma definida por $\|\{a_k\}_{k\in N}\|=\left(\sum_{k=1}^{\infty}|a_k|^2\right)^{\frac{1}{2}}$

Consideremos el conjunto $M=\left\{(0,\dots,0,p_{i_0},\dots,p_{i_s},0,\dots)\in l^1 \mid p_{i_k}\geq 0,\ \sum_{k=0}^{s}p_{i_k}=1\right\}\subset l^1$ y observemos que $\chi(\Omega)\subset M$

Las realizaciones de χ constituyen la información que extraemos de las realizaciones de la intensidad I. Para que esta información pueda ser fiable deberemos suponer que las distribuciones de probabilidad que nos proporciona la variable χ no son exageradamente arbitrarias sino que son coherentes con los resultados de I. Es decir, los comportamientos probabilísticos de I y χ han de estar relacionados de manera razonable. Se supondrá, por hipótesis, que la esperanza $E(\chi)=\int_\Omega\chi(\omega)dP$ de la variable χ coincide con la distribución de probabilidad de la variable I. Esto es:

$$E(\chi)=P_b.$$

Esta hipótesis puede justificarse ya que $E(\chi)\in M$, y la probabilidad que asigna la distribución de probabilidad $E(\chi)$ a cada i_k, con $k=0,\dots,s$, es $E(\chi)(i_k)=\int_\Omega\chi(\omega)(i_k)dP$. Es decir, $E(\chi)$ asigna a cada posible valor de intensidad i_k la probabilidad esperada de que $I=i_k$. Parece adecuado entonces considerar que el comportamiento probabilístico de χ sea tal que dicha probabilidad esperada coincida con la probabilidad de i_k dada por P_b.

Por la ley de los grandes números, para variables aleatorias valoradas en espacios de Hilbert, la distribución promedio muestral $\bar{\chi}_m=\frac{1}{m}\sum_{i=1}^{m}\chi_i$ converge en probabilidad

a P_b, lo cual sugiere estimar b mediante el valor que minimiza, en l^2, la distancia cuadrática $||\bar{\chi}_m(\omega) - P_b||^2$ entre $\bar{\chi}_m = \frac{1}{m}\sum_{i=1}^{m}\chi_i$ y P_b.

Obsérvese que los valores que minimizan esta distancia cuadrática son los mismos que minimizan la función

$$\mathcal{L}(\chi_1,\ldots,\chi_m;b) = \sum_{i=1}^{m}||\chi_i - P_b||^2 = \sum_{i=1}^{m} < \chi_i - P_b, \chi_i - P_b > .$$

Llamaremos $\hat{b}_m$ al estimador de b tal que $min_{\{b\}}\mathcal{L}(\chi_1,\ldots,\chi_m;b) = \mathcal{L}(\chi_1,\ldots,\chi_m;\hat{b}_m)$ obtenido a través de la ecuación $\frac{\partial\mathcal{L}}{\partial b} = 0$.

Suponiendo que $E\left[\left(\frac{\partial}{\partial b}||\chi - P_b||^2\right)^2\right]$ es finita y teniendo en cuenta que la función $\mathcal{L}(\chi_1,\ldots,\chi_m;b)$ es tal que $\frac{1}{n}\frac{\partial^3}{\partial b^3}\mathcal{L}(b)$ está acotada en el entorno de todo punto $\bar{b}$, se comprueba, mediante argumentos paralelos a los de los teoremas 2.1, 2.2 y 2.3 de [17] cap. 6, que la ecuación $\frac{\partial}{\partial b}\mathcal{L}(\chi_1,\ldots,\chi_m;b) = 0$ tiene, con probabilidad tendiendo a uno cuando $m \to \infty$, una raíz $\hat{b}_m$ que corresponde a un mínimo de $\mathcal{L}(\chi_1,\ldots,\chi_m;b)$ cumpliéndose además que la sucesión $\sqrt{m}(\hat{b}_m(\chi_1,\ldots,\chi_m) - b)$ converge en ley a una variable normal de esperanza cero y varianza

$$\frac{E\left(< \frac{\partial P_b}{\partial b}, \chi - P_b >^2\right)}{||\frac{\partial P_b}{\partial b}||^4}$$

6. CASO PRÁCTICO DE ESTIMACIÓN DEL PARÁMETRO DE RICHTER

Hemos aplicado el desarrollo anterior a la estimación del parámetro de Richter para dos catálogos sísmicos. Uno de ellos es un catálogo obtenido por simulación y el otro es el catálogo presentado en el apartado **4** correspondiente a la zona del Pirineo Oriental.

El catálogo del Pirineo corresponde al período comprendido entre 1940 y 1987 en el que se registraron 20 terremotos de intensidades comprendidas entre 5 y 10. En el catálogo simulado, en un período de 750 años, se han registrado 362 terremotos de intensidades comprendidas también entre 5 y 10. En el caso del Pirineo Oriental se asocian los modelos de imprecisión de la tabla 2 tal como se explicó en el apartado **4** y en el caso del catálogo simulado la asociación se hace por sorteo, en proporciones preestablecidas y con $b = 1$. La figura 3 representa el catálogo simulado de manera análoga a como se representó el catálogo del Pirineo Oriental, en la figura 2.

En primer lugar se ha realizado la estimación mediante las técnicas estándar de máxima verosimilitud y Bayes considerando los datos como precisos.

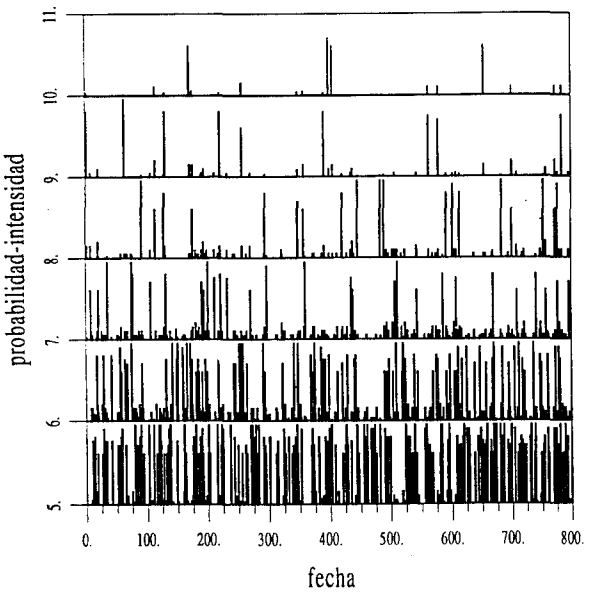


Figura 3

Catálogo Impreciso Simulado

En segundo lugar se ha procedido a la estimación de b después de haber modelado la imprecisión de los datos muestrales según las distribuciones de la tabla 2. Las estimas se han realizado por simulación y mediante $\mathcal{L}$. Tal como ya se apuntó anteriormente, la estima bayesiana con imprecisión no es en este caso viable.

Por último se ha utilizado la función $\mathcal{L}$ para estimar b cuando suponemos que la imprecisión se anula. Partíamos de unos registros puntuales iniciales a los que se había asignado un modelo de imprecisión. Si ahora sustituimos los modelos de imprecisión por distribuciones de probabilidad que asocian probabilidad uno al valor registrado y cero a los demás valores, estamos anulando la imprecisión introducida anteriormente, pero a pesar de ello podemos llevar a cabo la estimación de b con la misma técnica ideada para tratar el caso de datos imprecisos. Los resultados se presentan en la tabla 3.

Las filas ML(datos precisos) y Bayes(datos precisos) corresponden a las estimaciones de b cuando se supone que en los registros no ha habido imprecisión y se

acepta como buena la intensidad consignada en el catálogo. Se dan aquí (Columna Est. punt.) las estimas de máxima verosimilitud y bayesiana del parámetro. Se indica también (Columna Std. A) la desviación típica asintótica para el caso del estimador máximo verosímil y la desviación típica de la distribución a posteriori para el caso de la estima bayesiana.

En las dos filas siguientes se tienen los resultados de las diferentes estimaciones del parámetro cuando se supone que los registros de que se dispone son imprecisos y que además dicha imprecisión se cuantifica a través de los modelos de imprecisión de la tabla 2. En este caso la estimación de b se lleva a cabo por simulación, primero y luego utilizando $\mathcal{L}$. Como se indicó en **4**, la obtención de la estima simulada se realiza mediante el promedio de los distintos estimadores máximo verosímiles obtenidos para cada muestra simulada.

Tabla 3

	Est. punt.	Std. A	Std. BS	$d(l^2)$
ML(datos precisos)	0.8935	0.04882		
Bayes(datos precisos)	0.8963	0.04890		
Sim(datos imprecisos)	0.8158			0.02067
$\mathcal{L}$ (datos imprecisos)	0.7920	0.04564	0.04492	0.01679
$\mathcal{L}_\delta$ (imprec. nula)	0.8935	0.06318	0.06044	0.03104

(a) Estimación del parámetro b de Richter. Catalogo Simulado

	Est. punt.	Std. A	Std. BS	$d(l^2)$
ML(datos precisos)	1.253	0.2991		
Bayes(datos precisos)	1.334	0.3129		
Sim(datos imprecisos)	1.244			0.1098
$\mathcal{L}$ (datos imprecisos)	1.034	0.2951	0.3003	0.07458
$\mathcal{L}_\delta$ (imprec. nula)	1.040	0.3172	0.3191	0.08376

(b) Estimación del parámetro b de Richter. Pirineo Oriental

La segunda columna (Std. A) da la desviación típica asintótica estimada del estimador $\hat{b}$, obtenido con la función $\mathcal{L}$, según su expresión dada en el apartado **5**, en donde $E\left(< \frac{\partial P_b}{\partial b}, \chi - P_b >^2\right)$ se estima mediante $\frac{1}{n}\sum_{i=1}^{n}\left(< \frac{\partial P_b}{\partial b}_{b=\hat{b}}, \chi_i - P_{\hat{b}} >^2\right)$ y $||\frac{\partial P_b}{\partial b}||^4$ mediante $||\frac{\partial P_b}{\partial b}_{b=\hat{b}}||^4$.

En la tercera columna (Std. BS) se tiene la estimación de la desviación típica de $\hat{b}$ obtenida mediante bootstrap sobre la muestra imprecisa, en la que se ha muestreado mil veces. Observemos que estas dos estimaciones son prácticamente coincidentes.

En la última columna se aporta una medida del grado de cumplimiento de la hipótesis $E(\chi) = P_b$. Para verificar el cumplimiento de esta hipótesis es preciso conocer el verdadero valor del parámetro b y la esperanza de χ; ello evidentemente no es posible y, en principio, aceptaremos que en nuestro caso y para nuestros datos la hipótesis se cumple. A pesar de ello, al converger $\bar{\chi}_m$ a $E(\chi)$, la distancia en l^2 entre $\bar{\chi}_m$ y $P_{\hat{b}}$, con $\hat{b}$ el estimador correspondiente de b, nos proporciona una cierta medida del grado de cumplimiento de esta hipótesis. Indicamos esta distancia en las tablas mediante $d(l^2)$. Por tanto, cuanto menor sea el valor que aparece en la cuarta columna más podemos confiar en el cumplimiento de la hipótesis establecida. Los valores numéricos obtenidos indican que en el caso del catálogo simulado las distribuciones $P_{\hat{b}}$ son prácticamente iguales a la distribución $\bar{\chi}$, y en el caso del catálogo de Catalunya tampoco son muy diferentes. Podemos suponer cómodamente, entonces, que la hipótesis se cumple en ambos casos.

En la cuarta fila ($\mathcal{L}_\delta$) se presentan los resultados análogos a los de la tercera, pero en este caso, reduciendo la imprecisión a cero. Es decir, cuando suponemos que los datos son precisos y lo expresamos considerando que la muestra obtenida la constituyen distribuciones con probabilidad uno en los registros iniciales. Se recogen aquí los resultado de aplicar la función $\mathcal{L}$ anterior a este caso, siendo el significado de los distintos apartados análogo al de los correspondientes apartados de la tercera fila.

Observemos que tanto en el caso del catálogo simulado como en el real se obtiene un valor estimado de b menor cuando se considera los datos imprecisos que cuando se hace nula la imprecisión. Ello es coherente con el hecho de que b mide el decrecimiento de la frecuencia relativa con la intensidad ($logP(I \geq i) = -b(i - i_0)$). La proporción de terremotos es más grande para intensidades pequeñas y cuando se modela la imprecisión tal como hemos hecho aquí la cesión de probabilidades que se produce de unas intensidades a otras redunda en un aumento de las probabilidades en las intensidades superiores y, en consecuencia, en un decrecimiento de la probabilidad con la intensidad y en la obtención de una estima menor que la obtenida con datos precisos.

BIBLIOGRAFÍA

[1] **Arstein, Z.** (1974). «On the Calculus of Closed Set-Valued Functions». *Indiana Univ. Math. J.*, **24**, 433–441.

[2] **Ash, R.B.** (1972). *Real Analysis and Probability*. Academic Press.

[3] **Box, G.E.P** y **Tiao, G.C.** (1973). *Bayesian Inference in Statistical analysis*. Addison-Wesley.

[4] **Butlletí Sismológic 1985-1990.** Servei Geològic de Catalunya. Departament de Política Territorial i Obres Públiques. Generalitat de Catalunya.

[5] **Campbell, K.W.** (1982). «Bayesian Analysis of Extreme Earthquake Ocurrences. Part I. Probabilistic Hazard Model». *Bull. Seism. Soc. Am.*, **72**, 1689–1706.

[6] **Chen, D., Dong, W.** y **Shah, C.** (1988). «Earthquake Recurrence Relationships from Fuzzy Earthquake Magnitudes». *Soil Dynamics and Earthquake Engineering*, **7**, 136–142.

[7] **Cornell, C.A.** (1968). «Engineering Seismic Risk Analysis». *Bull. Seism. Soc. Am.*, **58**, 1583–1606.

[8] **Efron, B.** y **Tibshirani, R.J.** (1993). «An Introduction to the Bootstrap». *Monographs on Statistics and Applied Probability*. Chapman Hall.

[9] **Egozcue, J.J., Barbat, A., Canas, J.A., Miquel, J.** y **Banda, E.** (1991). «A Method to Estimate Intensity Ocurrence Probabilities in Low Seismic Activity Areas». *Earthquake Eng. Struc. Dyn.*, **20**.

[10] **Egozcue, J.J., Díez, P.** y **Muñoz, P.** (1991). «Bayesian Hazard Assessement Using Imprecise Data». *Proc. of CERRA-ICASP*, **6**, Mexico DF.

[11] **Esteva, L.** (1969). «Seismicity Prediction: A Bayesian Approach». *Proc. 4^{th} World Conf. Earthquake Eng.*, Santiago de Chile.

[12] **Feng, D., Gu, J., Lin, M., Xu, S.** y **Yu, X.** (1985). «Assessement of Earthquake Hazard by Simultaneous Use of the Statistical Method and the Method of Fuzzy Mathematics». *Pageoph.*, **122**, 982–997.

[13] **Gradshteyn, I.S.** y **Ryzhik, I.M.** (1980). *Tables of Integrals, Series, and Products*. Academic Press.

[14] **Kijko, A.** y **Sellevoll, M.A.** (1990). «Estimation of Earthquake Hazard Parameters for Incomplete and Uncertain Data Files». *Natural Hazards*, **3**, 1–13.

[15] **Kruse, R.** y **Meyer, K.D.** (1987). *Statistics with Vague Data*. D. Raidel Publishing Company.

[16] **Laha, R.G.** y **Rohatgi, V.K.** (1979). *Probability Theory*. Wiley.

[17] **Lehmann, E.L.** (1983). *Theory of Point Estimation*. Wiley.

[18] **McGuire, R.K.** (1977). «Effects of Uncertainty in Seismicity on Estimates of Seismic Hazard for the East Coast of United States». *Bull. Seism. Soc. Am.*, **67**, 827–848.

[19] **McGuire, R.K.** «FORTRAN Computer Program for Seismic Risk Analysis». *U.S.G.S. Open File Report*, 67–76.

[20] **Matheron, G.** (1975). *Random Sets and Integral Geometry*. Wiley.

[21] **Mezcua, J.** y **Martínez, J.M.** (1983). «Sismicidad del area Ibero-Mogrebí». *Inst. Geogr. Nacional*. Madrid.

[22] **Morgat, C.P.** y **Shah, H.C.** (1979). «A Bayesian Model for Seismic Hazard Mapping». *Bull. Seism. Soc. Am.*, **69**, 1237–1251.

[23] **Padgett, W.J.** y **Taylor, R.L.** (1973). «Laws of Large Numbers for Normed Linear Spaces and Certain Fréchet Spaces». *Lecture Notes 360*. Springer-Verlag.

[24] **Patwardhan, A.S., Kulkarni, R.B.** y **Tocher, D.** (1980). «A Semi-Markov Model for Characterizing Recurrence of Great Earthquakes». *Bull. Seism. Soc. Am.*, **70**, 323–347.

[25] **Press, S.J.** (1989). *Bayesian Statistics: Principles, Models, and Applications*. Wiley.

[26] **Puri, M.L.** y **Ralescu, D.A.** (1986). «Fuzzy Random Variables». *J. Math. Anal. Appl.*, **114**, 409–422.

[27] **Richter, C.F.** (1958). *Elementary Seismology*. Freeman.

[28] **Shlien, S.** y **Toksoz, N.** (1975). «A Branching Poisson-Markov Model of Earthquake Ocurrences». *Geophys. J. R. Soc.*, **42**, 49–59.

[29] **Udías, A.** y **Mezcua, J.** (1986). *Fundamentos de Geofísica*. Alhambra.

[30] **Vere-Jones, D.** (1975). «Stocasthic Models for Earthquakes Sequences». *Geophys. J. R. Soc.*, **42**, 811–826.

[31] **Zadeh, L.A.** (1968). «Probability Mesures of Fuzzy Events». *J. Math. Anal. Appl.*, **10**, 421–427.

ENGLISH SUMMARY

ESTIMATION OF RICHTER *b* PARAMETER BASED ON IMPRECISE MEASURES OF EPICENTRAL INTENSITY

S. FORCADA*

J.J. EGOZCUE*

Universitat Politècnica de Catalunya

In this paper the problem of seismic Richter b parameter estimation starting from imprecise intensity data when the imprecision is quantified by a distribution of probability is considered. An imprecise data value of intensity is perceived as a realization of a random probability measure which is represented on the Hilbert space $l^2(R)$. In distinction to Bayes approach geometrical and probabilistic arguments in the Hilbert space lead to the construction of an assymptotically normal and centered estimator of Richter b parameter which provide good results through easy calculations.

Keywords: Point estimation, maximum likelihood, imprecise data, random elements, Hilbert spaces, epicentral intensity.

* S. Forcada y J.J. Egozcue. Departament de Matemàtica Aplicada III. Universitat Politècnica de Catalunya. TERRASSA.

– Received february 1995.

– Accepted march 1996.

We aim to estimate Richter's exponential parameter, which caracterize the seismic activity in areas of low seismicity when working with imprecise data.

An epicentral intensity value $I \in \{i_0, \ldots, i_s\}$ is assigned to each given seismic movement. The probability distribution of the random variable $I : (\Omega, \mathcal{A}, P) \longrightarrow (R, \mathcal{B}_R)$, is determined by the probabilities

$$p_i = P(I = i), \qquad i = i_0, \ldots, i_s \qquad \text{with} \qquad \sum_{i=i_0}^{i_s} p_i = 1.$$

By hypothesis the exponential accumulated model derived from the Richter's law for seismic intensities is assumed to be:

$$P(I \geq i) = e^{-b(i-i_0)} \qquad i = i_0, \ldots, i_s$$

$$P(I \geq i_0) = 1, \qquad P(I > i_s) = 0$$

This leads to the probability function P_b defined by

$$p_i = P(I = i) = e^{-b(i-i_0)}(1 - e^{-b}), \qquad i = i_0, \ldots, i_{s-1}$$

$$p_s = P(I = i_s) = e^{-b(i_s - i_0)}.$$

and the Richter's parameter, b, characterizes the distribution of the random variable I.

The available data for estimating b are given by the seismic catalogue of the area in which the estimated epicentral intensities are specified. This measurements of seismic intensity are, by their very nature, inaccurate and weakely defined. The estimation of the parameter is carried out taking into account in the procedure the data uncertainties. The uncertainty of the data is quantified by attributing to each intensity measurement a probability distribution over the range of intensities. If the intensity stated in the seismic catalogue for the event ω is k, the uncertainty on the observed intensity value is modeled by assigning a distribution of probability $\chi(\omega)$ to ω. This $\chi(\omega)$ is a probability function defined on $i_0, \ldots, i_s$. For each $i = i_0, \ldots, i_s$ the value $\chi(\omega)(i)$ is the probability of i being the true intensity of the event ω which in the catalogue was indicated as $I = k$. For m events $\omega_1, \ldots, \omega_m$ we will obtain a sample $\chi_1 = \chi(\omega_1), \ldots, \chi_m = \chi(\omega_m)$ which will be called imprecise sample.

We represent each probability distribution, $(p_{i_0}, \ldots, p_{i_s})$, on the range of intensities, by the sequence $(0, \ldots, 0, p_{i_0}, \ldots, p_{i_s}, 0, \ldots)$ of the space $l^1(R) \subset l^2(R)$ Hence, the imprecise sample

$$\chi_1 = (\chi_1(i_0), \ldots, \chi_1(i_s)), \ldots, \chi_m = (\chi_m(i_0), \ldots, \chi_m(i_s))$$

of I corresponds to a sample of a random variable

$$\chi : (\Omega, \mathcal{A}, P) \longrightarrow (l^2(R), \mathcal{B}_2)$$

being $\mathcal{B}_2$ the σ-algebra generated by the open subsets of the Hilbert space $l^2(R)$

The probabilistic behaviour of random variables I and χ are related by the hypothesis

$$E(\chi) = P_b$$

where P_b denotes the probability measure associated to I.

The estimation, $\hat{b}_m$, of parameter b is obtained through the minimisation of the function

$$\mathcal{L}(\chi_1, \ldots, \chi_m; b) = \sum_{i=1}^{m} ||\chi_i - P_b||^2$$

which represents the sum of squared distances in $l^2(R)$ between the theoretical distribution of the intensity variable, I, and each element of the imprecise sample.The norm in $l^2(R)$ is $||\{a_k\}_{k\in N}|| = \left(\sum_{k=1}^{\infty} |a_k|^2\right)^{\frac{1}{2}}$

It can be shown that the estimation obtained in such a manner provides, under certain reasonable practicable hypotheses, an assimptotically centred and normal estimator.

The method is applied to the estimation of Richter parameter for two specific seismic catalogues, one real and corresponding to an area in the Eastern Pyrenees and a simulated one. We have given a model of uncertainty to each intensity value stated in the Seismic Catalogue following the information provided by the Servei Geologic de la Generalitat de Catalunya. We apply also the method to estimate b with the data when considered accurate. When imprecise data are replaced by the probability distribution over $\{i_0, \ldots, i_s\}$ which assigns probability one to the intensity in the seismic catalogue and zero to the other intensity values the method applies. Both cases (imprecise data and accurate data) are compared.

QÜESTIIÓ, Vol 20, 3 pp. 449-482, 1996

LA COMBINACIÓ DE TÈCNIQUES DE GEOMETRIA DIFERENCIAL AMB ANÀLISI MULTIVARIANT CLÀSSICA: UNA APLICACIÓ A LA CARACTERITZACIÓ DE LES COMARQUES CATALANES

SERGI VIVES*

ÀNGEL VILLARROYA*

Universitat de Barcelona

En aquest treball s'estudien les característiques i les relacions de les 41 comarques de Catalunya basant-se en diverses variables socio-econòmiques. La metodologia utilitzada combina tècniques clàssiques d'anàlisi de dades amb un nou mètode per a la representació simultània de poblacions i variables, basat en tècniques de geometria diferencial, anomenat IDA (Intrinsic Data Analysis). Els resultats obtinguts han permès classificar les comarques de Catalunya en 9 grups que es poden reagrupar en 4 grans blocs (agrícola, turístic, administratiu i industrial), que presenten una clara gradació geogràfica.

Combining geometric-differential techniques with classic multivariate analysis: an application to characterize catalan regions

Keywords: Intrinsic Data Analysis; distància de Rao; Anàlisi de Correspondències; Anàlisi de conglomerats.

Classificació AMS: 62H25, 62H30.

*S. Vives i À. Villarroya. Dept. d'Estadística. Universitat de Barcelona. Av. Diagonal, 645. 08028 Barcelona. Espanya.

- Article rebut el gener de 1995.
- Acceptat el setembre de 1996.

1. INTRODUCCIÓ

Des del treball pioner de Rao (1945), molts autors han desenvolupat una branca de l'estadística basada en aplicacions de la geometria diferencial (Efron, 1975; Atkinson i Mitchell, 1981; Burbea i Rao, 1982, 1984). Malauradament, molts d'aquests treballs són molt teòrics i s'aprecia una mancança d'aplicacions pràctiques. En aquest treball es mostra un exemple de com algunes d'aquestes tècniques, per si mateixes o combinades amb altres de *clàssiques*, constitueixen unes eines molt potents de fàcil aplicació i interpretació.

Els objectes estadístics d'aquest treball els formen les *comarques* de Catalunya. La divisió territorial de Catalunya en comarques va ser proposada per primera vegada el 1936 per la Generalitat de Catalunya, encara que aquesta divisió no va gaudir de poders administratius fins als anys 80. Els criteris originals en què es va basar aquesta divisió van ser fonamentalment comercials (proximitat al mercat més proper), sense deixar de banda aspectes geogràfics i històrics, importants, però de segon ordre. No cal dir que l'origen d'aquestes comarques va comportar problemes, ja que algunes entitats geogràfiques van quedar fora de la divisió original dividides entre les comarques oficials, com és el cas del Lluçanès i el Moianès, que es coneixen com a subcomarques. Amb la nova ampliació a les 41 comarques actuals, l'any 1988, s'han solucionat alguns d'aquests problemes.

En aquest treball ens plantegem la pregunta de quines són les característiques que defineixen les comarques de Catalunya i les seves relacions. Com que des d'un punt de vista antropològic un dels factors que més influeixen en el caràcter d'una població és l'activitat a què es dediquen els seus habitants, serà aquesta variable la que farem servir de base en el nostre estudi.

2. METODOLOGIA UTILITZADA

Com ja s'ha dit, les poblacions objecte del nostre estudi són les 41 comarques en què actualment es troba dividida Catalunya, i la variable principal del nostre estudi és la X: *Població activa per grups professionals* segons el cens de 1991 (dades facilitades per l'Institut d'Estadística de Catalunya), dividida en vuit categories:

X_1: Professionals i tècnics	X_2: Personal directiu
X_3: Serveis administratius	X_4: Comerciants i venedors
X_5: Hoteleria i altres	X_6: Agricultura i pesca
X_7: Indústria	X_8: Forces armades

L'elecció d'aquesta variable es fonamenta en el fet que quan es vol determinar el caràcter o, fins i tot, el grau de desenvolupament d'una població des d'un punt de vista antropològic, un primer factor a estudiar és com es reparteixen el recursos humans entre les diferents tasques que cal realitzar.

Una primera aproximació per determinar les semblances i les diferències entre les comarques podria ser la seva representació gràfica en un espai de dimensió reduïda segons el vector X, mitjançant el mètode denominat IDA, Instrinsic Data Analysis (Ríos *et al.*, 1994), com s'explica a la secció següent, però del qual volem destacar que, a diferència de la majoria de mètodes de reducció de la dimensió, es tracta d'una metodologia no específica d'un model estadístic, sinó que es pot aplicar a qualsevol amb moments de segon ordre finits.

L'IDA ens permet visualitzar les analogies entre les diferents comarques, però si el que volem és realitzar-ne una classificació en una sèrie de grups cal utilitzar alguna de les tècniques conegudes amb el nom d'anàlisi de conglomerats (*cluster analysis*). Sota aquest nom s'engloben un conjunt de tècniques que tenen per finalitat, a partir d'un conjunt inicial, obtenir una classificació dels objectes que el formen en un conjunt de grups que, dintre de cadascun, tinguin la màxima homogeneïtat possible, mentre que entre si l'heterogeneïtat sigui màxima respecte a les variables estudiades. En el nostre cas vam triar una tècnica aglomerativa jeràrquica denominada *average linkage clustering*, coneguda originalment com a UPGMA (Sokal i Sneath, 1963), que és una de les més utilitzades en la pràctica. L'UPGMA parteix d'una matriu d'interdistàncies entre els objectes a classificar i dóna com a resultat una classificació jeràrquica en forma de dendrograma. En el nostre cas, i d'acord amb la natura de les dades, vàrem triar la distància de *Bhattacharyya*, que presenta l'avantatge addicional de coincidir, excepte constants, amb la distància que utilitza l'IDA per al model multinomial.

Així doncs, el següent pas va consistir a calcular la matriu D d'interdistàncies entre totes les comarques. De manera que l'element d_{ij} de la matriu D representa la distància de Bhattacharyya entre la i-èsima i la j-èsima comarca calculada segons l'expressió:

$$(1) \qquad d_{ij} = d(p^{(i)}, p^{(j)}) = \operatorname{acos}\left(\sum_{k=1}^{8} \sqrt{p_{ik}p_{jk}}\right) \qquad i,j = 1,\ldots,41$$

on p_{ik} representa la probabilitat que un individu qualsevol de la i-èsima comarca pertany al k-èsim grup professional ($k = 1,\ldots,8$). Sobre la matriu D es va aplicar l'UPGMA per tal d'obtenir una classificació jeràrquica de les comarques.

Finalment, segons el dendrograma obtingut per l'UPGMA i la representació gràfica mitjançant l'IDA es van agrupar les 41 comarques en una sèrie de *clusters*.

L'últim pas d'aquest treball va consistir en l'estudi de les característiques de la classificació obtinguda i les relacions entre els diferents *clusters* obtinguts, tant per la

variable original X com per altres variables d'interès social (procedents del cens del 1991 o de l'any més proper disponible): nivell d'estudis, saldo migratori, densitat, renda per càpita, distribució per edats de la població activa, etc. A l'Apèndix es troben les dades originals, algunes de les quals procedeixen de l'Anuari Estadístic de Catalunya de 1992, i d'altres, encara no publicades en el moment d'elaborar aquest treball, ens han estat facilitades per l'Institut d'Estadística de Catalunya.

3. REPRESENTACIÓ MITJANÇANT L'IDA

3.1. Generalitats

Qualsevol mètode de representació gràfica pressuposa una determinada mètrica. Així, per exemple, l'Anàlisi de Components Principals (ACP) es basa en la distància euclidiana i l'Anàlisi de Correspondències (AC) en la distància khi-quadrat. L'elecció de la mètrica és arbitraria i dependrà de les propietats que es considerin més importants. L'IDA es basa en una mètrica riemanniana segons un enfocament que sintetitzem a continuació.

Donat un model paramètric amb moments de segon ordre finits, podem dotar-lo amb una estructura de varietat riemanniana, V, on el tensor mètric ve donat per la matriu d'informació de Fisher. Els punts d'aquesta varietat (V) representen poblacions estadístiques (mesures de probabilitat) que tenen com a coordenades els valors dels seus paràmetres (per exemple en el cas del model normal univariant, $N(\mu,\sigma)$ $\mu \in \mathbb{R}$, $\sigma > 0$, les coordenades de les poblacions a la varietat corresponent serien els parells (μ,σ)) i on la mesura de dissimilaritat entre poblacions ve donada per la distància riemanniana corresponent (coneguda en aquest cas com a distància de Rao). Aquesta aproximació permet utilitzar les eines de la geometria diferencial amb les seves propietats d'invariància versus canvis de coordenades (és a dir, versus transformacions admissibles dels paràmetres i de les variables). Una descripció molt més detallada d'aquest enfocament pot trobar-se a Amari (1985), Atkinson i Mitchell (1981), Bandorff-Nielsen (1984) Burbea i Rao (1982) i Oller (1989) entre d'altres.

Donades r poblacions pertanyents a un model probabilístic amb n paràmetres, l'IDA considera els passos següents:

- Representació de les poblacions $p^{(1)},\ldots,p^{(r)}$ a la varietat riemanniana V.
- Determinació de q, l'origen de la representació final, que es recomana que sigui el baricentre (centre de masses riemannià) de les poblacions.
- Mapatge de les poblacions des de V fins a V_q, l'espai tangent a la varietat al punt q, amb la mínima distorsió possible. Aquest pas es justifica pel fet que

V_q és un espai euclidià i, per tant, la posterior representació a un espai de dimensió reduïda mitjançant les tècniques clàssiques, com l'ACP, ja és factible. El mapatge des de V fins a V_q es realitza mitjançant la denominada *inversa de la funció exponencial al punt q* ($\exp_q^-$) que presenta la propietat de conservar les interdistàncies entre q i les poblacions.

- Finalment apliquem l'ACP als punts $m^{(1)},\ldots,m^{(r)}$ que representen les poblacions a l'espai V_q, tenint en compte que la matriu de la mètrica a V_q ve donada per G_q, la matriu d'informació de Fisher a q. Per tant, l'ACP es resumeix en sol·lucionar la següent diagonalització:

$$Cu_i = \lambda_i G_q^- u_i \tag{2}$$

on C representa la matriu de covariàncies (o qualsevol múltiple d'aquesta matriu) de les coordenades de les poblacions a V_q i u_i és el vector propi (ortonormal) associat al i-èsim valor propi λ_i. De manera que la projecció de V_q a $\mathbb{R}^s$ $(s \leq n)$ es realitza mitjançant:

$$X = MU \tag{3}$$

on M és la matriu $r \times n$ que té a les seves files les coordenades de les poblacions a V_q $(m^{(i)})$, U és la matriu $n \times s$ que té a les seves columnes els s primers vectors propis i X és la matriu $r \times s$ que té a les seves files les coordenades de les poblacions a $\mathbb{R}^s$ $(x^{(i)})$.

Esquema de la representació de les poblacions mitjançant l'IDA.

Varietat V	$\xrightarrow{\exp_q^-}$	Espai tangent V_q	$\xrightarrow{ACP}$	$\mathbb{R}^s$
$p^{(1)},\ldots,p^{(r)}$		$m^{(1)},\ldots,m^{(r)}$		$x^{(1)},\ldots,x^{(r)}$

Les variables es representen per les corbes de màxima variació de les seves esperances. Així, donada la variable X (amb esperança finita), podem definir el camp vectorial $\mathrm{grad}(E(x))$ (el gradient de l'esperança d'X) i associat a aquest tenim el flux integral tal que les seves corbes indiquen, localment, les direccions de màxima variació de l'esperança d'X. Amb això tenim les variables representades com a corbes a la varietat V i ja només cal repetir els passos indicats per a les poblacions: transport a V_q mitjançant $\exp_q^-$ i finalment projecció a $\mathbb{R}^s$.

Una descripció molt més acurada de l'IDA (incloent la representació de regions confidencials, efecte dels individus sobre l'estimació, etc.) pot trobar-se a Ríos *et al.* (1994).

3.2. Aplicació de l'IDA al nostre exemple

Tenim $r = 41$ poblacions (comarques) i considerem el model de Bernoulli multivariant (multinomial de mida $N = 1$) amb funció de densitat:

$$f_i(x_1,\ldots,x_8) = p_{i1}^{x_1} \cdot \ldots \cdot p_{i8}^{x_8} \qquad i = 1,\ldots,41 \tag{4}$$

$$x_j = \{0,1\} \qquad p_{ij} \in [0,1] \quad \text{i} \quad \sum_{j=1}^{8} p_{ij} = 1$$

on x_j $(j = 1,\ldots,8)$ són les variables indicadores de cada classe (grup professional) i p_{ij} és la probabilitat que un individu qualsevol de la i-èsima població pertany a la j-èsima classe. Les p_{ij} són, per tant, els paràmetres del nostre model. La geometria d'aquest model és ben coneguda i pot trobar-se a Atkison i Mitchell (1981), Oller (1982).

a) Representació de les poblacions

Pas a1: Representem les 41 poblacions (comarques) $p^{(1)},\ldots,p^{(41)}$ a la varietat V. Les coordenades estan donades pels paràmetres del model multinomial (les probabilitats de cada classe), és dir, les proporcions (p_{ij}) de cada grup professional a la població corresponent:

$$p^{(i)} = (p_{i1},\ldots,p_{i8}) \qquad i = 1,\ldots,41 \tag{5}$$

Pas a2: Determinem q, el baricentre de les poblacions, és a dir, el punt q de la varietat $(q \in V)$ que minimitza la suma de distàncies al quadrat a les poblacions:

$$\sum_{k=1}^{41} \left(d\left(p^{(i)},q\right)\right)^2 \leq \sum_{k=1}^{41} \left(d\left(p^{(i)},v\right)\right)^2 \qquad \forall v \in V \tag{6}$$

on $d\left(p^{(i)},q\right)$ representa la distància riemanniana (distància de Rao) entre $p^{(i)}$ i q, que en el model multinomial té la forma:

$$d\left(p^{(i)},q\right) = 2 \operatorname{acos}\left(\sum_{k=1}^{8} \sqrt{p_{ik}q_k}\right) \tag{7}$$

Les coordenades de q es resolen (6) numèricament i en el nostre cas s'obtingué:

$$q = (0.104, 0.020, 0.113, 0.118, 0.098, 0.107, 0.437, 0.003) \tag{8}$$

Pas a3: Transportem els punts $p^{(i)}$ a l'espai tangent a la varietat en el punt q, és dir a V_q, mitjançant la denominada *inversa de la funció exponencial* al punt q

$(\exp_q^-(\cdot))$. En aquest model, si tenim el punt $p^{(i)}$ de la varietat, les seves coordenades a l'espai tangent $(m^{(i)} = (m_{i1}, \ldots, m_{i7}) \in V_q)$ es calculen com:

$$m_{i\alpha} = \rho(i)\frac{\sqrt{p_{i\alpha}q_\alpha} - q_\alpha \cos(\rho(i)/2)}{\sin(\rho(i)/2)} \qquad \alpha = 1, \ldots, 7 \tag{9}$$

on $\rho(i)$ simbolitza la distància entre $p^{(i)}$ i q, és a dir $\rho(i) = d(p^{(i)}, q)$. Cal fer notar que la dimensió de V_q és 7, això és una conseqüència que la vuitena coordenada a V és redundant, ja que la suma de totes elles ha de ser 1.

Pas a4: Projecció de les poblacions des de V_q fins a l'espai de dimensió reduïda, en el nostre cas $\mathbb{R}^2$. Primer calculem la matriu d'informació de Fisher al punt q (G_q), que en aquest model té la forma:

$$G_q = g_{ij} = \left(\frac{\delta_{ij}}{q_i} - \frac{1}{q_8}\right) \qquad i, j = 1, \ldots, 7 \tag{10}$$

Definim com a M la matriu 41×7 que té a les files les coordenades de les poblacions i calculem C la matriu de covariàncies de les poblacions a V_q. Efectuem la diagonalització: $Cu_i = \lambda_i G_q^- u_i$ tal com es va indicar a (2). Com que estem interessats en una representació bidimensional, construïm la matriu U amb els dos primers vectors propis (u_1, u_2) i obtenim:

$$U' = \begin{pmatrix} 0.767 & 1.171 & 1.181 & 0.940 & 0.941 & -2.375 & 0.605 \\ 2.948 & 2.645 & 3.301 & 3.265 & 2.062 & 3.392 & 4.844 \end{pmatrix} \tag{11}$$

Finalment, les coordenades de la i-èsima població a $\mathbb{R}^2$ $(x^{(i)})$ es correspondrà amb la i-èsima fila de la matriu X calculada com: $X = MU$.

Si prenem com a exemple la comarca del Barcelonès $(p^{(13)})$, hem seguit els passos següents:

$$p^{(13)} = (0.171, 0.029, 0.214, 0.148, 0.112, 0.004, 0.320, 0,001)$$

$$m^{(13)} = \exp_q^-(p^{(13)}) = (0.071, 0.011, 0.098, 0.042, 0.024, -0.164, -0.080)$$

$$x^{(13)} = m^{(13)}U = (0.587, -0.196)$$

b) Representació de les variables

Pas b1: La corba integral, associada amb la variable X_i $i = 1, \ldots, i = 1, \ldots, 8$ que s'origina a q ve donada per:

$$\pi_i^{\alpha}(t) = \frac{q_\alpha(1+\delta_{i\alpha}(e^t-1))}{1+q_i(e^t-1)} \qquad \alpha = 1,\ldots,7 \tag{12}$$

on δ_{ij} representen les deltes de Kronecker. Per tant, pot transportar-se a l'espai tangent V_q mitjançant exp_q^-, repetint l'explicat per a les poblacions.

4. RESULTATS

La representació gràfica obtinguda mitjançant l'IDA es mostra a la figura 1, en què s'ha pres com a origen el baricentre de les poblacions, que es podria interpretar com la *comarca mitjana*. De la figura 1 es desprèn que el factor més important en la diferenciació de les comarques és l'agrícola (X_6), ja que la fletxa corresponent a aquesta variable és pràcticament paral·lela al primer component principal (que explica el 72.03 % de la variabilitat total) i és la de major longitud. Les comarques situades a l'esquerra són les que, proporcionalment, més recursos humans dediquen a les tasques agràries i de pesca. D'altra banda, la fletxa que correspon al component industrial (X_7) és pràcticament paral·lela al segon component principal (que explica el 17.20 % de la variabilitat total) i la segona de major longitud, indicant-nos que aquesta variable és la segona en importància en la discriminació entre les comarques segons els seus recursos humans. Les comarques representades en la part superior del gràfic són les que tenen una proporció més gran d'habitants dedicats a tasques industrials. El component *forces armades* (X_8) creix en la mateixa direcció, però en sentit oposat al component industrial, encara que qualsevol possible interpretació d'aquest fet seria qüestionable per l'escassa importància relativa d'aquest component. La resta de variables ($X_1,\ldots,X_5$), que corresponen majoritàriament a les activitats de tipus *serveis*, donen una informació aparentment redundant, ja que totes creixen segons una direcció i sentit molt semblant (inferior dreta de la fig. 1), sent la variable X_5 la de major longitud d'aquest grup de variables.

L'excel·lent relació entre les dades originals i la representació gràfica a través de l'IDA (fig. 1) es demostra pel fet que aquesta explica pràcticament el 90 % de la variabilitat original i pel coeficient de correlació entre les 820 interdistàncies originals de les 41 comarques i les interdistàncies de la figura 1, que té un valor de 0.9888.

La figura 1 ja ens permet visualitzar la formació de grups, però per establir una agrupació més objectiva vàrem utilitzar l'UPGMA. El dendrograma resultant es mostra a la figura 2. Es va tallar a un nivell de $d = 0.294$, ja que d'aquesta manera vàrem obtenir un nombre de *clusters* o *grups* raonable, nou, que s'han identificat a la fig. 1, mostrant la bona concordança entre els resultats de l'UPGMA i de l'IDA. A la taula 1 es detalla la classificació obtinguda i la numeració dels nou *clusters*; a la figura 3 es representen les mitjanes de la variable X per a cada *grup*.

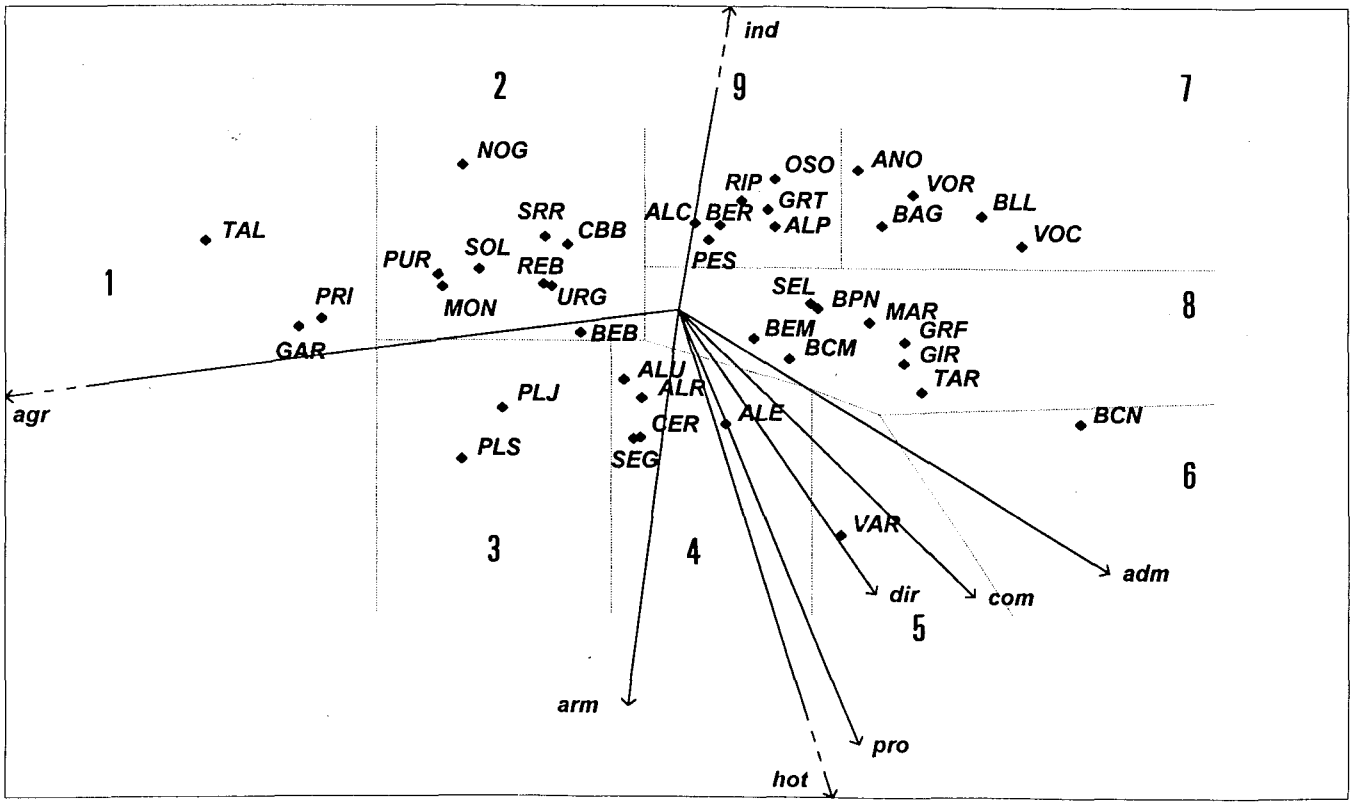


Figura 1

Representació bidimensional de les 41 comarques de Catalunya i de les 8 variables (grups professionals) obtinguda mitjançant l'IDA, corresponent al cens de 1991. L'origen de la representació correspon al baricentre de les poblacions. Cada comarca s'identifica per un codi de tres lletres (majúscules) que es mostra a la Taula 1. Les fletxes representen els grups professionals (identificats per les tres primeres lletres) i les línies discontínues separen els 9 grups en què es classifiquen les comarques.

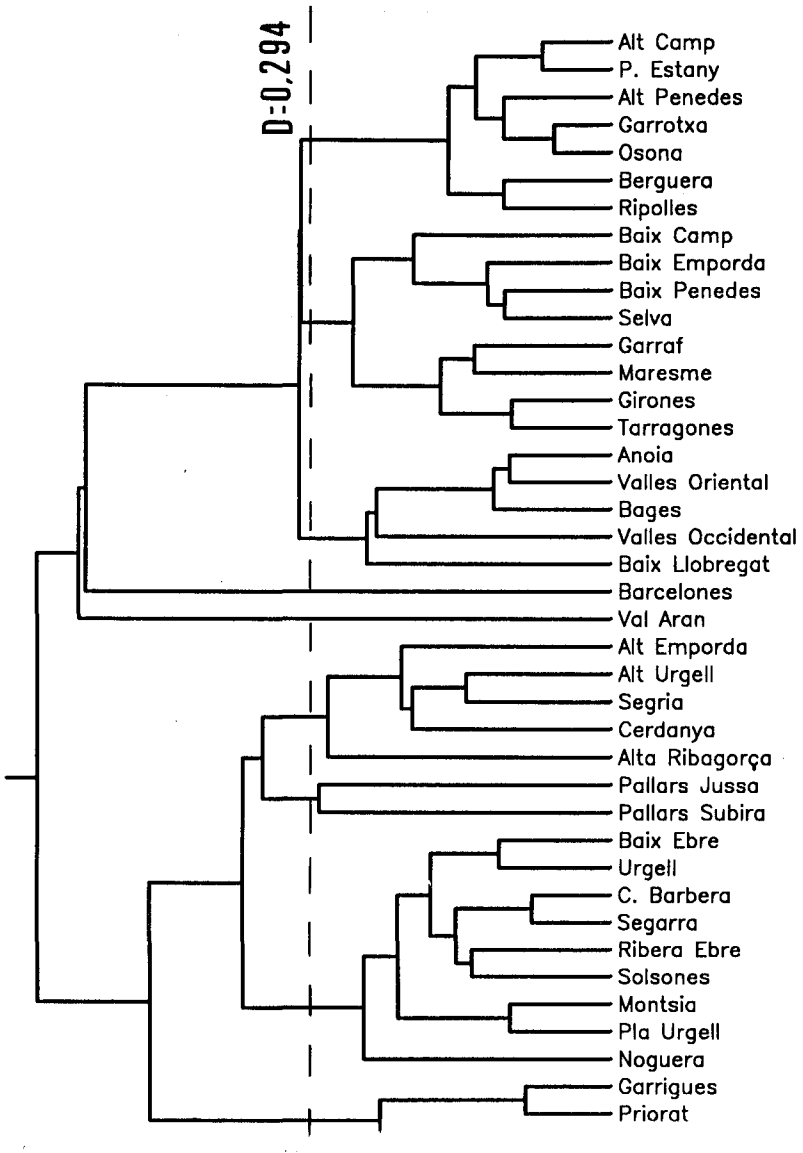


Figura 2

Dendrograma obtingut per l'UPGMA per les 41 comarques utilitzant la distància de Bhattacharyya. S'ha indicat la línia de tall, situada a d = 0.294, que dona els nou grups en què hem agrupat les comarques.

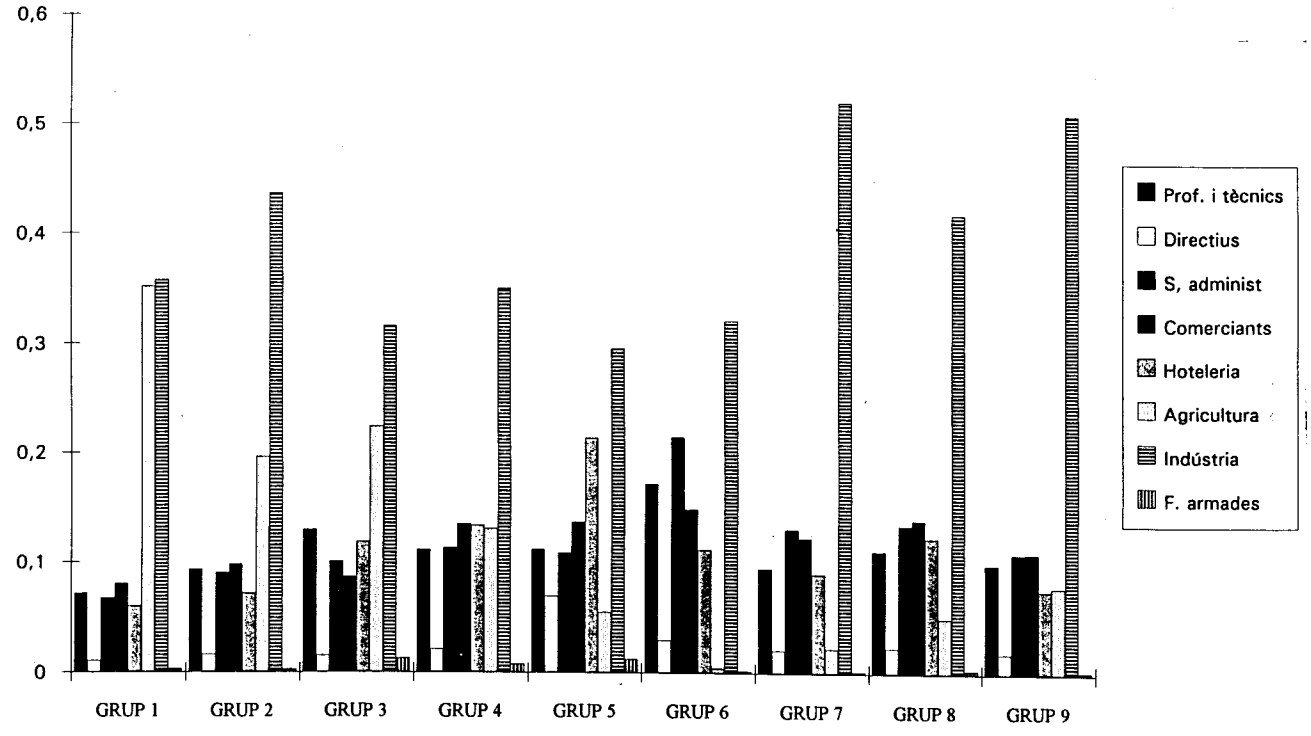


Figura 3

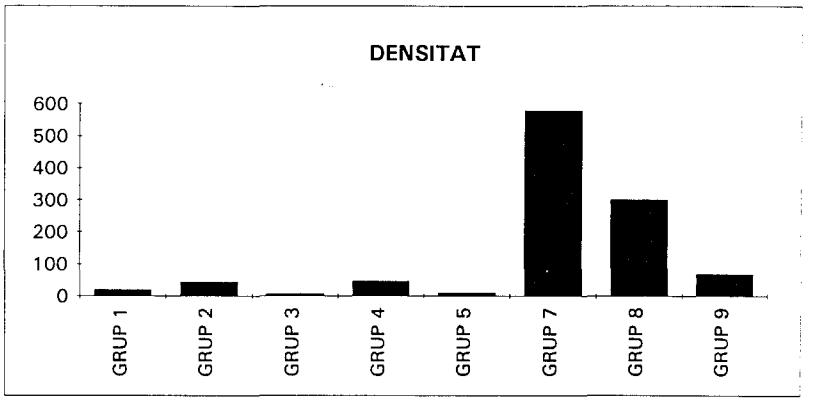
Proporcions mitges dels vuit grups professionals, variable **X**, pels nou grups en què s'han agrupat les quaranta-i-una comarques.

L'última part d'aquest treball va consistir a estudiar les característiques dels nou grups obtinguts. De les diverses variables estudiades, les que finalment van resultar més informatives van ser: la densitat (h./km^2), el saldo migratori, la renda familiar per càpita i la distribució per grups d'edat de la població activa. Els valors d'aquestes variables per a cada *grup* es troben a la figura 4.

5. DISCUSSIÓ I CONCLUSIONS

Amb vista a altres interpretacions i generalitzacions, cal no oblidar que en el nostre treball hem volgut donar el mateix pes a totes les comarques, ja que ens interessava, justament, estudiar-ne les relacions i característiques independentment del nombre d'habitants. És evident que si haguéssim treballat en nombres absoluts la comarca del Barcelonès hauria disfressat la pràctica totalitat dels resultats, ja que hi viuen el 38% dels habitants de Catalunya. En aquest sentit, una alternativa a la representació mitjançant l'IDA hauria estat l'Anàlisi de Correspondències (AC). És ben conegut que l'AC, per fer servir la distància khi quadrat, depèn de les mides mostrals. Quan es vol evitar aquesta dependència alguns autors (Greenacre, 1984; Greenacre,1993) recomanen realitzar l'AC sobre la taula de freqüències relatives (obtinguda dividint cada fila pel nombre total d'individus de la comarca corresponent). Els resultats, així obtinguts, es mostren a la figura 5 segons la denominada forma biplot en la que les files (comarques) es representen segons les seves coordenades principals i les fletxes representen les columnes (grups professionals) segons la direcció dels seus vèrtexs, però reescalades mitjançant la multiplicació de les seves coordenades estàndards per les freqüències relatives de cada grup professional (veure els capítols 10 i 13 de Greenacre, 1993, per a una excel·lent explicació). Una altra característica de la distància khi quadrat és que la distància entre dues poblacions també depèn de la resta de poblacions que considerem, ja que a la fórmula de la distància khi quadrat intervenen les sumes parcials tant per files com per columnes. Això pot provocar que la configuració obtinguda amb l'AC per a un conjunt de poblacions es vegi modificada si s'afegeix un segon grup de poblacions, tal com es mostra a l'exemple 4.2.1 de Rao (1995). En canvi, l'IDA utilitza la distància de Rao que no depèn de mides mostrals i la distància entre dues poblacions només depèn dels valors observats per a aquestes, però no de la resta de poblacions considerades. De fet, en estar basat en tècniques de geometria diferencial, l'IDA és invariant versus transformacions admissibles de les variables i/o els paràmetres. Així mateix, com per a la segona part del treball, l'elaboració d'una classificació jeràrquica de les comarques, van triar la distància de Bhattacharyya, semblava més coherent triar l'IDA com a mètode de representació, encara que, evidentment, l'AC hauria estat una alternativa (la similitud entre les figures 1 i 5 és molt clara), com també ho hauria estat el mètode descrit a Rao (1995) i la distància d'Hellinger.

a)



b)

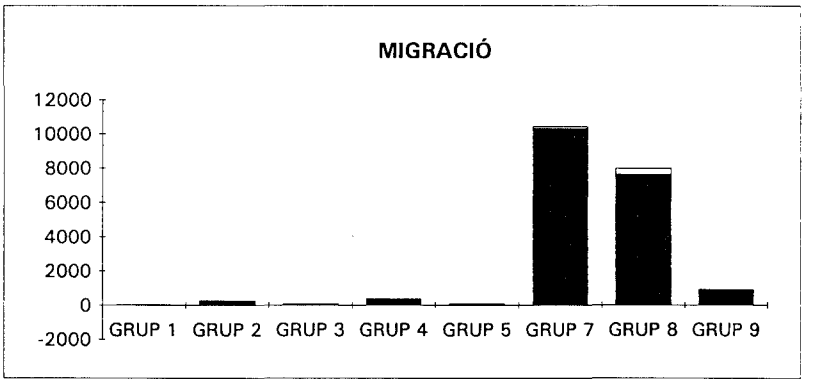


Figura 4

Algunes característiques socio-demogràfiques pels 9 grups en què hem classificat les comarques de Catalunya:

a) Densitat (h./km^2),

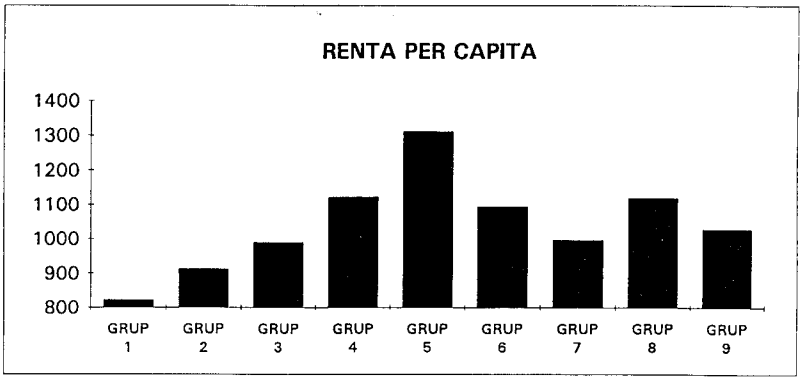
b) Saldo migratori total,

c) Renda familiar disponible per càpita, i

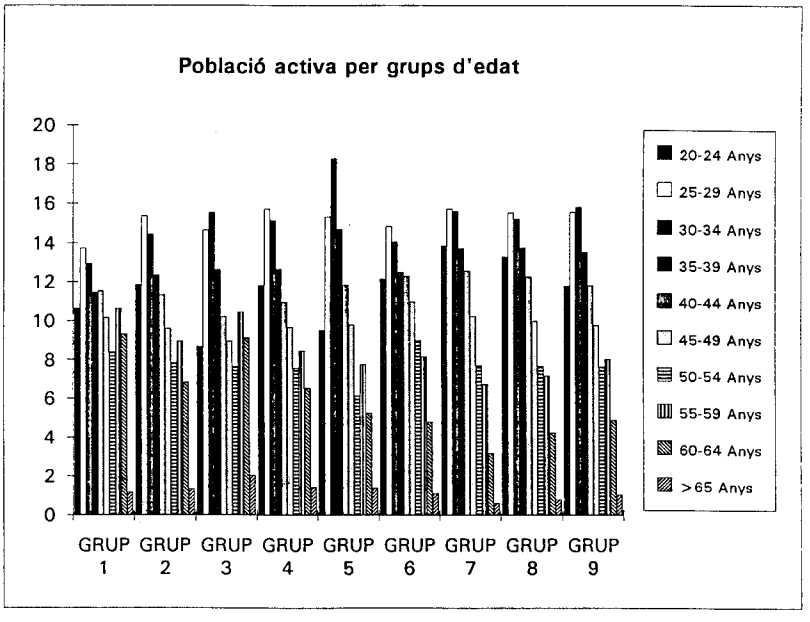
d) Distribució de la població activa, en percentatge, per grups d'edat.

En *a)* i *b)* s'ha omès el grup 6, corresponent al Barcelonès, pels seus valors atípics: densitat de 16.091 hab/Km^2 i saldo migratori negatiu de 21121 habitants, dividit en 19330 cap a Catalunya i 1791 cap a la resta d'Espanya.

c)


RENTA PER CAPITA
1400
1300
1200
1100
1000
900
800
GRUP 1
GRUP 2
GRUP 3
GRUP 4
GRUP 5
GRUP 6
GRUP 7
GRUP 8
GRUP 9

d)


Població activa per grups d'edat
20
18
16
14
12
10
8
6
4
2
0
GRUP 1
GRUP 2
GRUP 3
GRUP 4
GRUP 5
GRUP 6
GRUP 7
GRUP 8
GRUP 9
20-24 Anys
25-29 Anys
30-34 Anys
35-39 Anys
40-44 Anys
45-49 Anys
50-54 Anys
55-59 Anys
60-64 Anys
>65 Anys

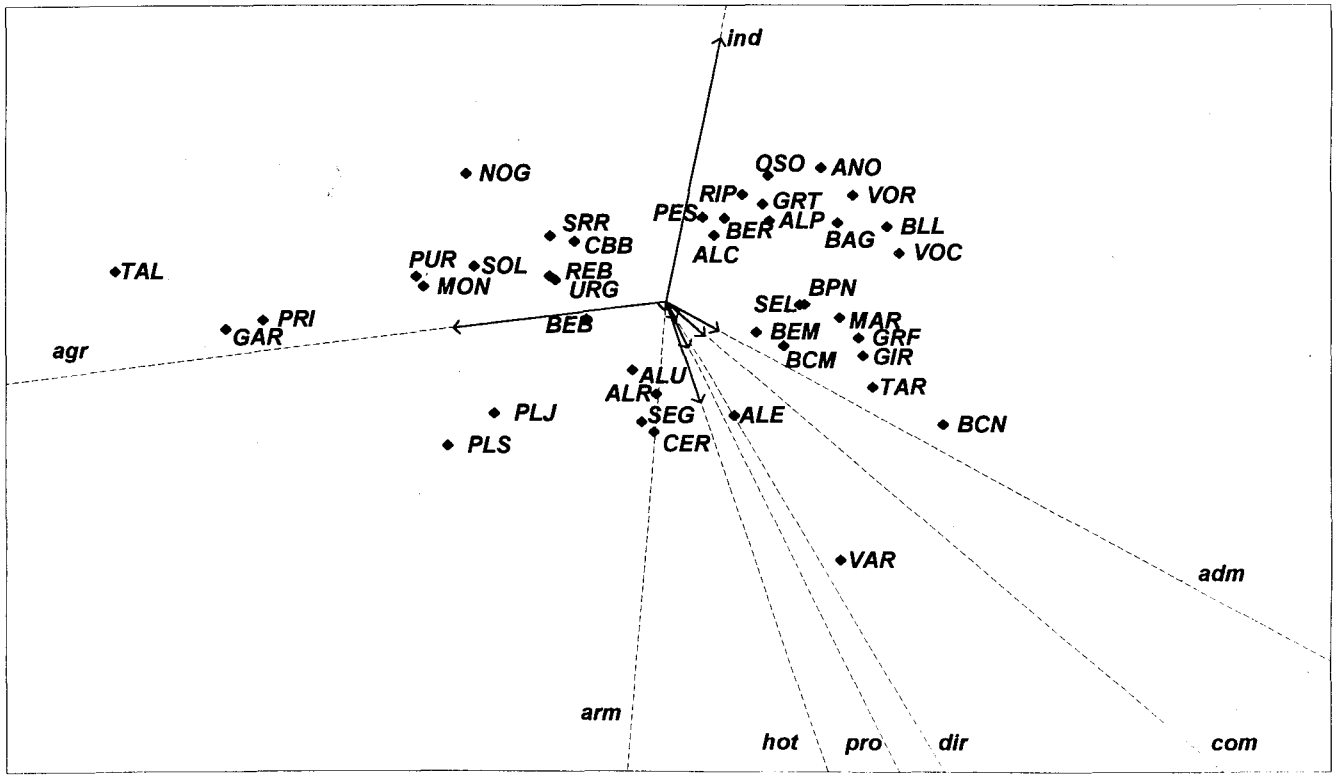


Figura 5

Representació bidimensional de les 41 comarques de Catalunya i de les 8 variables (grups professionals) obtinguda mitjançant l'Anàlisi Factorial de Correspondències (AC) aplicat a les mateixes dades que la figura 1. L'AC es mostra en la seva forma biplot (veure el text per més detalls) aplicat sobre les freqüències relatives. Les fletxes representen els grups professionals i s'han prolongat amb línies discontínues per tal de visualitzar amb claredat les direccions.

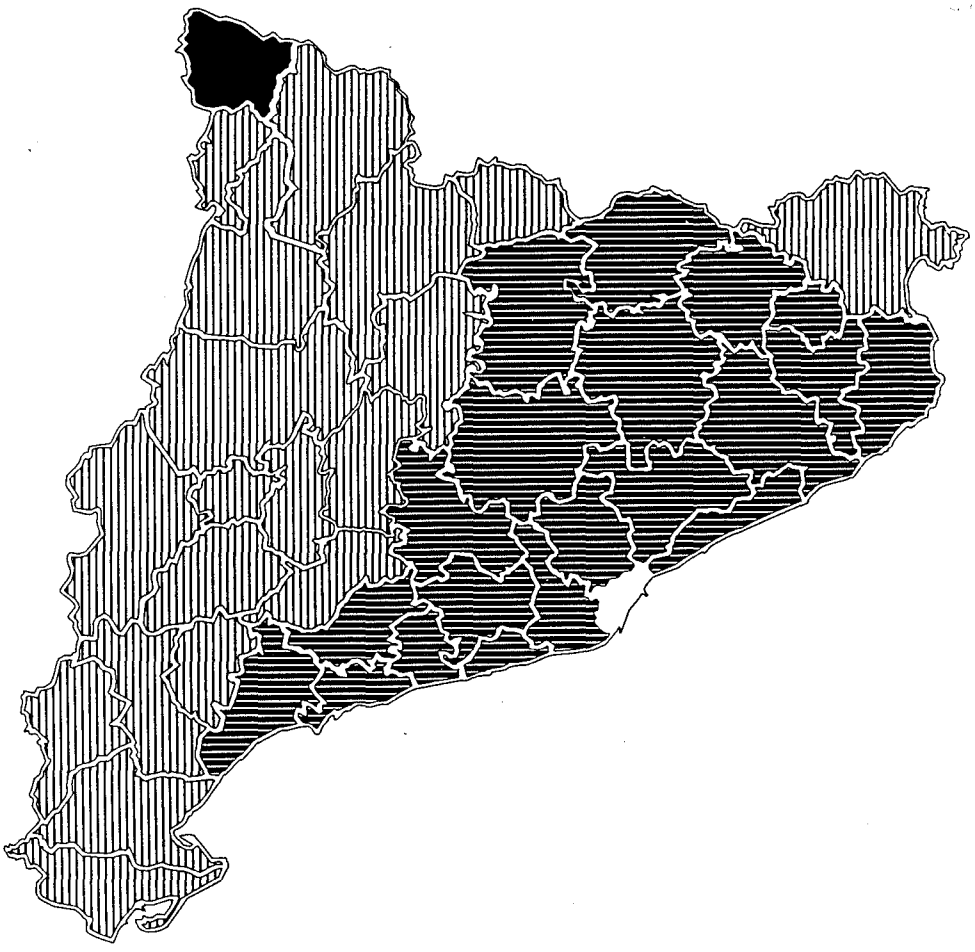


Figura 6

Distribució geogràfica dels blocs en que hem classificat les comarques de Catalunya:
a) BLOC AGRARI, identificat amb línies verticals,
c) BLOC TURÍSTIC, en negre
b) BLOC ADMINISTRATIU, en blanc
c) BLOC INDUSTRIAL, amb línies horitzontals

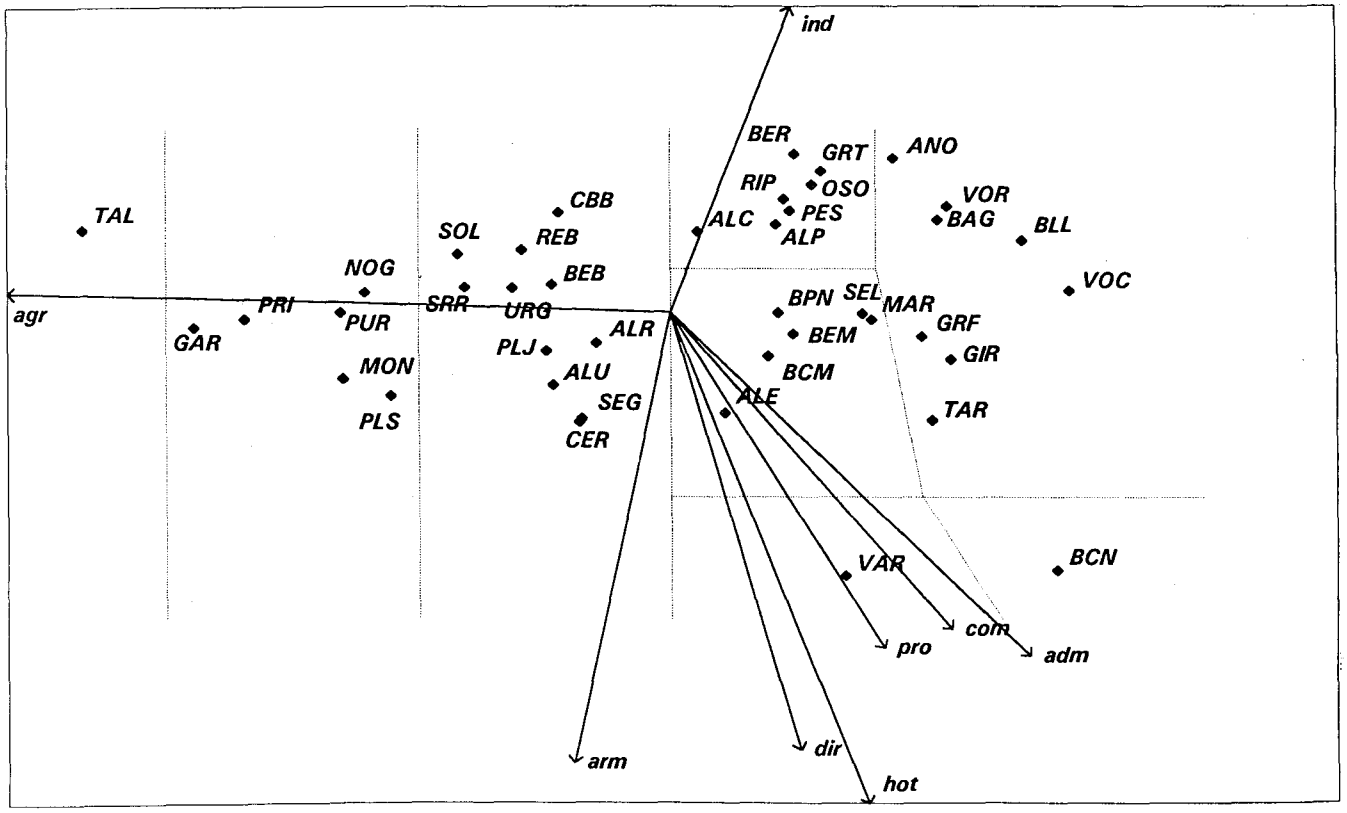


Figura 7

Representació bidimensional de les 41 comarques de Catalunya i de les 8 variables (grups professionals) obtinguda mitjançant l'IDA, corresponent al cens de 1986. La terminologia utilitzada és la mateixa que la de la fig. 1.

Taula 1

Classificació de les comarques de Catalunya obtinguda segons l'IDA (model multinomial) i l'UPGMA (distancia de Bhattacharyya).

	Grup 1		**Grup 6**
GAR:	Garrigues	BCN:	Barcelonès
PRI:	Priorat		
TAL:	Terra Alta	**Grup 7**	
		ANO:	Anoia
	Grup 2	BAG:	Bages
BEB:	Baix Ebre	BLL:	Baix Llobregat
CBB:	Conca de Barberà	VOC:	Vallès Occidental
MON:	Montsià	VOR:	Vallès Oriental
NOG:	Noguera		
PUR:	Pla d'Urgell		**Grup 8**
REB:	Ribera d'Ebre	BCM:	Baix Camp
SRR:	Segarra	BEM:	Baix Empordà
SOL:	Solsonès	GRF:	Garraf
URG:	Urgell	GIR:	Gironès
	Grup 3	MAR:	Maresme
PLJ:	Pallars Jussà	SEL:	Selva
PLS:	Pallars Sobirà	TAR:	Tarragonès
	Grup 4		**Grup 9**
ALE:	Alt Empordà	ALC:	Alt Camp
ALR:	Alta Ribagorça	ALP:	Alt Penedès
ALU:	Alt Urgell	BER:	Berguedà
CER:	Cerdanya	GRT:	Garrotxa
SEG:	Segrià	OSO:	Osona
		PES:	Pla de l'Estany
	Grup 5	RIP:	Ripollès
VAR:	Val d'Aran		

Un resultat important del present estudi és que la divisió comarcal de Catalunya es pot interpretar en termes de les variables aquí estudiades (a més del factor geogràfic que comentarem posteriorment) i d'acord amb altres estudis basats en diferents tipus de dades (Calapell i Hernández, 1993). Així mateix, també és remarcable que sigui el component agrícola (X_6) el més important en la seva diferenciació (figura 1), ja que cal no oblidar que l'existència de mercats agrícoles va ser precisament un dels factors claus en la divisió original del territori de Catalunya en comarques. Per tant, a pesar de l'evolució soferta i de la creixent industrialització del seu territori (hem vist

que en tots els grups era sempre el component industrial el majoritari), el component agrícola segueix sent el que millor explica les diferències intercomarcals.

Com hem vist a la secció de *Resultats*, hem classificat les comarques de Catalunya en nou *grups* que, a la vista dels resultats anteriors, es poden agrupar en quatre grans blocs amb les característiques essencials següents:

- **Bloc agrícola**: Comprèn els *clusters* 1, 2, 3 (situats a l'esquerra de la fig 1) i 4 (part central inferior) en què el component X_6 (agricultura i pesca) té més importància (figures 1 i 3), encara que convé no oblidar que el sector a què més recursos humans dediquen els nou *grups* és l'industrial. Aquest bloc també es caracteritza, en general, per un saldo migratori molt baix (positiu o negatiu), una densitat baixa i una renda familiar per càpita baixa. Cal assenyalar que el *grup* 4, encara que d'acord amb l'UPGMA s'engloba clarament en aquest bloc, presenta certes característiques particulars: els components d'X corresponents al sector serveis presenten unes freqüències relativament altes (figura 3) i la renda familiar per càpita és superior a la de la resta dels grups (figura 4c). Aquestes peculiaritats concorden perfectament amb la seva posició a la figura 1, molt menys extrema que la de la resta de grups d'aquest bloc, i podrien justificar-se per la importància del sector turístic en aquestes comarques.

- **Bloc turístic**: Format pel *cluster* 5 (part inferior central de la fig. 1) que inclou únicament la Vall d'Aran. Es caracteritza per la importància del sector *serveis* (professionals (X_1), administratius (X_3), comerciants (X_4) i, especialment, hotelers (X_5)), que s'explicaria per la importància del turisme en aquesta comarca. Així mateix, es caracteritza per una elevada renda per càpita (fig. 4c) i una densitat baixa.

- **Bloc administratiu**: Inclou el *cluster* 6 (part inferior dreta de la fig. 1), format únicament pel Barcelonès (la comarca que inclou Barcelona), amb unes connotacions molt especials. Els sectors professionals i comercials estan àmpliament representats, però és sobretot la importància del sector administratiu el que més el diferencia de la resta. Aquest fet s'explica fàcilment pel seu paper de capital administrativa del principat. Altres característiques d'aquest bloc són: densitat molt elevada (16.901 h./km^2), saldo migratori molt negatiu dirigit principalment a la resta de Catalunya i pràctica absència del component agricultura i pesca. També resulta destacable el fet que, en contra de les idees preestablertes que es poden tenir, en el Barcelonès el percentatge de la població activa que es dedica a la indústria és inferior al dels *grups* que l'envolten geogràficament, probablement a causa de la descentralització creixent d'aquest sector i la seva importància administrativa.

- **Bloc industrial**: Inclou els *grups* 7, 8, i 9 (part superior dreta de la fig. 1), caracteritzats per un component industrial molt alt, mentre que el component

agrícola és poc important. El saldo migratori és positiu (provinent de la resta d'Espanya i de Catalunya), la densitat elevada (encara que no comparable a la del Barcelonès) i la renda per càpita intermèdia.

Respecte a la distribució per grups d'edat (fig. 4d), cal indicar que les diferències no són tant clares com a les variables ja comentades, encara que s'observen algunes peculiaritats. Així, en el *bloc agrícola* s'observa una distribució bimodal amb dos màxims: el primer corresponent als grups de 25-29 i 30-34 anys, i el segon situat als 55-59 anys. En el *bloc turístic* el segon màxim gairebé no s'aprecia, i es confirma aquesta tendència en els *blocs administratiu i industrial,* en què el segon màxim ha desaparegut totalment. També cal tenir en compte que aquesta divisió en blocs no és absoluta, sinó que presenta una gradació, de manera que els grups (i per tant les comarques incloses) que a la fig. 1 es troben pròxims, encara que pertanyin a blocs diferents, poden presentar certes similituds. Així, prenent com a exemple el *bloc* agrícola, queda clar que el *cluster* 1 (el més extrem del bloc) presenta totes les característiques que defineixen aquest bloc, mentre que el *cluster* 4 (de posició més intermèdia) ja hem vist que presenta certes similituds amb les d'altres blocs.

Es destacable que aquests quatre *blocs* presenten una clara gradació geogràfica, tal com es mostra a la figura 6. Així, el *bloc industrial* correspon a les comarques orientals, de manera que observem clarament una línia divisòria amb les comarques de ponent que corresponen al *bloc agrícola* amb l'única excepció de l'Alt Empordà. Els dos, formats només per una comarca *turística i administrativa,* es troben intercalats entre els blocs anteriors.

Un aspecte interessant, amb vista a estudis posteriors, és el seguiment de l'evolució de la variable X al llarg del temps. En aquest sentit es va repetir aquest estudi amb les dades de l'any 1986 (Anuari Estadístic de Catalunya, 1990), ja que no es disposa de dades per comarques de censos anteriors. A la figura 7 se sintetitzen els resultats obtinguts per l'IDA i l'UPGMA (tallant també per d = 0.294), i s'obté una classificació en 8 *grups*. Com era d'esperar, la interpretació de la figura 6 és molt semblant a la de la figura 1, de manera que continua sent la variable X_6 (agricultura i pesca) la que millor discrimina les relacions entre les comarques, seguida d'X_7 (indústria). Sintetitzant els resultats podem dir que la classificació en blocs s'ha mantingut pràcticament inalterable i només una comarca, l'Alt Empordà, ha canviat de bloc. Respecte als grups, l'any 1986, el Barcelonès i la Vall d'Aran també es presentaven en blocs individuals; els grups del *bloc industrial* presenten una gran estabilitat. Per contra, els grups agrícoles són els que més han evolucionat durant aquest període, de manera que a l'any 1986 només es distingien tres grups de composició clarament diferent a la dels quatre actuals. Resultarà interessant comprovar si la tendència del grup 4 cap al sector serveis es confirma en propers censos.

Per últim, respecte a la metodologia utilitzada, volem destacar la fàcil interpretació de la representació obtinguda pel mètode IDA, com també la seva concordança amb els

resultats obtinguts a partir de l'UPGMA. Tot això mostra com les tècniques basades en l'aplicació de la geometria diferencial a l'estadística poden ser utilitzades sense gaires dificultats i obtenir resultats òptims en casos pràctics. Amb aquest treball esperem contribuir a disminuir la reticència a la utilització d'aquestes tècniques en el camp aplicat, justificada, en part, pel caràcter tan teòric que solen tenir les publicacions d'aquesta branca de l'estadística.

AGRAÏMENTS

Els autors volen agrair al Sr. Jordi Oliveras i Prats, director de l'Institut d'Estadística de Catalunya de la Generalitat de Catalunya, el fet d'haver-nos facilitat moltes de les dades utilitzades en aquest treball corresponents al cens de 1991 i que encara no han estat publicades.

APÈNDIX

Dades originals empleades

Població ocupada per grups professionals (1991).

Dades facilitades per l'Institut d'Estadística de Catalunya.

Comarca	Professionals i tècnics	Personal directiu	Serveis administratius	Comerciants i venedors	Hoteleria i altres	Agricultura i pesca	Indústria	Forces armades
Alt Camp	1231	243	1446	1420	875	1265	6286	25
Alt Empordà	2948	793	5040	5510	4823	3509	12083	317
Alt Penedès	2419	502	3667	3077	2000	1827	13118	36
Alt Urgell	778	135	835	1020	798	1068	2777	79
Alta Ribagorça	175	23	98	131	199	163	469	1
Anoia	2764	614	3462	3556	2408	1124	17472	43
Bages	6274	1022	6485	7095	4570	1755	28255	171
Baix Camp	5699	989	6165	7029	5221	3270	18436	110
Baix Ebre	2446	383	2311	2808	1994	3682	8846	65
Baix Empordà	2810	737	3716	4900	4635	2747	14519	127
Baix Llobregat	12371	4009	31296	26849	24955	2605	110826	274
Baix Penedès	1116	320	1705	1997	1762	785	6305	49
Barcelonès	146521	24845	182813	126740	95496	3462	274395	1258
Berguerà	1373	164	1207	1555	1131	1129	6910	78
Cerdanya	492	116	462	679	786	670	1695	38
Conca de Barberà	563	124	636	631	488	1068	3018	7
Garraf	3484	549	3419	3875	3559	836	11448	43
Garrigues	539	79	524	619	424	2338	2286	13
Garrotxa	1909	390	2064	2037	1420	1264	9712	32
Gironès	7315	1187	8884	7173	5127	1727	19917	269
Maresma	12837	3475	15056	15560	10867	4504	45818	189
Montsià	1329	282	1600	2046	1394	4588	7716	77
Noguera	1131	185	931	1226	824	3215	7911	35
Osona	4901	901	5277	5423	3238	3076	26436	50
Pallars Jussà	567	79	479	465	410	955	1530	101
Pallars Sobirà	280	27	200	148	307	497	620	6
Pla d'Urgell	863	169	1019	1020	597	2570	4200	24
Pla de l'Estany	923	187	1036	881	587	804	4004	8
Priorat	287	34	245	255	232	1063	1179	10
Ribera d'Ebre	936	75	684	657	592	1318	3263	27
Ripollès	1012	193	905	1106	1006	801	5908	27
Segarra	654	125	653	560	415	1152	3023	6
Segrià	7841	1279	8280	8294	6253	8678	18970	577
Selva	2776	744	4106	4720	5758	2149	17562	66
Solsonès	431	61	330	315	348	900	1854	6
Tarragonès	8047	1201	9403	7294	7309	1640	21352	348
Terra Alta	217	41	220	324	209	1757	1710	16
Urgell	1020	235	1099	1431	758	1991	4699	31
Val d'Aran	295	182	286	360	562	143	779	32
Vallès Occidental	28614	5383	34772	31343	21310	1610	114191	231
Vallès Oriental	9550	2250	13548	11619	8395	2499	54530	122

Densitat de població (any 1991).

Dades reproduïdes de l'Anuari Estadístic de Catalunya de 1992.

Comarca	hab./Km^2
Alt Camp	62.5
Alt Empordà	67.6
Alt Penedès	117.9
Alt Urgell	13.0
Alta Ribagorça	8.2
Anoia	95.1
Bages	117.5
Baix Camp	189.3
Baix Ebre	65.4
Baix Empordà	128.4
Baix Llobregat	1254.2
Baix Penedès	128.9
Barcelonès	16091.0
Berguerà	33.0
Cerdanya	22.7
Conca de Barberà	27.7
Garraf	417.8
Garrigues	24.3
Garrotxa	62.7
Gironès	218.7
Maresma	738.5
Montsià	76.6
Noguera	20.1
Osona	92.9
Pallars Jussà	10.0
Pallars Sobirà	4.0
Pla d'Urgell	94.6
Pla de l'Estany	80.2
Priorat	19.1
Ribera d'Ebre	27.9
Ripollès	28.3
Segarra	23.6
Segrià	116.9
Selva	98.7
Solsonès	10.8
Tarragonès	491.5
Terra Alta	17.5
Urgell	50.8
Val d'Aran	10.0
Vallès Occidental	1118.9
Vallès Oriental	308.2

Moviments migratoris (any 1990).

Dades reproduïdes de l'Anuari Estadístic de Catalunya de 1992.

Comarca	Amb la resta de Catalunya	Amb la resta d'Espanya	Saldo total
Alt Camp	116	11	127
Alt Empordà	174	237	411
Alt Penedès	489	30	519
Alt Urgell	67	120	187
Alta Ribagorça	10	10	20
Anoia	407	89	496
Bages	150	−44	106
Baix Camp	536	385	921
Baix Ebre	42	−5	37
Baix Empordà	444	333	777
Baix Llobregat	3236	186	3422
Baix Penedès	786	55	841
Barcelonès	−19330	−1791	−21121
Berguerà	−79	−2	−81
Cerdanya	47	24	71
Conca de Barberà	25	−30	−5
Garraf	1064	97	1161
Garrigues	−7	−9	−16
Garrotxa	52	58	110
Gironès	515	295	810
Maresma	3137	276	3413
Montsià	91	50	141
Noguera	1	13	14
Osona	152	41	193
Pallars Jussà	35	−10	25
Pallars Sobirà	39	−8	31
Pla d'Urgell	−16	−6	−22
Pla de l'Estany	186	−6	180
Priorat	−4	−2	−6
Ribera d'Ebre	−64	−20	−84
Ripollès	−76	−22	−98
Segarra	49	19	68
Segrià	7	35	42
Selva	722	472	1194
Solsonès	72	−15	57
Tarragonès	386	597	983
Terra Alta	0	−17	−17
Urgell	44	32	76
Val d'Aran	0	70	70
Vallès Occidental	3351	312	3663
Vallès Oriental	3144	395	3539

Renda familiar disponible per càpita (any 1989).

Dades reproduïdes de l'Anuari Estadístic de Catalunya de 1992.

Comarca	Milers de PTA
Alt Camp	990.3
Alt Empordà	1221.6
Alt Penedès	1063.1
Alt Urgell	978.6
Alta Ribagorça	1123.1
Anoia	1009.3
Bages	961.5
Baix Camp	1016.9
Baix Ebre	993.8
Baix Empordà	1202.2
Baix Llobregat	959.1
Baix Penedès	1153.7
Barcelonès	1095.6
Berguerà	895.1
Cerdanya	1340.2
Conca de Barberà	916.0
Garraf	1103.0
Garrigues	776.7
Garrotxa	1090.2
Gironès	1103.1
Maresma	1126.6
Montsià	1001.7
Noguera	874.2
Osona	1060.2
Pallars Jussà	953.8
Pallars Sobirà	1026.0
Pla d'Urgell	908.0
Pla de l'Estany	1111.7
Priorat	834.5
Ribera d'Ebre	918.3
Ripollès	994.6
Segarra	811.2
Segrià	945.8
Selva	1171.3
Solsonès	910.6
Tarragonès	1084.4
Terra Alta	853.5
Urgell	873.3
Val d'Aran	1311.8
Vallès Occidental	981.7
Vallès Oriental	1078.4

Població activa per grups d'edat (any 1991).

Dades facilitades per l'Institut d'Estadística de Catalunya.

Comarca	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	> 65
Alt Camp	1530	1936	2012	1622	1385	1253	900	888	607	111
Alt Empordà	4274	5162	4704	4216	3906	3268	2648	2755	1902	366
Alt Penedès	3356	3958	3847	3334	2925	2442	1779	1847	1356	237
Alt Urgell	863	1064	1088	891	814	696	500	622	523	138
Alta Ribagorça	129	208	210	163	110	110	94	104	81	21
Anoia	3859	4673	4597	4046	3788	3131	2324	2236	1054	213
Bages	6629	8348	8687	7280	6520	5321	3939	3963	1949	434
Baix Camp	5878	7014	6880	5888	5480	4446	3351	3153	2071	369
Baix Ebre	2748	3156	2908	2520	2613	2207	1884	1926	1269	258
Baix Empordà	4423	5049	4705	4194	3750	3201	2510	2551	1679	290
Baix Llobregat	31717	32662	30854	28056	27066	22686	17570	13218	5682	870
Baix Penedès	1902	2108	1895	1742	1637	1302	1065	987	575	117
Barcelonès	100357	122492	116025	103243	101586	90528	74228	67333	39678	9326
Berguerà	1425	2122	2257	1857	1572	1232	915	891	554	174
Cerdanya	592	699	693	630	544	491	352	383	267	79
Conca de Barberà	743	963	881	742	694	632	477	584	495	88
Garraf	3488	3970	4067	3698	3184	2566	2086	1747	882	169
Garrigues	744	958	857	788	726	684	518	707	538	62
Garrotxa	1881	2621	2821	2510	2243	1796	1492	1700	916	198
Gironès	6284	7488	7661	7080	6037	4810	3536	3527	2166	490
Maresma	13019	15634	15642	14736	13548	10879	8136	7222	3652	748
Montsià	2355	2635	2396	2128	2027	1834	1584	1664	1302	200
Noguera	1350	1873	1693	1458	1311	1108	917	1122	989	147
Osona	5767	7501	7599	6578	5577	4385	3480	3594	1827	470
Pallars Jussà	400	670	617	532	496	412	374	517	399	94
Pallars Sobirà	173	294	354	275	192	179	143	192	192	41
Pla d'Urgell	920	1278	1258	1112	931	767	587	653	475	118
Pla de l'Estany	1330	1646	1437	1201	1052	919	768	884	651	113
Priorat	324	448	430	359	371	322	258	348	303	46
Ribera d'Ebre	701	1057	1134	951	853	750	615	715	506	66
Ripollès	1036	1492	1644	1506	1384	1121	931	946	444	102
Segarra	779	1093	944	782	710	523	422	570	469	88
Segrià	6973	9248	8892	7633	6820	5671	4444	4443	2994	557
Selva	4912	5760	5599	4836	4238	3553	2739	2568	1492	242
Solsonès	477	623	601	472	467	370	349	352	290	87
Tarragonès	6879	8273	8412	7722	6818	5501	4037	3773	2348	391
Terra Alta	453	547	537	485	518	433	402	445	448	50
Urgell	1354	1762	1531	1274	1154	1031	782	1042	791	125
Val d'Aran	245	395	471	379	305	252	158	200	135	36
Vallès Occidental	31836	35907	35347	32083	28660	22408	16614	13421	6426	1220
Vallès Oriental	14194	15082	15229	13231	11841	9578	7326	6226	3110	575

Població ocupada per grups professionals (1986).

Dades reproduïdes de l'Anuari Estadístic de Catalunya de 1991.

Comarca	Professionals i tècnics	Personal directiu	Serveis administratius	Comerciants i venedors	Hoteleria i altres	Agricultura i pesca	Indústria	Forces armades	Altres
Alt Camp	1028	280	913	1137	654	1417	5455	15	334
Alt Empordà	2153	564	3418	4757	3673	3485	10448	287	871
Alt Penedès	1785	435	2369	2086	1219	1906	10757	10	707
Alt Urgell	649	238	593	580	515	1345	2528	98	447
Alta Ribagorça	124	20	65	91	96	180	427	4	69
Anoia	2032	301	2432	2831	1661	1226	15092	12	496
Bages	4704	903	4519	5924	3758	1738	25562	111	633
Baix Camp	4554	1030	4128	5014	3823	3920	14590	128	564
Baix Ebre	1975	377	1518	2248	1407	4392	7788	47	366
Baix Empordà	2229	721	2658	4193	3999	2722	11844	65	278
Baix Llobregat	15002	3059	18646	17558	16478	2775	86143	224	2118
Baix Penedès	798	228	985	1456	1111	973	4724	24	322
Barcelonès	109438	40542	122925	79951	65488	2466	209762	1499	83063
Berguedà	1098	128	842	1220	778	996	7802	57	399
Cerdanya	324	153	291	478	468	761	1411	50	214
Conca de Barberà	457	137	464	484	388	1304	2896	8	144
Garraf	2326	340	1995	2888	2416	783	9452	16	813
Garrigues	437	94	341	415	300	3011	1760	21	349
Garrotxa	1496	774	1510	1616	960	1321	9940	8	253
Gironès	6128	1364	6015	4815	3615	1387	17713	272	802
Maresme	8747	3122	9632	10882	8190	5001	38721	105	1960
Montsià	921	246	978	1343	862	5089	5142	63	2291
Noguera	908	157	699	1055	604	3848	3826	40	378
Osona	3904	763	3611	4293	2392	3108	24107	19	1465
Pallars Jussà	426	166	338	378	305	903	1552	124	77
Pallars Sobirà	191	68	115	136	178	613	591	9	100
Pla d'Urgell	602	149	494	679	355	2917	3128	32	889
Pla de l'Estany	783	278	793	874	457	737	4014	1	98
Priorat	217	32	163	182	151	1260	919	12	206
Ribera d'Ebre	772	81	459	560	379	1547	2870	18	218
Ripollès	925	212	575	1032	635	803	5536	25	350
Segarra	501	95	506	418	352	1527	2090	11	171
Segrià	6116	1581	5848	5490	3899	9758	15699	470	1707
Selva	2143	769	2959	4289	5673	2177	15148	32	370
Solsonès	305	83	228	246	239	1000	1504	11	107
Tarragonès	6503	1267	6663	4968	5950	1806	16696	216	800
Terra Alta	169	46	131	208	155	2413	1210	15	112
Urgell	783	219	734	1207	544	2341	3721	37	250
Val d'Aran	237	116	189	399	460	137	622	36	77
Vallès Occidental	22082	3824	20615	20135	14845	1217	87170	121	5564
Vallès Oriental	6644	1893	7985	8050	5657	2503	42332	72	1341

REFERÈNCIES

[1] **Amari,S.** (1985). «Differential–Geometrical Methods in Statistics». *Lecture notes on statistics*, **28**. Springer-Verlag, New York.

[2] **Atkinson, C.** and **Mitchell, A. F. S.** (1981). «Rao's Distance Measure». *Sankhyā*, **43, A**, 345–365.

[3] **Bandorff-Nielsen, O.E.** (1984). «Differential geomtry in statistical inference». Capater Differential and integral geometry in statistical inference Volume 10 of *Lecture Notes - Monographs Series* Institute of Mathematical Statistics, Hayward, California.

[4] **Burbea, J.** and **Rao, C. R.** (1982). «Entropy differential metric, distance and divergence measures in probability spaces: a unified approach». *J. Multivarate Anal.*, **12**, 575–596.

[5] **Burbea, J.** and **Rao, C. R.** (1984). «Differential metrics in probability spaces». *Probability Math. Statist.*, **3**, 241–258.

[6] **Calafell, F.** and **Hernández, M.** (1993). «Multivariate approach to matrimonial mobility in Catalonia». *Human Biology*, **65**, 731–742.

[7] **Efron** (1975). «Defining the curvature of a statistical problem (with application to second order efficiency) (with discussion)». *Annals of Statistics*, **3**, 1189–1242.

[8] **Generalitat de Catalunya.** *Anuari Estadístic de Catalunya de 1990.*

[9] **Generalitat de Catalunya.** *Anuari Estadístic de Catalunya de 1992.*

[10] **Greenacre, M.J.** (1984). *Theory and applications of Correspondence Analysis.* Academic Press, London.

[11] **Greenacre, M.J.** (1993). *Correspondence analyisis in practice.* London: Academic Press.

[12] **Oller, J. M.** (1982). *Utilización de métricas riemannianas en análisis de datos multidimensionales y su aplicación a la Biología.* Barcelona: Publicaciones de Bioestadística.

[13] **Oller, J.M.** (1989). «Statistical data analysis and inference». Chapter: *Some geometrical aspects of data analysis and statistics.* Elsevier science publishers B.V. North Holland, Amsterdam.

[14] **Rao, C. R.** (1945). «Information and accuracy attainable in the estimation of parameters». *Bull. Calcurra Math. Soc.*, **37**, 81–91.

[15] **Rao, C.R.** (1995). «A review of canonical coordinates and an alternative to correspondence analysis using hellinger distance». *Qüestiió*, **19**, 23–64.

[16] **Ríos, M., Villarroya, A.** and **Oller, J. M.** (1994). «Intrinsic Data Analysis: a new method for simultaneous representation of populations and variables». *Mathematical Preprint Series*, **160**, Universitat de Barcelona.

[17] **Sokal, R.** and **Sneath, P.H.A.** (1963). *Principles of Numerical Taxonomy*. W. H. Freeman and Co.

ENGLISH SUMMARY

COMBINING GEOMETRIC-DIFFERENTIAL TECHNIQUES WITH CLASSIC MULTIVARIATE ANALYSIS: AN APPLICATION TO CHARACTERIZE CATALAN REGIONS

SERGI VIVES*

ÀNGEL VILLARROYA*

Universitat de Barcelona

In this work the characteristics and relationships, from a social-economic point of view, of the 41 **comarques** *which compose Catalonia are studied. The methodology used is a combination of classical data analysis techniques and a new method for graphical display of statistical models, called IDA (Intrinsic Data Analysis), based on differential geometry tools. This approach has allowed us to classify the* **comarques** *of Catalonia into 9 clusters which, in their turn, can be grouped into four blocks: agricultural, tourist, administrative and industrial.*

Keywords: Intrinsic Data Analysis; distància de Rao; Anàlisi de Correspondències; Anàlisi de conglomerats.

AMS Classification: 62H25, 62H30.

*S. Vives i À. Villarroya. Dept. d'Estadística. Universitat de Barcelona. Av. Diagonal, 645. 08028 Barcelona. Espanya.

–Received january 1995.

–Accepted septembre de 1996.

The use of differential geometry in statistics is a fruitful branch of this science but is hardly used in practical applications, probably due to the theoretical character of published papers about this subject. This paper illustrates how such techniques, alone or together with classical ones, can be used to solve practical problems easily.

The populations of this study are the 41 *comarques* (geographic and administrative divisions) of Catalonia. Originally the criteria used to divide the region were maninly commercial ones, based on proximity to agricultural markets. The main variable of our paper is the random variable X: *Active population in professional groups*, according to the census of 1991, where the groups are:

X_1: Professionals and technicians	X_2: Managing directors
X_3: Civil servants	X_4: Traders and salesmen
X_5: Hotels and others	X_6: Agriculture and fishing
X_7: Industry	X_8: Army

This variable was chosen because, from an anthropological point of view, the way in which human resources are distributed is one of the most important distinctive traits of a population.

To study the relationships between *comarques* the first step was their graphical representation in a plane by means of a new method called intrinsic data anaylis (IDA; Ríos *et al.*, 1994) based on the application of differential geometry tools to statistics. IDA allows the joint representation of populations and variables and can be applied to any statistical model; in our case, according to the nature of variable X, the multinomial model was used. The result is shown in figure 1, which explains 90% of original variability. From this display we can deduce that the most important factor in the differentiation of *comarques* is X_6 (the agricultural component), followed by X_7 (the industrial component). In figures 5a and 5b displays are shown of the same data using correspondence analysis.

The next step consisted of grouping the *comarques* into a reasonable number of meaningful groups. To do it we used the cluster analysis technique known as *average linkage clustering* (also called UPGMA), applied to the interdistance matrix calculated by means of the Bhattacharyya distance, every *comarca* being identified by the value of X. The selection of the Bhattacharyya distance was due to its good properties and to the fact that it coincides with the Rao distance for the multinomial model, the same that IDA uses. UPGMA results are given in figure 2, showing that the dendogram has been cut at level $d = 0.294$ reulting in nine clusters. In table 1 the *comarca* nomenclature and the group composition is shown. In figure 3 the mean values of vector X for each group can be seen.

At this point we proceeded to study the characteristics of the nine clusters. Several variables were studied, but the most informative turned out to be: density, migratory balance, *per capita* family income and age distribution of the active population. The observed mean values for each cluster are displayed in figure 4.

From previous results some conclusions can be made. First of all, when professional groups of the active population are considered, the agrarian component (X_6) is the most important one for the *comarques* differenciation, although the industry is the major component in all the groups (see figures 1 and 3). Another important result is that the 41 *comarques* can be classified into 9 clusters which, in their turn, can be grouped into the following four blocks:

- **Agricultural block**: Composed of clusters 1, 2, 3 (placed on the left side of figure 1) and 4 (central-down placed) characterized by the great importance of X_6, a low migratory balance (positive or negative), a low density and low *per capita* family income, although cluster 4 presents some characteristics similar to those of the other blocks.

- **Touristic block**: Composed of cluster 5, which only includes the *Val d'Aran comarca*, with great importance of service components and with the higher *per capita* family income.

- **Administrative block**: Composed of cluster 6, Barcelones, the *comarca* which includes Barcelona city) with very special characteristics: it includes 38 % of Catalonia's citizens, very high population density, very negative migratory balance and a very small agricultural component.

- **Industrial Block** : Composed of clusters 7, 8 and 9 (upper right side of figure 1) with a very high industrial component and an absence of the agrarian component. Migratory balance is positive, population density high and family income intermediate.

Till this point we have not considered the geographic location of the *comarques*, but when the four previous blocks are placed on a map of Catalonia (figure 6) it is observed that the agricultural block is placed on the west side (with only one exception) and clearly separated from the *comarques* of the industrial block situated on the east side. Service and administrative blocks are situated inside the agricultural and industrial blocks, respectively.

Finally, results corresponding to census of 1991 were compared with those from the census of 1986, the only other census with the same data available (figure 7). The industrial block is the most stable during this period, while the agricultural block changes the most.

Investigació Operativa

QÜESTIIÓ, Vol 20, 3 pp. 485-501, 1996

UNA VARIANTE DEL ALGORITMO DE AHUJA-ORLIN PARA PROBLEMAS DE FLUJO MÁXIMO: EXPERIENCIAS COMPUTACIONALES Y COMPARACIONES

SEDEÑO NODA, A.A.*

GONZÁLEZ MARTÍN, C.*

Universidad de La Laguna

En este trabajo se introduce una variante del algoritmo de escalado de Ahuja y Orlin, con la misma complejidad computacional teórica, para resolver problemas de flujo máximo en redes sin circuitos. Como se constata en las experiencias computacionales que hemos realizado sobre problemas generados aleatoriamente, en el noventa por ciento de los casos el tiempo de CPU del nuevo procedimiento es significativamente inferior.

A variant of Ahuja-Orlin's algorithm for maximum flow problems: comparison and computational experiences

Keywords: Problema de flujo máximo, complejidad computacional, etiqueta distancia, experiencias computacionales.

*Departamento de Estadística, Investigación Operativa y Computación (DEIOC). Universidad de La Laguna. 38271 La Laguna, Tenerife (España).

- Article rebut el setembre de 1995.
- Acceptat el juliol de 1996.

1. INTRODUCCIÓN

El problema de flujo máximo es uno de los problemas más relevantes del análisis de redes y ha sido estudiado exhaustivamente (ver, por ejemplo, Ahuja, Magnanti y Orlin (1989b, 1993), Nicoloso y Simeone (1992), Murty (1992), Nemhauser y Wolsey (1988),...). Fue introducido por Fulkerson y Dantzig (1955) y resuelto por Ford y Fulkerson (1956) mediante su conocido algoritmo de caminos incrementales.

A partir de entonces, se han desarrollado un gran número de procedimientos que, a pesar de la eficiencia práctica del algoritmo anterior, tienden a mejorar la complejidad computacional teórica. Dichos procedimientos aportan nuevas ideas, entre las cuales caben destacarse dos: el concepto de *caminos incrementales mínimos* y el de *preflujo*. Un algoritmo que utiliza ambas ideas es el de Ahuja y Orlin (1989a), incorporando además un procedimiento de escala del exceso de flujo. En este trabajo mostramos que la herramienta etiqueta distancia utilizada en este algoritmo conduce, dependiendo de la morfología de la red, a un comportamiento empírico cercano al estimado para la complejidad computacional del caso peor. Esto implica, en muchos casos, un mayor consumo en el tiempo de ejecución de éste frente a los algoritmos clásicos de Ford y Fulkerson (1956), Dinic (1970), Edmonds y Karp (1972), Malhotra, Kumar y Maheshwari (1978).

Por estas razones, hemos centrado nuestro interés en estudiar algunas modificaciones del procedimiento de Ahuja y Orlin, en la pretensión de conseguir mejoras en su eficiencia práctica. En este trabajo estudiamos una de esas variantes para los casos de redes sin circuitos. Las redes sin circuitos aparecen en multitud de aplicaciones como, por ejemplo, los problemas de flujos en redes dinámicas, problemas de planificación con máquinas paralelas (especializadas o no especializadas), sistemas de abastecimientos de aguas donde la fuerza motriz del bien que circula es de origen natural (gravedad), redes de telecomunicación jerarquizadas, una vez establecido el sentido de la transmisión de datos (origen-destino), etc...

Hacemos énfasis en un análisis comparativo del comportamiento empírico de este nuevo algoritmo frente al procedimiento introducido previamente.

2. FORMALIZACIÓN DEL PROBLEMA Y CONCEPTOS BÁSICOS

Dada una red dirigida conexa y sin circuitos $N = (V,A)$, con $V = \{1,\ldots,n\}$ el conjunto de vértices y $A = \{(i,j)/i \in V', j \in V'', i \neq j, V', V'' \subseteq V\}$ el conjunto de arcos, denotaremos por n el cardinal de V y por m el cardinal de A. Sean $s,t \in V$, dos vértices distinguidos de la red denominados respectivamente, *fuente* y *sumidero*. Para cualquier vértice i definimos $\text{Pred}(i) = \{j \in V/(j,i) \in A\}$, al conjunto de vértices

predecesores de i, y $\text{Suc}(i) = \{j \in V/(i,j) \in A\}$, al conjunto de vértices sucesores de i. Asociado con cada arco $(i,j) \in A$ se tiene una capacidad u_{ij}, que indica el límite superior de la cantidad de flujo que puede soportar el arco y l_{ij} la menor cantidad de flujo que debe pasar necesariamente por el mismo arco. Sin pérdida de generalidad, podemos suponer que l_{ij} es igual a cero para cada arco. Denotaremos por U a $\max\{u_{ij}/(i,j) \in A\}$. En el problema de Flujo Máximo se desea enviar la mayor cantidad de flujo de la fuente al sumidero satisfaciendo las restricciones de capacidad y las de conservación de flujo en los nodos. En lenguaje matemático, este problema se puede plantear en los siguientes términos:

$$\max f$$

$$\text{(1)} \qquad \text{sujeto a:} \quad \sum_{j \in \text{Suc}(i)} x_{ij} - \sum_{j \in \text{Pred}(i)} x_{ji} = \begin{cases} f & \text{si } i = s \\ 0 & i \in V - \{s,t\} \\ -f & \text{si } i = t \end{cases}$$

$$\text{(2)} \qquad 0 \leqslant x_{ij} \leqslant u_{ij}, \quad (i,j) \in A$$

Un flujo factible en N es un vector $x = \{x_{ij}/(i,j) \in A\}$, que satisface (1) y (2), donde x_{ij} denota las unidades de flujo que circulan de i a j sobre el arco (i,j), y f es el valor del flujo de s a t.

Un *preflujo* x es una función $x : A \longrightarrow \mathbb{R}^+$ satisface (2) y la siguiente relajación de (1):

$$e(i) = \sum_{j \in \text{Pred}(i)} x_{ji} - \sum_{j \in \text{Suc}(i)} x_{ij} \geqslant 0, \quad i \in V - \{s,t\}$$

Dado un preflujo x, se define, para cada $i \in V - \{s,t\}$, $e(i)$ como el exceso del nodo i. Un nodo con exceso positivo es llamado nodo *activo*.

Dado cualquier par de nodos $i, j \in V$ definimos la *capacidad residual* del arco (i,j) con respecto a un preflujo x, como $r_{ij} = u_{ij} - x_{ij} + x_{ji}$ (debemos entender que si $(i,j) \neq A$, entonces $u_{ij} = 0$). La capacidad residual de (i,j) representa la cantidad máxima de flujo adicional que puede ser enviada desde el nodo i al j usando los arcos (i,j) y (j,i) de A (si existen ambos). Llamaremos *red residual* o *incremental* a aquella que consiste únicamente de los arcos con capacidades residuales positivas.

Para cada nodo i, definimos $A(i)$ como el conjunto de arcos de la red residual que salen de i.

Una función distancia $d : V \longrightarrow \mathbb{Z}^+$ para un preflujo x es una función del conjunto de nodos a los enteros no negativos. Diremos que d es *válida*, si satisface las dos condiciones siguientes:

a) $d(t) = 0$

b) $d(i) \leqslant d(j) + 1 \quad \forall (i,j) \in A,\ r_{ij} > 0$

Por inducción se demuestra que $d(i)$ es una cota inferior de la longitud del camino más corto de i a t en la red residual, entendiendo por longitud de un camino el número de arcos del mismo. Si para cada i, la etiqueta distancia $d(i)$ es igual a la longitud del camino de longitud mínima entre i y t en la red residual, entonces la denominaremos etiqueta distancia *exacta*.

Un arco (i,j) en la red residual es llamado *admisible* si satisface que $d(i) = d(j) + 1$.

3. ALGORITMO DE AHUJA Y ORLIN

En cada paso, este algoritmo mantiene un preflujo y envía flujo hacia el sumidero, desde los nodos más cercanos con exceso positivo, a través de arcos admisibles. Para estimar qué nodo activo está más cerca del sumidero y qué arcos de los que salen de éste son admisibles, se utiliza la etiqueta distancia. Este procedimiento usa un escalado del exceso. La idea básica consiste en enviar flujo desde nodos activos con exceso suficientemente grande a nodos con excesos suficientemente pequeños, sin que los excesos de estos últimos se hagan demasiado grandes. El algoritmo recibe por ello el nombre de *algoritmo de escala del exceso*. El algoritmo realiza $K = [\log U] + 1$ iteraciones de escalado. Para cada iteración de escalado, se define el *exceso dominante* como el menor entero Δ potencia de dos que satisface que $e(i) \leqslant \Delta,\ \forall i \in \mathbb{N},\ i \neq s,t$. En cada iteración de escalado se consideran nodos con exceso mayor que $\Delta/2$; y entre todos estos, se seleccionan uno de menor etiqueta distancia. Una vez seleccionado, se envia un flujo igual a $\delta = \min\{e(i),\ r_{ij},\ \Delta - e(j)\}$, si el arco (i,j) es admisible y se actualizan convenientemente los excesos de los nodos i y j. Un envío es *saturante* si $\delta = r_{ij}$, es decir, si la cantidad de flujo enviado a través del arco (i,j) coincide con su residuo; en otro caso, es un envío *no saturante*. Una iteración de escalado termina cuando no hay nodos con excesos mayores que $\Delta/2$ y una nueva iteración empieza con Δ igual a $\Delta/2$. Después de $\log U$ iteraciones de escalado todos los nodos tendrán un exceso igual a cero y obtendremos un flujo máximo. Para seleccionar un nodo activo con un exceso mayor que $\Delta/2$ y con la menor etiqueta distancia, se mantienen unas listas o colas $L(r) = \{i \in \mathbb{N} : e(i) > \Delta/2 \text{ y } d(i) = r\}$ para $r = 1, \ldots, 2n-1$. Un esquema del algoritmo es el siguiente:

Algoritmo de Ahuja-Orlin (1989a)
begin
Para cada arco $(s,j) \in A \Rightarrow x_{sj} = u_{sj}$, $x_{ij} = 0$ el resto;
$d(t) := 0$, determina las etiquetas distancias iniciales $d(i)$ mediante un recorrido en amplitud desde el nodo t; $d(s) := n$;
$K = [\log U] + 1$;
for $k = 1$ to K do (bucle logarítmico)
begin
$\Delta := 2^{K-k}$;
for cada nodo do $i \in V$ **do if** $e(i) > \Delta/2$ **then** Suma i a $L(d(i))$;
level:= 1;
while *level* $< 2n$ **do**
begin
if L *(level)* = $\emptyset$ **then** *level*:= *level* + 1
else
begin
Selecciona un nodo $i \in L$ *(level)*;
if hay un arco admisible $(i,j) \in A(i)$ **then**
begin
Envía $\delta = \min\{e(i), r_{ij}, \Delta - e(j)\}$ de i a j;
Actualiza las capacidades residuales y $e(i)$, $e(j)$;
if $e(i) \leqslant \Delta/2$ **then** Borra i de L *(Level)*;
if $(e(j) > \Delta/2)$ **and** $(j <> s,t)$ **then**
Suma j a L *(Level*-1*)*, *Level*:=*Level*-1;
end
else
begin
Borra i de L *(Level)*;
$d(i) = \min\{d(j) + 1 : (i,j) \in A(i)$ y $r_{ij} > 0\}$
Suma i a $L(d(i))$;
end
end
end
end
end;

La Complejidad del algoritmo de Ahuja y Orlin es $O(nm + n^2 \log U)$, donde $n^2 \log U$ es el número de envíos no saturantes y nm es el número de envíos saturantes. El número de veces que se actualizan las etiquetas distancias de los nodos es del orden de $O(n^2)$.

En la experiencia práctica sobre problemas generados aleatoriamente, se constata un comportamiento dispar respecto del tiempo de ejecución de este algoritmo. De-

pendiendo del problema particular, este puede ser resuelto de forma rápida o emplear un número de iteraciones cercana a las estimadas para el peor caso. Esto es así ya que, para el primer caso, el número de actualizaciones de la etiqueta distancia es $O(n)$, mientras que en el segundo es del orden de $O(n^2)$. Conviene recordar que cada vez que se necesita actualizar la etiqueta distancia de un nodo i, es porque este es un nodo activo y se debe disminuir su exceso de flujo. Este exceso de flujo es contribución del envío de flujo a través de arcos que habrán sido saturados o no. El tener que deshacerse de este exceso implica el envío de nuevos flujos a través de arcos que emanen de i, lo que implica un incremento obvio en el número de envíos saturantes y no saturantes. Si la etiqueta distancia de un nodo necesita ser actualizada $O(n)$ veces, es por que se tiene que enviar flujo desde este nodo $O(n)$ veces. Si esto ocurre para $O(n)$ nodos, el algoritmo necesita un número de iteraciones cercana a la de su complejidad teórica. Dicho de otro modo, es fácil que este algoritmo realice $O(nm + n^2 \log U)$ iteraciones incluso cuando la red de partida no sea complicada. Además, incluso habiendo encontrado el valor del flujo máximo, se necesitan realizar un número de iteraciones adicionales para eliminar los excesos de los nodos que no alcanzan a t.

Estimamos que este comportamiento es debido al uso de la herramienta etiqueta distancia, cuya forma de actuar muestra una dependencia clara con la morfología de la red. Dicha herramienta hace imprevisible el comportamiento de estos algoritmos frente a varios problemas con la misma especificación de los parámetros que lo definen, es decir, frente al número de nodos, número de arcos y capacidad máxima. Un análisis detallado de estas consideraciones aparece en González Martín, González Sierra y Sedeño Noda (1995).

A continuación se muestra la tabla 1 en la que se aprecia el comportamiento del algoritmo de Ahuja y Orlin para problemas particulares con la misma especificación.

Tabla 1

Valores de Ret y CPU para el algoritmo de Ahuja y Orlin

n	m	U	RET	CPU	n	m	U	RET	CPU
100	2000	32000	9493	1.17	100	2000	32000	38	0,02
100	2000	1E+09	9357	1.24	100	2000	1E+09	41	0,02
100	2000	2E+09	9372	1.10	100	2000	2E+09	67	0,03
200	7000	32000	38672	7.45	200	7000	32000	60	0,03
200	7000	1E+09	38777	6.82	200	7000	1E+09	68	0,05
200	7000	2E+09	38484	8.90	200	7000	2E+09	51	0,04
300	20000	32000	88122	23.55	300	20000	32000	95	0,05
300	20000	1E+09	88163	25.12	300	20000	1E+09	125	0,12
300	20000	2E+09	87875	26.7	300	20000	2E+09	131	0.12
400	50000	32000	157701	70	400	50000	32000	208	0.22
400	50000	1E+09	157859	71.88	400	50000	1E+09	218	0.03
400	50000	2E+09	157356	73	400	50000	2E+09	266	0.35

El algoritmo ha sido instrumentado en Pascal estándar y ha sido ejecutado sobre una estación de trabajo HP9000 serie 715/33. Los problemas han sido generados aleatoriamente recibiendo como entrada la especificación de parámetros antes reseñada. En esta tabla se muestra el número de nodos (n), de aristas (m) y el mayor valor de la capacidad (U) de la red generada aleatoriamente y, a continuación, el número de actualizaciones de la etiqueta distancia (RET) y el tiempo empleado por el algoritmo en segundos (CPU). Se ve claramente que para la misma especificación, el algoritmo se comporta de manera diferente.

Ahuja y Orlin (1989a) en la presentación de su algoritmo reconocen que, en la práctica, un cuello de botella potencial de éste es el número de actualizaciones de la etiqueta distancia. En particular, el algoritmo identifica que la red residual no contiene ningún camino del nodo i a t sólo cuando $d(i)$ es mayor que $n-2$. Evidentemente un nodo desconectado del sumidero con un exceso positivo, debe enviar este exceso de vuelta a la fuente. Pero para que esto ocurra, la etiqueta distancia de este nodo debe ser incrementada hasta un valor mayor que $n-2$ y, en el peor de los casos, se incrementa en una unidad cada vez. Lo ideal sería que la etiqueta distancia de un nodo activo no fuera actualizada mientras sea posible enviar flujo desde este nodo hacia el sumidero, es decir, mientras t sea alcanzable desde este nodo. Dicho de otro modo, se hace necesario relajar la condición de que para enviar flujo a través de un arco (i,j) las etiquetas distancias sean $d(i) = d(j)+1$ y $r_{ij} > 0$, de tal manera que sólo sea necesario actualizar la etiqueta distancia de un nodo cuando su exceso sea positivo y este nodo no alcance a t. El problema es que para poder relajar esta condición, los valores de las etiquetas distancias deben ser los adecuados. Una posible solución es que para todo nodo conectado al sumidero el valor de $d(i)$ sea el número de arcos del camino más largo de i a t en la red incremental. Así, la condición se cambiaría por $d(i) > d(j)$ y $r_{ij} > 0$ y nos evitaríamos tener que actualizar la etiqueta $d(i)$ hasta que $d(i) = d(j)+1$. La cuestión es que para poder realizarlo, debemos ser capaces de ordenar el conjunto de nodos de acuerdo con la cercanía de estos al sumidero y este orden sólo se consigue de manera total si la red no contiene circuitos.

4. VARIANTE DEL ALGORITMO DE AHUJA Y ORLIN PARA REDES SIN CIRCUITOS

Si la red no contiene circuitos, es posible ordenar el conjunto de nodos de acuerdo con la cercanía de estos a t. La función *orden*, $\text{ord} : N \longrightarrow \mathbb{Z}^+$, es una función definida en el conjunto de nodos y que toma valores en los enteros no negativos tal que $\text{ord}(t) = 0$ y $\text{ord}(i)$ es el número de arcos del camino más largo de i a t. Los valores de esta función pueden calcularse en $O(m)$ mediante un recorrido en profundidad, empezando en el nodo t. Obviamente si N contiene circuitos no es posible tal orden. Además, un $i \in \mathbb{N}$, $i \neq t$ tal que $\text{ord}(i) = 0$ implica que no hay camino de i a t. Proponemos utilizar la función orden en lugar de la función

etiqueta distancia permitiendo enviar flujo desde un nodo i a un nodo j siempre que ord(i) > ord(j). En cada etapa consideraremos el nodo activo i de menor ord(i) y el arco (i,j) de $A(i)$ tal que j sea el de menor etiqueta distancia. De esta manera solo tendremos que actualizar la etiqueta distancia de un nodo activo cuando haya que devolver este exceso hacia la fuente. El algoritmo se desarrolla según el siguiente esquema:

Algoritmo para el caso de N sin circuitos

begin

$d(t) := 0, d(s) := n$; determina el orden ord(i) $\forall i \neq s,t$ mediante un recorrido en profundidad desde el nodo t;

$d(i) =$ ord(i);

Para cada arco $(s,j) \in A : d(j) > 0$ *or* $j = t \Rightarrow x_{sj} = u_{sj}, x_{ij} = 0$ el resto;

$K = [\log U] + 1$;

for $k = 1$ to K **do** (bucle logarítmico)

begin

$\Delta := 2^{K-k}$

for cada nodo $i \in V$ **do if** $e(i) > \Delta/2$ **then** Suma i a $L(d(i))$;

level:= 1;

while *level* $< 2n$ **do**

begin

if L *(level)* = $\emptyset$ **then** *level*:= *level* + 1

else

begin

Selecciona un nodo $i \in L$ *(level)*;

if $\exists j : d(j) = \min_{p \in \text{suc}(i)} \{d(p) : d(i) > d(p)$ y $(d(p) > 0$ o $p = t)\}$

and $r_{ij} > 0$ **then**

begin

Envía $\delta = \min\{e(i), r_{ij}, \Delta - e(j)\}$ **de** i **a** j;

Actualiza las capacidades residuales y $e(i), e(j)$;

if $e(i) \leqslant \Delta/2$ **then Borra** i **de** L ***(Level)***;

if $(e(j) > \Delta/2)$ **and** $(j <> s,t)$ **then Suma** j **a** $L(d(j))$, ***Level***:=$d(j)$;

end

else

begin

Borra i **de** L ***(Level)***;

$d(i) = \min\{d(j) + 1 : (i,j) \in A(i)$ **y** $r_{ij} > 0$ **y** $(d(j) >= d(i))\}$

Suma i **a** $L(d(i))$;

end

end

end

end

end;

Sólo se actualizará la etiqueta distancia de un nodo cuando no haya camino desde este al sumidero. La figura 1 ilustra los pasos del algoritmo. En La figura 1 d) se puede observar que una vez que el nodo 3 no alcanza a t, su valor de etiqueta distancia es mayor que la del nodo fuente, de tal manera que en el siguiente envío desde este nodo, nos podemos deshacer de su exceso. Además, no se enviará flujo de 2 a 3, a través del arco (2, 3), ya que la etiqueta distancia del 2 es menor que la del 3. Esto es así, ya que el orden en el que se envían flujos es siempre a través de los nodos más cercanos a t, y en este caso 3 está muy alejado de t. En el Algoritmo de Ahuja y Orlin se realizan una serie de pasos equivalentes al de la figura anterior, pero efectuando actualizaciones de la etiqueta distancia adicionales y, por lo tanto, un número mayor de envíos de flujo.

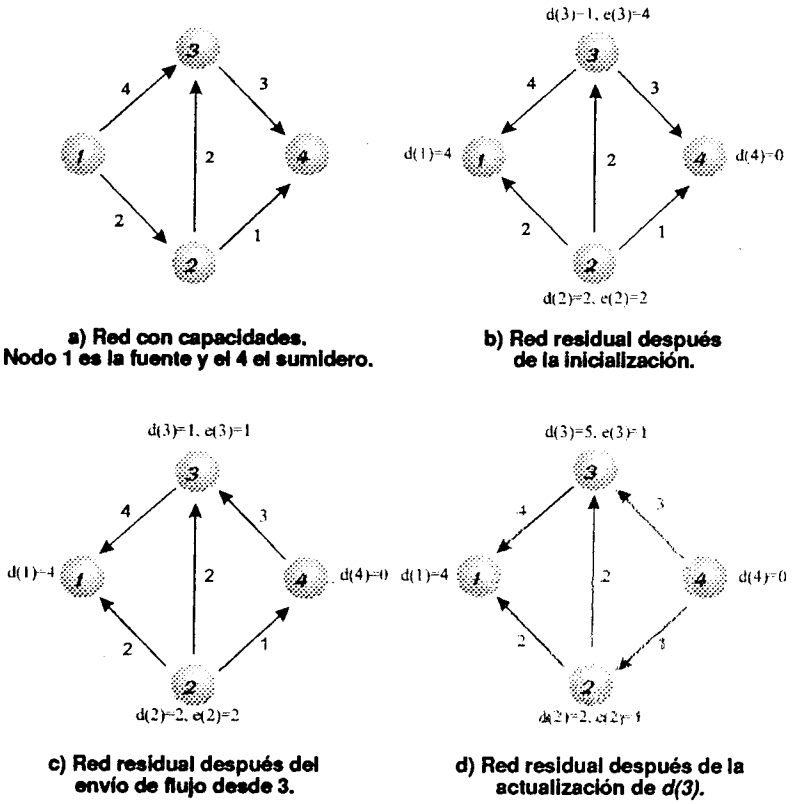


Figura 1. *Pasos del algoritmo con la nueva etiqueta distancia*

5. ASPECTOS COMPUTACIONALES

La complejidad de este algoritmo es la misma que el de Ahuja y Orlin, es decir, $O(nm+n^2 \log U)$. Esto es evidente, ya que si bien la etiqueta distancia $d(i)$ representa la longitud del camino más largo de i a t, el valor de ésta vuelve a estar acotado por $2n$ y cada envío no saturante desde un nodo i a un j es al menos de $\Delta/2$ unidades. Por lo tanto el número de envíos no saturantes vuelve a estar acotado por $8n^2$ (ver Ahuja y Orlin (1989a)). Por otro lado, la condición de enviar flujo de un nodo i a un nodo j es $d(i) > d(j)$, lo cual reduce en la práctica el número de actualizaciones de las etiquetas distancia. Conviene recordar que cada vez que se actualizaba la etiqueta distancia de algún nodo en el algoritmo de Ahuja y Orlin era porque los arcos que salían de este nodo habían sido saturados. Además, por ser este nodo activo, necesita realizar nuevos envíos para deshacerse de sus excesos. Por ello, disminuir el número de actualizaciones de d equivale a disminuir el número de envíos, tanto saturantes como no saturantes. Estos comentarios se pueden constatar en la práctica.

Más adelante se muestran los resultados obtenidos a la hora de resolver una serie de problemas generados aleatoriamente. Los algoritmos han sido instrumentados en Pascal estándar y han sido ejecutados sobre una HP9000 serie 715/33. Como medida del esfuerzo de ejecución de los dos algoritmos (A&O= Ahuja y Orlin y VA&O= variante de Ahuja y Orlin) hemos tomado los siguientes valores: el tiempo de CPU(seg), el número de envíos saturantes (SAT), el número de envíos no saturantes (NSAT) y el número de actualizaciones de las etiquetas distancia (RET).

El estudio práctico se ha llevado a cabo generando aleatoriamente 150 problemas con las siguiente especificaciones para los nodos n, las aristas m y la capacidad máxima U. El rango de los nodos se estableció en 400, 600 y 800. Para cada valor del rango de nodos, el rango de aristas se estableció en $5n, 10n, 15n, 20n, 25n, 30n, 35n, 40n, 45n, 50n$. Para cada valor de n y m, se generaron 5 problemas cuyas capacidades fueron tomadas de forma aleatoria y uniforme en el intervalo $[1, U]$, con U tomando los valores $10^5, 10^6, 10^7, 10^8, 10^9$. De los 150 problemas, en 136 el método denominado VA&O mejora notablemente el tiempo de ejecución necesario para resolver un problema, ya que disminuye el número de envíos no saturantes, saturantes y el número de actualizaciones de la etiqueta distancia. En los 14 restantes ambos métodos son similares.

En la tabla 2, se muestran los tiempos de CPU en segundos, así como los valores de NSAT, SAT y RET promediados según los rangos de variación de las aristas y las capacidades máximas. Se observa que tanto los tiempos de CPU, como el número de envíos no saturantes, saturantes y el número de actualizaciones de la etiqueta distancia es bastante inferior en el método VA&O que en el A&O.

En la tabla 3, se muestran los tiempos de CPU en segundos, así como los valores de NSAT, SAT y RET promediados según los rangos de variación de la relación aristas/nodos,obteniendo así, el comportamiento de los métodos según la densidad de

la red. En esta tabla se vuelven a observar que los resultados obtenidos por el método VA&O mejoran los del A&O.

Tabla 2

Promedio para aristas y capacidad máxima

n	Algoritmo	CPU	NSAT	SAT	RET
400	A&O	19.55	86207	41427	107711
400	VA&O	4.95	19938	9560	24138
600	A&O	44.13	196744	78267	233445
600	VA&O	16.03	57996	26329	69495
800	A&O	89.54	397405	142229	454264
800	VA&O	27.50	103464	42069	118071

Tabla 3

Promedio según la densidad de la red

m/n	Algoritmo	CPU	NSAT	SAT	RET
5	A&O	30,49	253345	71685	247749
5	VA&O	7,56	61443	19184	56463
10	A&O	28,59	207936	63009	219200
10	VA&O	8,38	61496	21430	61962
15	A&O	29,58	186855	66106	206936
15	VA&O	10,56	64631	24578	69001
20	A&O	51,69	267268	95176	303507
20	VA&O	14,12	65010	25137	72555
25	A&O	52,83	245555	98958	290648
25	VA&O	16,53	62170	27285	73990
30	A&O	57,56	243672	94328	287650
30	VA&O	17,87	66006	31665	81735
35	A&O	50,69	196325	76843	240288
35	VA&O	16,28	57101	27725	71494
40	A&O	67,72	232683	102357	299528
40	VA&O	23,40	58414	30867	76662
45	A&O	66,71	191643	98229	249777
45	VA&O	20,81	49720	22178	63242
50	A&O	66,71	191643	98229	249777
50	VA&O	26,09	58667	29808	78574

En las figuras anteriores aparecen unas gráficas del tiempo de CPU, en segundos, empleado por cada algoritmo para cada problema en particular, agrupando según el número de nodos. La figura 2 se refiere a los problemas con 400 nodos, la figura 3 a los de 600 nodos y la figura 4 a los de 800 nodos. En estas figuras, se puede observar que el VA&O emplea menor tiempo de CPU que el A&O en la mayoría de los problemas. En aquellos, en que A&O es más rápido que VA&O, se aprecia que el tiempo de CPU es similar.

6. CONCLUSIONES

El método de Ahuja y Orlin y, en general, aquellos que utilizan la función etiqueta distancia son sensibles a la morfología de la red y, por lo tanto, a la numeración de los nodos. Esto hace que el comportamiento práctico de los mismos precise de un número de iteraciones similar a las de su cota del caso peor, aún cuando las redes no sean complicadas. Esto se pone de manifiesto en la tabla 1 y es tratado de manera rigurosa en el estudio estadístico contenido en González Martín, González Sierra y Sedeño Noda (1995).

El uso de unas etiquetas distancias adecuadas, o dicho de otro modo, el orden de consideración adecuado de los nodos de una red, proporciona una mayor robustez práctica a los métodos mencionados anteriormente. El orden o etiquetado presentado en este trabajo, válido para redes sin circuitos, mejora notablemente el comportamiento empírico de unos de estos algoritmos.

7. BIBLIOGRAFÍA

[1] **Ahuja, R.** and **Orlin, J.B.** (1989a). «A fast and simple algorithm for the maximum flow problem». *Operations Research*, **37**, 748–759.

[2] **Ahuja, R., Magnanti, T.** and **Orlin, J.B.** (1989b).«Network flows». En *Optimizacion. Handbooks in Operations Research and Management Science*. **Vol. 1**, 211–369. G.L. Nemhauser, A.H.G. Rinnooy Kan, M.J. Todd (eds.). North Holland.

[3] **Ahuja, R., Magnanti, T.** and **Orlin, J.B.** (1993). *Network Flows*. Prentice-Hall, Inc.

[4] **Dinic, E.A.** (1970). «Algorithms for solution of aproblem of maximum flow in networks with power estimation». *Soviet Mathematical Doklady*, **11**, 1277–1280.

[5] **Edmonds, J.** and **Karp, R.M.** (1972). «Theoretical improvements in algorithmic efficiency of network flow problems». *Journal of ACM*, **19**, 248–264.

[6] **Ford, L.R.** and **Fulkerson, D.R.** (1956). «Maximal flow through a network». *Canadian Journal of Mathematics*, **8**, 399–404.

[7] **Goldberg, A.V.** and **Tarjan, R.E.** (1986). «A new approach to the maximum flow problem». *Proc. 18th ACM Symp. on the Theory of Computation*, 136–146.

[8] **González Martín, C., González Sierra, M.A.** and **Sedeño Noda, A.A.** (1995). *An algorithimic study of the maximum flow problem: a comparative statistical analysis*. Documento de trabajo.

[9] **Karzanov, A.V.** (1974). «Determining the maximal flow in a network by the method of preflows». *Soviet Mathematical Doklady*, **15**, 434–437.

[10] **Malhotra, V.M., Kumar, M.P.** and **Maheshwari, S.N.** (1978). «An $O(|V|^3)$ algorithm for finding maximum flows in networks». *Inform. Process. Lett.*, **7**, 277–278.

[11] **Murty, K.** (1992). *Network programming*. Prentice-Hall, Inc.

[12] **Nemhauser, G.** and **Wolsey, L.** (1988). *Integer and combinatorial optimization*. John Wiley and Sons, Inc.

[13] **Nicoloso, S.** and **Simeone, B.** (1992). *Classical and contemporary methods in network optimizacion*, part I: Network Flows.

ENGLISH SUMMARY

A VARIANT OF AHUJA-ORLIN'S ALGORITHM FOR MAXIMUM FLOW PROBLEMS: COMPARISON AND COMPUTATIONAL EXPERIENCES

SEDEÑO NODA, A.A.*

GONZÁLEZ MARTÍN, C.*

Universidad de La Laguna

In this paper, it is introduced a variant of Ahuja and Orlin's scaled algorithm, with an equivalent theoretical complexity, for solving maximum flow problems on networks without circuits. As it can be deduced from computational experiences made with randomly generated problems, in the ninety per cent of the cases, the CPU time of this new procedure is significantly less.

Keywords: Maximum flow problem; computational complexity; distance label; computational experiments.

*Departamento de Estadística, Investigación Operativa y Computación (DEIOC). Universidad de La Laguna. 38271. La Laguna, Tenerife (España).

-Received september 1995.

-Accepted july 1996.

In this paper, it is introduced a variant of Ahuja and Orlin's scaled algorithm, with an equivalent theoretical complexity, for solving maximum flow problems on networks without circuits. As it can be deduced from computational experiences made with randomly generated problems, in the ninety per cent of the cases, the CPU time of this new procedure is significantly less.

The Ahuja and Orlin's algorithm uses the *shortest augmenting path* idea presented by Edmonds and Karp, and the *Preflow* idea, introduced by Karzanov for solving the maximum flow problem. This algorithm keeps a preflow in each iteration and sends flow to the sink node from the nearest nodes with positive excess through admissible arcs. The distance label is used for estimating which active node is nearer the sink and which adjacent arcs of this one are admissible. This algorithm uses and excess scaled. The basic idea is to send flow from active nodes with excesses enough little, avoiding to create with big excess again. This algorithm is named excess scaled algorithm for this reason. The complexity of Ahuja and Orlin's algorithm is $O(nm+n^2 \log U)$, where $n^2 \log U$ is the number of non saturating sends and nm is the number of saturating sedns. The number of times that the distance labels are updated is $O(n^2)$ (1989a).

We have proved experimentaly that the behavior of this algorithm depends on the network morphology. Sometimes it needs much more time for solving a problem that others where the timeu sed is lower. Then, the algorithm makes a number of iterations near the estimated for the worst case according to the kind of network. We think that this behavior is caused for the use of the distance label tool, which depends clearly on the netsork morphology. This tool makes unforable the behavior of this algorithm with different problems the same specification: number of nodes, number of arcs and maximum capacity. The results are shown in González Martín, González Sierra and Sedeño Noda (1995).

Ahuja and Orlin (1989a), in the presentation of their algorithm, recognize that, in practice, a potential bottle neck is the number of updates of the label distance. Particularly, the algorithm detects that there is not path from node i to node t on the residual network only when the label distance of i is greater than $n-2$. Evidently, a disconnected node from the sink with a positive excess, musts send back it to the source. However, until this happens, the distance label of this node must be incremented until a value greater than $n-2$ and, in the worst case, it is increased in one unit each time. The best way would be that the distance label of an active node keeps its value while it was possible to send flow from this node to the sink; while t was reachable from this node. A solution is that, for each node i connected to the sink, the value of the distance label was the number of arcs of the longest path from i to t on the incremental network. Then, the condition of admissibility would be relaxed for decreasing the number of updates of the distance label. For realizing this, we must to order the set of nodes according to their nearness to the sink, and this order only is possible if the network has not circuits.

In this work we present a modification of the excess scaled algorithm, using a new distance function named *Order*, for networks without circuits. This one leads to an algorithm with the same complexity of Ahuja and Orlin's algorithm but with a practical behavior much better. We introduce also, an experimental comparison of the algorithms on randomly generated problems.

Estadística Oficial:

Mètodes per a la preservació del secret estadístic

QÜESTIIÓ, Vol 20, 3 pp. 505-521, 1996

PRIVACY HOMOMORPHISMS FOR STATISTICAL CONFIDENTIALITY†

JOSEP DOMINGO I FERRER*

Universitat Rovira i Virgili

When publishing contingency tables which contain official statistics, a need to preserve statistical confidentiality arises. Statistical disclosure of individual units must be prevented. There is a wide choice of techniques to achieve this anonymization: cell suppression, cell perturbation, etc. In this paper, we tackle the problem of using anonymized data to compute exact statistics; our approach is based on privacy homomorphisms, which are encryption transformations such that the decryption of a function of ciphers is a (possibly different) function of the corresponding clear messages. A new privacy homomorphism is presented and combined with some anonymization techniques, in order for a classified level to retrieve exact statistics from statistics computed on disclosure-protected data at an unclassified level.

Keywords: Statistical confidentiality, privacy homomorphisms.

†This work is partly supported by the Spanish CICYT under grant no. TIC95-0903-C02-02.

*Estadística i Investigació Operativa, Dep. Enginyeria Química, Universitat Rovira i Virgili, Autovia de Salou s/n, 43006 Tarragona, **e-mail:** jdomingo@etse.urv.es

–Article rebut el juny de 1995.

–Acceptat l'abril de 1996.

1. INTRODUCTION

Statistical confidentiality attempts to keep individual information anonymous when releasing macrodata (contingency tables) and microdata (individual records). Such data can be released by publishing printed reports or following a set of queries to a statistical database. In the first case, off-line protection is enough, whereas in the second case on-line protection is required. While off-line statistical confidentiality techniques seem well developed, on-line techniques do not appear as very promising (see [Adam (1989)] for a survey) even if they have been long studied: as queries are done interactively, one can very often devise an adaptive query strategy to get a particular information out of the database. Another taxonomy of disclosure control methods [Schackis (1993)] is based on whether the data to be protected are macrodata or microdata; for macrodata, available methods are cell suppression, change of the classification scheme, rounding and random perturbation; for microdata, the choice is between data reduction and data modification methods. Rather than following one of the above mentioned references to extensively review all available methods, we prefer to recall here three operating principles underlying, if not all methods, at least a good deal of them:

① ***Random perturbations***. The basic idea is to distort microdata/macrodata by adding a small (often zero-mean) perturbation. This technique suffices for anonymizing off-line contingency tables, and can be improved by performing secondary compensations in order to keep marginal sums unchanged [Appel (1992)], [Turmo (1993)]. For databases which can be queried repeatedly, zero-mean random perturbations are far from secure: the true answer can be derived by computing the average of a sufficiently large number of perturbed answers to the same query. Error inoculation is a possible solution: the perturbation used for a value is not computed each time the record is used for a computation, but is determined by a fixed perturbation factor stored along with the value.

② ***Data suppression or query set size control***. Statistics affecting more than K individual records o less than $N-K$ are not supplied, where N is the total number of records in the database. This strategy is useful for anonymizing off-line contingency tables, where it amounts to eliminating cells with very low or high frequencies (disclosure cells). In this case, secondary suppressions will probably have to be performed; otherwise, marginal sums could be used to determine primary suppressions. It is desirable to minimize the total value of suppressed cells, which can be achieved through the use of linear programming [Cox (1992)]. For on-line statistical databases, data suppression amounts to controlling the query set size (queries having very small or very large query sets are not answered). Unfortunately, this technique is rather ineffective for on-line protection: Schlörer showed long ago that an individual record can be isolated by an iterative technique such as the «tracker» [Schlörer (1979)], even for K near $N/2$.

③ ***Random samples***. In this approach, statistics are not computed on the original data, but on a random sample of them —a random query subset in a database—. This is a rather secure technique, but can only provide estimates of marginal sums. If samples are small, then there is a clear loss of quality. A related alternative is using a re-sampling method, such as bootstrapping (see [Heer (1992)] and subsection 4.3).

Disclosure control methods yield data only usable for consultation and approximate computation at an unclassified level. In this paper, we will show a cryptographic way for a classified level (statistical institute) to take advantage of unclassified (*e. g.* subcontracted) computation on disclosure-protected data. The idea is to extract, with little classified effort, exact statistics from approximate statistics computed on disclosure-protected data at an unclassified level. Future research will be directed to exporting our results to on-line scenarios. In section 2, we introduce privacy homomorphisms (PH), which are the basic tool used in this paper. In section 3, we present a new PH which has better properties than the PHs known so far. In section 4, the use of PHs for multilevel computation on microdata and macrodata is discussed; more specifically, we illustrate the combination of PHs with disclosure control methods based on random perturbation, data suppression and resampling. Section 5 contains some final remarks.

2. PRIVACY HOMOMORPHISMS

Privacy homomorphisms (PHs from now on) were formally introduced in [Rivest (1978b)] as a tool for processing encrypted data. Basically, they are encryption functions $E_k : T \longrightarrow T'$ which allow to perform a set F' of operations on encrypted data (in T') without knowledge of the decryption function D_k. Knowledge of D_k allows to recover the same outcome that would be obtained if the corresponding set F of operations had been used on clear data (in T). The security gain is obvious because classified data can be encrypted, processed by an unclassified computing facility, and the result decrypted by the classified level. Next, we include some simple examples of PHs

Example 1. An exponential cipher such as RSA [Rivest (1978a)] is a PH. Let $m = pq$, where p and q are two large secret primes (about 100 decimal digits each). In this case,

$$T = T' = Z/(m)$$

$$E_k(a) = a^e \bmod m$$

$$D_k(a') = (a')^d \bmod m$$

where $Z/(m)$ is the set of integers modulo m, d is secret and $ed \bmod \phi(m) = 1$, with $\phi(m) = (p-1)(q-1)$ being Euler's totient function. Clearly,

$$D_k(E_k(a)) = a^{ed} \bmod m = a^{1+t\phi(m)} \bmod m = a$$

where Euler's theorem is used in the last step. Now, let $F = F' = \{\star\}$, where $\star$ denotes the modular multiplication over $Z/(m)$. The following homomorphic property holds

$$E_k(a) \star E_k(b) = (a^e \bmod m)(b^e \bmod m) \bmod m = (a \star b)^e \bmod m = E_k(a \star b)$$

This homomorphism allows only one operation, but appears to be very secure. Finding D_k from E_k, *i. e.* finding d from e, seems to be equivalent to factoring a large modulus m —no polynomial-time algorithm for factoring has been published up-to-date—. An additional interesting property relates to the preservation of the equality predicate, because it holds that

$$E_k(a) = E_k(b) \text{ if and only if } a = b$$

□

Example 2. In [Rivest (1978b)], the following PH is given. Let p and q be two large secret primes (100 decimal digits each). Consider the set of cleartext data $T = Z/(m)$ and the set of cleartext operations $F = \{+_m, -_m, \times_m\}$ consisting, respectively, of the addition, subtraction and multiplication modulo m, with $m = pq$. Define the encryption key $k = (p,q)$ and $E_k(a) = [a \bmod p, a \bmod q]$. Now, given k, and given $[b,c] \in Z/(p) \times Z/(q)$, computation of $D_k([b,c])$ is as follows: decryption consists of finding an $x \in Z/(m)$ such that

$$b = x \bmod p \quad \text{and} \quad c = x \bmod q \tag{1}$$

To compute x, we have from the first equation 1 that

$$x = b + pt$$

for some t. Then, substitution in the second equation 1 yields

$$b + pt \bmod q = c$$

Now, if p^{-1} is the multiplicative inverse of p modulo q —Euclid's extended algorithm can be used to invert p over $Z/(q)$—, we have

$$t = p^{-1}(c-b) \bmod q = d + qr$$

for some r. Finally

$$x = b + pd + (pq)r = b + pd \bmod m$$

The Chinese remainder theorem guarantees the uniqueness of x.

Thus, the set of ciphertext data is $T' = Z/(p) \times Z/(q)$, and the set F' of ciphertext operations consists of the componentwise modular addition, subtraction and multiplication. Since $k = (p,q)$ is only known at a classified level, unclassified operations on encrypted data are actually carried out over $Z \times Z$ (reduction over $Z/(p) \times Z/(q)$ being restricted to the classified decryptor). This fact implies that, for a given number in $Z/(m)$, there is an infinity of different enciphered versions; therefore, test for equality is not possible at an unclassified level. Also, the authors of [Rivest (1978b)] do not consider division as a valid operation. Whereas there is no problem in including the componentwise modular division in F', this operation cannot be formally included in F, because $Z/(m)$ is not a field. However, the only additional requirement for division to be possible is that the divisor be relatively prime to m. This is not a very restrictive condition for statistical data processing, because there are $\phi(m) = pq - p - q + 1$ numbers relatively prime to m in $Z/(m)$; if p and q are 100-digit primes, then $\phi(m)$ and m have about the same order of magnitude. The probability of getting a divisor not relatively prime to m is

$$\frac{m - \phi(m)}{m} = \frac{p + q - 1}{pq} \approx O(10^{-100})$$

So, dividing at the unclassified level involves finding the multiplicative inverse of the divisor over $Z/(m) \times Z/(m)$ (almost always possible), and then multiplying. The resulting quotient can be reduced by the classified level into $Z/(p) \times Z/(q)$, and then decrypted to get a quotient in $Z/(m)$. However, *this quotient coincides with the standard quotient over Z only if the division is exact* (with remainder 0). Thus it is better to leave divisions in rational form (as fractions).

While this homomorphism allows more operations than the one in example 1, it is less secure. Brickell and Yacobi [Brickell (1988)] have shown that it can be broken by a known cleartext attack. If the ciphertexts $[b_i, c_i]$ corresponding to some cleartexts a_i, for $1 \leq i \leq r$, are known (the ciphertexts being nontrivial, *i. e.* $b_i \neq a_i$ or $c_i \neq a_i$), then p and q can be found with a high probability, and any ciphertext decrypted. The idea is that if $p' = \gcd\{a_i - b_i : 1 \leq i \leq r\}$ and $q' = \gcd\{a_i - c_i : 1 \leq i \leq r\}$, then $p|p'$ and $q|q'$. There is a high probability that $p = p'$ and $q = q'$, even for small r; anyway, as r increases, so does the probability of finding p and q. Still, this homomorphism can safely be used as long as no corresponding nontrivial ciphertext and cleartext are released.

□

3. A NEW PRIVACY HOMOMORPHISM

We propose in this section a new privacy homomorphism which is similar to the one of example 2 (it has the same sets T, T', F and F'), but entails two significant improvements

- Small values are nontrivially encrypted.
- The new PH is able to withstand a general known-cleartext attack, and specifically the [Brickell (1988)] attack.

3.1. The basic idea

When p, q, $m = pq$ are very large integers, a small value a is very likely to have the same representation over $Z/(m)$, $Z/(p)$ and $Z/(q)$, that is

$$a \bmod m = a \bmod p = a \bmod q \quad \text{if } a < \min(p,q)$$

This is an undesirable feature, because the homomorphism of example 2 leaves the cleartext unencrypted (trivial ciphertext). A possible solution is to multiply a by a pair of secret constants r_p and r_q such that $r_p < p$ and $r_q < q$ (the encryption key is now extended to $k = (p,q,r_p,r_q)$). A further improvement to deter ciphertext-only attacks based on frequency analysis is to secretly and randomly split a into $a_{.1},\cdots,a_{.n}$, such that $a_{.j} \in Z/(m)$ and $\sum_{j=1}^{n} a_{.j} \bmod m = a$. Thus the privacy homomorphism we propose is

Public parameters n, m (actually m can be made secret, to increase security).

Secret key p, q large primes, such that $pq = m$. Also, $r_p \in Z/(p)$, such that it generates a large multiplicative subgroup in $Z/(p) - \{0\}$. Also, r_q with similar properties with respect to $Z/(q)$.

Encryption Randomly split $a \in Z/(m)$ into secret $a_{.1},\cdots,a_{.n}$ such that

$$a = \sum_{j=1}^{n} a_{.j} \bmod m \quad \text{and} \quad a_{.j} \in Z/(m).$$

Compute

$$E_k(a) = ([a_{.1}r_p \bmod p, a_{.1}r_q \bmod q], [a_{.2}r_p^2 \bmod p, a_{.2}r_q^2 \bmod q], \cdots \quad (2)$$
$$\cdots, [a_{.n}r_p^n \bmod p, a_{.n}r_q^n \bmod q])$$

Decryption Compute the scalar product of the j-th [mod p, mod q] pair by $[r_p^{-j} \bmod p, r_q^{-j} \bmod q]$ to retrieve the $[a_{.j} \bmod p, a_{.j} \bmod q]$. Add up to get $[a \bmod p, a \bmod q]$. Use the Chinese remainder theorem to get $a \bmod m$.

3.2. The operations on encrypted data

As operations on encrypted values are carried out over $Z \times Z$ by the unclassified level, the use of r_p and r_q requires that the terms of the encrypted value having different r-degree be handled separately —the r-degree of a mod p, resp. mod q, term is the exponent of the power of r_p, resp. r_q, contained in the term—. This is necessary for the classified level to be able to multiply each term by r_p^{-1} (inverse of r_p over $Z/(p)$) and r_q^{-1} (inverse of r_q over $Z/(q)$) the right number of times, before adding all terms up, reducing the final result into $Z/(p) \times Z/(q)$, and decrypting into $Z/(m)$.

The only operation altering the r-degree is multiplication. If cleartext data x and y have been encrypted as $E_k(x)$ and $E_k(y)$, with r-degrees n_1 and n_2, then the product $E_k(x)E_k(y)$ has r-degree $n = n_1 + n_2$. The result may have terms $E_{k,j}(z)$ with degrees ranging from $j = 1$ to n and can be represented in vector notation as

$$(E_{k,1}(x), E_{k,2}(x), \cdots, E_{k,n}(x))$$

If we set $r = [r_p, r_q]$ then $E_k(x)$ can also be written as a polynomial of r

$$E_k(x)[r] = t_1 r + \cdots + t_n r^n$$

Although the coefficients t_j are in practice unknown to the unclassified level, the polynomial notation is useful to understand how algebraic operations should be carried out by this level using terms rather than coefficients

Addition and subtraction In vector notation, they are done componentwise over Z, which in polynomial notation means adding terms with the same degree.

Multiplication It works like in the case of polynomials: all terms are cross-multiplied in Z, with an j_1-th degree term by a j_2-th degree term yielding a $j_1 + j_2$-th degree term; finally, terms having the same degree are added up.

Division Cannot be carried out in general because the polynomials are a ring, but not a field. A good solution is to leave divisions in rational format by considering the field of rational functions, *i. e.* fractions whose numerator and denominator are polynomials. In this way, if a and b are two integers, we encrypt a/b as $\frac{E_k(a)}{E_k(b)}$.

Note. When addition or subtraction are performed on fractions with different denominators, numerators cannot be added or subtracted directly. The rules for ordinary fractions should be followed

$$\frac{E_k(a)}{E_k(b)} \pm \frac{E_k(c)}{E_k(d)} = \frac{E_k(a)E_k(d) \pm E_k(b)E_k(c)}{E_k(b)E_k(d)}$$

3.3. Security

We shall next quote two security results related to the proposed PH; for a detailed security analysis of the new PH from the cryptographic point of view, refer to [Domingo (1996)]. The first result is that our proposal appears to withstand any known-cleartext attack. In other words, even if an enemy has access to some *random* cleartexts and corresponding ciphertexts, she is not able to retrieve the secret key $k = (p, q, r_p, r_q)$ that would allow her to decipher arbitrary ciphertext data. A second result relates to plaintext splitting, which turns out to be essential to the security of the PH. This means that one should take $n \geq 2$. Without plaintext splitting ($n = 1$), the PH is shown to be insecure and can be broken by a known-plaintext attack, specifically an extended Brickell-Yacobi gcd attack.

4. MULTILEVEL COMPUTATION USING PRIVACY HOMOMORPHISMS

PHs enable multilevel processing of sensitive data. The idea is to have just a «small core» of resources for classified tasks at a statistical office. PHs make possible to subcontract external computing facilities without compromising statistical secrecy —external service providers being special instances of unclassified levels.

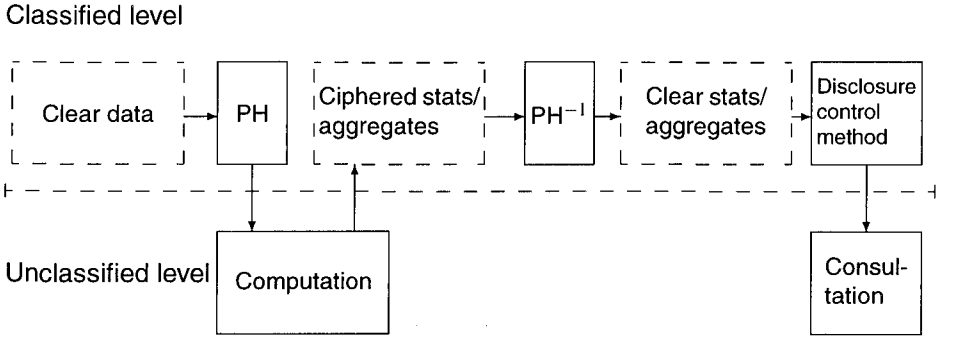


Figure 1.

Multilevel computation (scenario A).

At least two scenarios for using PHs are conceivable. In scenario A (see 1), sensitive micro/macrodata can be encrypted under a PH at a classified level and then be forwarded to an unclassified facility for computation of encrypted statistics or aggregate data; such results can later be decrypted at the classified level. Scenario A is quite straightforward and needs no further discussion. On the other hand, scenario B makes use of PHs in more subtle ways (see figure 2). After using a disclosure con-

trol method to protect clear exact micro/macrodata, some information generated by the method can be encrypted and forwarded to the unclassified facility. This facility takes such encrypted information and the disclosure-protected data as inputs to its computation. The outputs are forwarded to the classified level: this level *alone* can extract (with little effort) exact results from the outputs received from the unclassified level. Scenario B is the most relevant to this paper and will be developed in the following subsections for disclosure control methods based on random perturbation, data suppression and resampling.

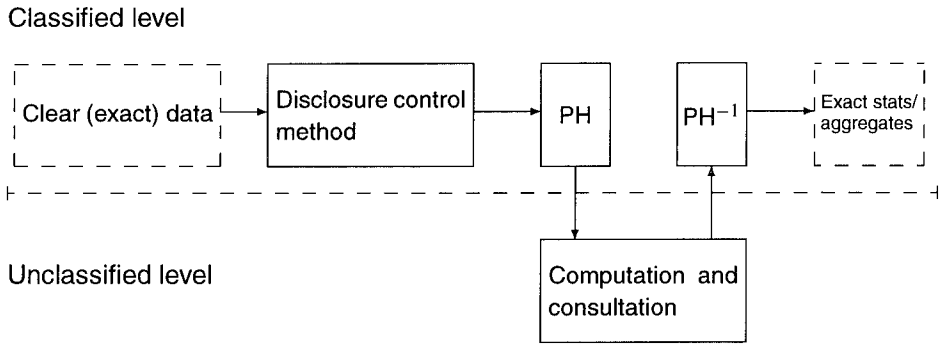


Figure 2.

Multilevel computation (scenario B).

4.1. Multilevel processing of randomly perturbed data

Assume that, at a classified level, a statistical office uses a random perturbation method on data $x_1, x_2, \cdots, x_n$ (which may be microdata or frequencies in a contingency table). This gives

$$x_i^* = x_i + \varepsilon_i \quad \forall i = 1, \cdots, n$$

where ε_i is a random value. Let E_k be the encrypting transformation of a PH with a set F of cleartext operations and a corresponding set of F' of ciphertext operations. Now, the classified level releases the pairs $(x_i^*, E_k(-\varepsilon_i))$ to the unclassified level for further processing. This level is able to perform on the x_i^* the operations in F and compute the encrypted perturbation of the result by using the operations in F' on the $E_k(-\varepsilon_i)$.

At any time, the classified level can take advantage of computations performed by the unclassified level on perturbed data, since restoration of the exact result involves

only decrypting the perturbation of the result. Remark that the unclassified level must content itself with perturbed statistics (on perturbed data), unless the classified level *decides* to reveal the clear perturbation corresponding to some perturbed statistic.

The homomorphism of section 3 can be used to have the unclassified level compute the encrypted perturbations corresponding to any kind of perturbed statistics. It is useful here to consider non-integer perturbations and non-integer data. Remark that integer perturbations are often too coarse, since in most applications perturbed statistics are required to reflect original statistics to some extent. So, if perturbations have at most u decimal positions, we multiply them by 10^u. In this case, a perturbation ε_i is encrypted as $\frac{E_k(\varepsilon_i \times 10^u)}{10^u}$; similarly, a perturbed value x_i^* is represented as $\frac{x_i^* \times 10^u}{10^u}$. Remark that, being public, the denominator 10^u need not be encrypted. Combining cleartext values with encrypted values in computations poses no arithmetical problems, as will be shown below. In this way, without loss of generality, we can assume in what follows that perturbations ε_i and perturbed values x_i^* have been converted to integers (the numerators of their corresponding fractions).

4.1.1. Encrypted perturbation algebra

Table 1 summarizes the computations on encrypted perturbations generated by elementary operations. Given that $x = x^* - \varepsilon_x$ and $y = y^* - \varepsilon_y$, deriving the clear perturbations for addition, subtraction and multiplication is straightforward. Division is not explicitly considered, because it follows from subsection 3.2 that it can be avoided if rational numbers are represented and handled as fractions.

Table 1

Encrypted perturbations corresponding to elementary operations

Perturbed operation	Clear perturbation	Encrypted perturbation
$x^* + y^*$	$-(\varepsilon_x + \varepsilon_y)$	$E_k(-\varepsilon_x) + E_k(-\varepsilon_y)$
$x^* - y^*$	$-(\varepsilon_x - \varepsilon_y)$	$E_k(-\varepsilon_x) - E_k(-\varepsilon_y)$
$x^* y^*$	$-x^*\varepsilon_y - y^*\varepsilon_x + \varepsilon_x\varepsilon_y$	$x^* E_k(-\varepsilon_y) + y^* E_k(-\varepsilon_x) + E_k(-\varepsilon_x)E_k(-\varepsilon_y)$

The new homomorphism has two remarkable properties which justify the formulae for encrypted perturbations in table 1.

- The fact that it is based on finite fields greatly facilitates computation at the unclassified level, which can operate over $\mathcal{Z} \times \mathcal{Z}$. Then the classified level, which knows

the underlying $Z/(p) \times Z/(q)$ vector space, is able to perform a suitable reduction on the result received from the unclassified level.

- Thanks to the previous property and to cleartext addition and multiplication being mapped to ciphertext addition and multiplication, respectively, the unclassified level can compute $E_k(-x^*\varepsilon_y - y^*\varepsilon_x)$ as $x^*E_k(-\varepsilon_y) + y^*E_k(-\varepsilon_x)$ over Z. This greatly facilitates unclassified computation of the encrypted perturbations.

4.1.2. Undoing the integer conversion

When the classified level receives the result of a computation from the unclassified level, it retrieves the clear perturbation and adds it to the perturbed result to obtain the exact result. If initial data and perturbations were converted to integers as discussed above, every result received from the unclassified level is a fraction; the numerator of the perturbation must be decrypted and thereafter divided over the real numbers by the decrypted denominator (be it a power of 10 or not), in order to get the right number of decimal positions.

4.1.3. Numerical examples

We next give two examples to illustrate computations with the proposed PH.

Example 3. (Multiplication) The following example is unrealistic because of the small size of p and q, and also because the amount of computation at the unclassified level is smaller than at the classified level, even if the latter could precompute the encrypted perturbations. Actually, the overhead introduced to perform a single multiplication is rather comical. For brevity and clarity, take $n = 2$, that is, perturbations are split into two parts during encryption.

Classified level

Let $p = 17$, $q = 13$, $r_p = 2$ and $r_q = 3$ be the secret key. One wants to have $x = -0.5$ and $y = 0.6$ multiplied by the unclassified level. Then random perturbations $\varepsilon_x = 0.3$ and $\varepsilon_y = -0.1$ are generated and the perturbed values $x^* = x + \varepsilon_x = -0.2$ and $y^* = y + \varepsilon_y = 0.5$ are obtained. In order to suppress decimal positions, perturbed data and perturbations are multiplied by 10, thus yielding the fractions

$$(x^*, -\varepsilon_x) = \left(\frac{\tilde{x}^*}{10}, \frac{-\tilde{\varepsilon}_x}{10}\right) = \left(\frac{-2}{10}, \frac{-3}{10}\right)$$

$$(y^*, -\varepsilon_y) = \left(\frac{\tilde{y}^*}{10}, \frac{-\tilde{\varepsilon}_y}{10}\right) = \left(\frac{5}{10}, \frac{1}{10}\right)$$

Numerators of perturbations are secretly and randomly split and then transformed according to the proposed PH, thus obtaining first and second r-degree terms

$$E_k(-\tilde{\varepsilon}_x) = E_k(-3) = E_k(1,-4) = ([2,3],[1,3])$$

$$E_k(-\tilde{\varepsilon}_y) = E_k(1) = E_k(-1,2) = ([15,10],[8,5])$$

Next, these encrypted perturbation numerators and perturbed data numerators $\tilde{x}^*, \tilde{y}^*$ are forwarded to the unclassified level, with a denominator 10 in all cases. Remark that encrypted perturbations could have been precomputed and/or reused from previous data.

Unclassified level

Compute $\tilde{x}^*\tilde{y}^* = -2 \times 5 = -10$ and the corresponding encrypted perturbation after table 1

$$\tilde{x}^* E_k(-\tilde{\varepsilon}_y) + \tilde{y}^* E_k(-\tilde{\varepsilon}_x) + E_k(-\tilde{\varepsilon}_x) E_k(-\tilde{\varepsilon}_y)$$

$$= -2 \times ([15,10],[8,5]) + 5 \times ([2,3],[1,3]) + ([2,3],[1,3]) \times ([15,10],[8,5])$$

$$= ([-2 \times 15 + 5 \times 2, -2 \times 10 + 5 \times 3],$$

$$[-2 \times 8 + 5 \times 1, -2 \times 5 + 5 \times 3] + [2 \times 15, 3 \times 10],$$

$$[2 \times 8 + 1 \times 15, 3 \times 5 + 3 \times 10], [1 \times 8, 3 \times 5])$$

$$= ([-20,-5],[19,35],[31,45],[8,15])$$

Notice that terms with different r-degree are dealt with as separate components. Return to the classified level with a denominator $10 \times 10 = 10^2$: i) the perturbed product; ii) its encrypted perturbation.

Classified level

Retrieve the clear perturbation $\tilde{\varepsilon}_{xy}$ of the product by computing

$$([-20 \times r_p^{-1} \bmod p, -5 \times r_q^{-1} \bmod q], [19 \times r_p^{-2} \bmod p, 35 \times r_q^{-2} \bmod q],$$

$$[31 \times r_p^{-3} \bmod p, 45 \times r_q^{-3} \bmod q], [8 \times r_p^{-4} \bmod p, 15 \times r_q^{-4} \bmod q])$$

$$= ([-20 \times 9 \bmod 17, -5 \times 9 \bmod 13], [19 \times 9^2 \bmod 17, 35 \times 9^2 \bmod 13],$$

$$[31 \times 9^3 \bmod 17, 45 \times 9^3 \bmod 13], [8 \times 9^4 \bmod 17, 15 \times 9^4 \bmod 13])$$

$$= ([7,7],[9,1],[6,6],[9,5]) = ([14,6])$$

where, in the last step, all terms have been added up over $\mathcal{Z}/(p) \times \mathcal{Z}/(q)$. Bearing in mind that $m = pq = 221$, use the Chinese remainder theorem on the pair $[14,6]$ to recover the perturbation $-\tilde{\varepsilon}_{xy} = 201 \bmod 221 = -20$ corresponding to $\tilde{x}^*\tilde{y}^*$. In the last equivalence, it has been implicity assumed that the upper half of $\mathcal{Z}/(221)$ represents negative integers: this assumption is valid if perturbations have an

absolute value much smaller than pq, which always occurs in real cases because pq is very large. Adding this perturbation to the perturbed product yields

$$\tilde{x}^*\tilde{y}^* - \tilde{\varepsilon}_{xy} = -10 - 20 = -30$$

Finally, divide -30 by the denominator 10^2 returned by the unclassified level. Thus the final result is $xy = -0.3$.

□

Example 4. (Average computation) This example is even more unrealistic as the previous one, but it illustrates a basic statistical operation such as an average computation —averages are the building blocks of most statistics, such as variance, skewness, kurtosis, etc. An average consists of an addition and a division. Again, take $n = 2$ (perturbations are split into two parts).

Classified level

Let $p = 17$, $q = 13$, $r_p = 2$ and $r_q = 3$ be the secret key. One wants to have the average of the sample $(x_1, x_2, x_3) = (0.3, 1.5, 1.0)$ computed by the unclassified level. Then three perturbations are generated to obtain the perturbed values (0.4, -0.1) for x_1, (1.2, 0.3) for x_2 and (0.9, 0.1) for x_3. In order to suppress decimal positions, these perturbed data are multiplied by 10, thus yielding the fractions

$$(x_1^*, -\varepsilon_1) = \left(\frac{\tilde{x}_1^*}{10}, \frac{-\tilde{\varepsilon}_1}{10}\right) = \left(\frac{4}{10}, \frac{-1}{10}\right)$$

$$(x_1^*, -\varepsilon_2) = \left(\frac{\tilde{x}_2^*}{10}, \frac{-\tilde{\varepsilon}_2}{10}\right) = \left(\frac{12}{10}, \frac{3}{10}\right)$$

$$(x_3^*, -\varepsilon_3) = \left(\frac{\tilde{x}_3^*}{10}, \frac{-\tilde{\varepsilon}_3}{10}\right) = \left(\frac{9}{10}, \frac{1}{10}\right)$$

Numerators of perturbations are secretly and randomly split and then transformed according to the new PH, thus obtaining

$$E_k(-\tilde{\varepsilon}_1) = E_k(-1) = E_k(1, -2) = ([2,3],[9,8])$$

$$E_k(-\tilde{\varepsilon}_2) = E_k(3) = E_k(-1,4) = ([15,10],[16,10])$$

$$E_k(-\tilde{\varepsilon}_3) = E_k(1) = E_k(-1,2) = ([15,10],[8,5])$$

Perturbed data and encrypted perturbation numerators are forwarded to the unclassified level, along with their denominators (10 in all cases). Remark that encrypted perturbations could have been precomputed or reused from a previous sample.

Unclassified level

Compute the perturbed average $\bar{x}^* = \frac{(\sum_{i=1}^{3} \tilde{x}_i^*)/10}{3} = \frac{(4+12+9)/10}{3} = \frac{25}{30}$. The encrypted perturbation of the numerator 25 of the average can be found by directly adding the numerators of encrypted perturbations (see table 1), because the denominator is 10 in all of them

$$\sum_{i=1}^{3} E_k(-\tilde{\varepsilon}_i)) = ([2,3],[9,8]) + ([15,10],[16,10]) + ([15,10],[8,5])$$

$$= ([2+15+15, 3+10+10], [9+16+8, 8+10+5]) = ([32,23],[33,23])$$

The denominator 30 of the average has perturbation 0, as it is the product of exact values 10 and 3. Return to the classified level: i) the perturbed average $\bar{x}^* = \frac{25}{30}$; ii) for the numerator 25, return the encrypted perturbation $([32,23],[33,23])$; iii) for the denominator 30, return perturbation 0.

Classified level

Retrieve the clear perturbation for the numerator of the average by computing

$$([32 \times r_p^{-1} \bmod p, 23 \times r_q^{-1} \bmod q], [33 \times r_p^{-2} \bmod p, 23 \times r_q^{-2} \bmod q])$$

$$= ([32 \times 9 \bmod 17, 23 \times 9 \bmod 13], [33 \times 9^2 \bmod 17, 23 \times 9^2 \bmod 13])$$

$$= ([16,12],[4,4]) = ([3,3])$$

where, in the last step, the first and the second r-degree terms have been added over $Z/(p) \times Z/(q)$. Bearing in mind that $m = pq = 221$, use the Chinese remainder theorem on the pair $[3,3]$ to recover the clear perturbation 3. The perturbation is taken as positive, because it belongs to the lower half of $Z/(221)$. Adding the perturbation 3 to the numerator of the perturbed average and the perturbation 0 to its denominator yields

$$\bar{x} = \frac{25+3}{30+0} = \frac{28}{30} (= 0.9\hat{3})$$

Remark that, in this last step, the amount of computation *is fixed and does not depend on the sample size.*

□

4.2. Multilevel processing with data suppression

The following disclosure-protected table is taken from [Cox (1992)]. Disclosure —sensitive— cells have been suppressed (primary suppressions); other cells have been secondarily suppressed to prevent inferences. A D-cell stands for a suppressed cell.

D	10	D	D	20	80
D	10	D	5	15	60
40	10	D	D	10	90
5	5	D	D	5	40
75	35	65	45	50	270

Now, an alternative approach to cell suppression is to use a PH to encrypt the D-cell values

$E_k(10)$	10	$E_k(25)$	$E_k(15)$	20	80
$E_k(20)$	10	$E_k(10)$	5	15	60
40	10	$E_k(20)$	$E_k(10)$	10	90
5	5	$E_k(10)$	$E_k(15)$	5	40
75	35	65	45	50	270

The unclassified level views the encrypted cells as if they had been suppressed, but it can perform on them the operations in F'. The classified level can take advantage of this work by decrypting the result. When using a PH breakable by a known-cleartext attack (such as the one in example 2), the unclassified level must separately operate on encrypted cells (operations in F') and unencrypted cells (operations in F); merging the results of both computational streams is up to the classified level. The use of a PH resistant to a known-cleartext attack (such as the ones in example 1 and section 3) allows the classified level to distribute the encrypted version of the whole table, along with the non-disclosure cells as cleartext for consultation purposes; thus the unclassified level can operate on the whole encrypted table using the operations in F', so that the only job left to the classified level is decryption of the final result.

4.3. Multilevel processing of resampled data

The basic resampling scheme dealt with in this subsection is from [Heer (1992)]. Assume that microdata $z_1, \cdots, z_n$ are aggregated to elaborate macrodata in the form of a contingency table x with I rows and J columns, which is produced according to certain specifications. Let x_{ij} be the original frequency in the i-th row and j-th column. In order to produce an anonymized table x^*, a bootstrap sample $z_1^*, \cdots, z_n^*$ is obtained by drawing from the original data $z_1, \cdots, z_n$, n times and with replacement. The bootstrap table x^* thus obtained is an estimation of the original table x and does not allow to get any precise information of x, due to its random error.

The main features of a bootstrap table are

- The overall frequency is preserved, since $\sum x_{ij} = \sum x^*_{ij} = n$.
- Each individual bootstrap frequency x^*_{ij} is a value drawn from a variable having a binomial distribution with mean x_{ij} and variance $x_{ij}(1 - x_{ij}/n)$. Therefore, x^* is an unbiased estimation of x.
- An original frequency $x_{ij} = 0$ is preserved, *i. e.* $x_{ij} = 0$ implies $x^*_{ij} = 0$. If this is undesirable, then a compensated perturbation method could be used on the original table before bootstrapping, in order to replace zero frequencies with small frequencies.

It can be seen from the previous paragraphs that x^* is actually a perturbed image of x, *i. e*

$$x^* = x + (x^* - x) = x + \varepsilon$$

where $\varepsilon = (\varepsilon_{ij})$ is a matrix of random zero-mean estimation errors. Therefore, privacy homomorphisms can be used in a way analogous to the one described in subsection 4.1. The classified level computes the matrix ε and releases the following pairs for unclassified processing

$$(x^*_{ij}, E_k(-\varepsilon_{ij})) \quad \forall i = 1, \cdots, I, \forall j = 1, \cdots, J$$

5. CONCLUSION

The use of privacy homomorphisms for multilevel processing of classified statistical data has been motivated. A new privacy homomorphism has been presented which exhibits good cryptographic and algebraic properties for statistical computation. It has been shown how to combine privacy homomorphisms and several well-known techniques for statistical confidentiality, so that the classified level can recover exact statistics from statistics obtained at an unclassified level on disclosure-protected data. Application of PHs for providing statistical confidentiality in on-line scenarios is currently being investigated. The key issue is that PHs useful for statistical data protection are not necessarily cryptographically unbreakable PHs: they only need be as strong as the statistical confidentiality techniques with which they are combined.

ACKNOWLEDGMENT

Thanks go to two anonymous referees for helpful suggestions that contributed to improve the presentation of this work.

REFERENCES

[1] **N.R. Adam** and **J.C. Wortmann** (1989). «Security-control methods for statistical databases: a comparative study». *ACM Computing Surveys*, **21, 4**, Dec. 1989, 515–556.

[2] **G. Appel** and **D.J. Hoffman** (1992). «Perturbation by compensation». Preproceedings of the *International Seminar on Statistical Confidentiality*, ISI-Eurostat, Dublin, Sep. 1992. Final version in the Proceedings published by Eurostat in 1993.

[3] **E. Brickell** and **Y. Yacobi** (1988). «On privacy homomorphisms», in *Advances in Cryptology-Eurocrypt'87*, eds. W. L. Price and D. Chaum, Springer-Verlag, 117–125.

[4] **L.H. Cox** (1992). «Solving confidentiality protection problems in tabulations using network optimization: a model for cell suppression in U. S. economic censuses». Preproceedings of the *International Seminar on Statistical Confidentiality*, ISI-Eurostat, Dublin, Sep. 1992. Final version in the Proceedings published by Eurostat in 1993.

[5] **J. Domingo** (1996). «A new privacy homomorphism and applications», working paper, Universitat Rovira i Virgili.

[6] **D. Schackis** (1993). *Manual on Disclosure Control Methods*, internal document, Eurostat (unit D3 - Research and development and statistical methods).

[7] **G. Heer** (1992). «A bootstrap procedure to preserve statistical confidentiality in contingency tables». Preproceedings of the *International Seminar on Statistical Confidentiality*, ISI-Eurostat, Dublin, Sep. 1992. Final version in the Proceedings published by Eurostat in 1993.

[8] **R.L. Rivest, A. Shamir** and **L. Adleman** (1978a). «A method for obtaining digital signatures and public-key cryptosystems». *Communications of the ACM*, **21**, Feb. 1978, pp. 120–126.

[9] **R.L. Rivest, L. Adleman** and **M.L. Dertouzos** (1978b). «On data banks and privacy homomorphisms», in *Foundations of Secure Computation*, eds. R. A. DeMillo et al., Academic Press, 169–179.

[10] **J. Schlörer** (1979). *Disclosure from Statistical Databases: Quantitative Aspects of Trackers*, Institut für Medizinische Statistik und Dokumentation, Universität Giessen, Mar. 1979.

[11] **J. Turmo** (1993). «Pertorbacions aleatòries amb compensació: una tècnica per a la protecció d'informació estadística confidencial». *Qüestiió*, **17, 3**, 413–435.

QÜESTIIÓ, Vol 20, 3 pp. 523-546, 1996

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

ALFONS GARÍN i RAMÍREZ*

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

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

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

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

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

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

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

1. INTRODUCCIÓ

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

1.1. El concepte de revelació estadística

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

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

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

ii) obtenció de valors superiors als acceptables;

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

1.2. Comissions de protecció de dades

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

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

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

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

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

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

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

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

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

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

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

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

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

a) es produirà una mostra de registres anònims,

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

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

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

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

4. Resultats i conclusions.

5. Estructura de l'arxiu final.

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

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

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

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

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

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

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

2.1. Motius de l'espionatge estadístic

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

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

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

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

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

2.2. Regles d'identificació

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

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

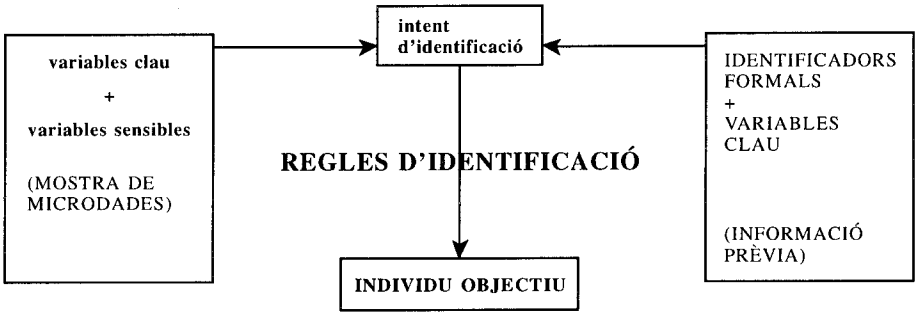
b) De l'arxiu estadístic de microdades (anònim), l'usuari triarà les variables que mantinguin coherència conceptual i de nivell de desagregació amb el seu arxiu d'identificadors formals, amb l'objectiu de realitzar una comparació de continguts entre les dues informacions.

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

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

2.3. Escenari d'identificació

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



Esquema 1

2.4. Components del risc de revelació

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

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

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

  i) els sistemes de codificació emprats en ambdós casos,

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2.5.1. Mètodes de càlcul de la proporció de poblacions úniques

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

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

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

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

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

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

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

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

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

3.1. Fases principals del procés

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

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

Així doncs, tenim:

N: dimensió de la població

f: factor de mostreig

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

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

III. Determinació de les variables clau.

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

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

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

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

VIII. Inferència de p(identificació).

3.2. Estimadors dels components de p(identificació)

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

Definim:

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

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

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

$p(a) = 1$;

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

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

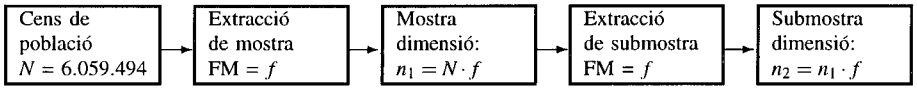
estimador de $p(d) = u/u_2$; la proporció de poblacions úniques vertaders a la submostra respecte a les observades a la submostra, s'utilitza com a estimador de la probabilitat de que una població única, observada a la mostra, sigui registre únic de la població; l'estimador de l'expressió (2) serà:

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

4. RESULTATS I CONCLUSIONS

4.1. Aplicació del procés

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



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

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

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

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

$E(n_1) = f \cdot N$

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

4.2. Determinació de les variables clau

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

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

variables de localització:

Localització de residència

Lloc de naixement

Altres variables:

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

Sexe

Estat civil

Ocupació o Ofici o Professió

Nivell d'estudis acabats

Relació amb l'activitat.

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

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

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

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

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

4.4. L'anàlisi dels tres estrats

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

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

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

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

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

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

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

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

Resultats del càlcul de p(identificació):

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

Taula 1

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

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

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

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

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

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

Característiques de la mostra

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

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

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

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

Variables presents en la mostra i nivells de desagregació

Es distingeixen dos tipus de variables:

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

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

(1) Variables de l'individu

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

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

(2) Variables de l'habitatge

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

REFERÈNCIAS METODOLÒGIQUES CONSULTADES

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

ENGLISH SUMMARY

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

ALFONS GARÍN i RAMÍREZ*

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

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

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

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

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

– Received april 1996.

– Accepted september 1996.

1. INTRODUCTION

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

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

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

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

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

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

Microdata dissemination policy

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

Data protection methodology

There are two general methods:

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

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

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

Model for studying disclosure risk

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

- The predictive model proposed by Duncan and Lambert [Duncan i Lambert (1986)]; a Bayesian model, in wich the identification of the target individual is not crucial. It is based on both the current and the previous predictive distributions, connected by an uncertainty function.
- The model, which we might call of identification , based on the analysis of the frequency of the variables that may be used to identify a certain individual. This method uses data from the available sample and drastically separates identificative variables from variables containing sensitive information, which won't be used as an identificative key.

The IEC follows the second method.

Identification rules

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

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

Main problem in estimating disclosure risk: population uniques

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

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

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

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

The problem of variables containing geographic references

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

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

Biometria

Consell Directiu de la «Región Española de la International Biometric Society» (1996)

President:	C.M. Cuadras
Vice-president:	E. Carbonell
Secretari:	F. López-Santovería
Tresorer:	M. Dolores Sánchez
Vocals:	J.L. Chorro
	R. Estarelles
	G. Gómez (corresponsal del Biometric Bulletin)
	M. Ríos

Adreça electrònica: http://www.iata.csic.es/IBSREsp/

QÜESTIIÓ, Vol 20, 3 pp. 549-562, 1996

SENSIBILIDAD DEL ANÁLISIS DE LA «ESTRUCTURA PROPIA» EN LA DETECCIÓN DE COLINEALIDAD

ESTARELLES, R.*

PRIETO, J.*

Universidad de Valencia

Muchos son los criterios, tanto de carácter formal como informal, que han ido proponiéndose para la detección de la colinealidad. Sin embargo, muchos de ellos no ofrecen un diagnóstico eficaz que permita evaluar el nivel de alteración que se produce en las estimaciones obtenidas, variables realmente implicadas... En el presente trabajo se presentan las ventajas de utilizar índices basados en la «estructura propia» de la matriz de predictores frente a otras estrategias formales de uso más frecuente. Este objetivo se pone de manifiesto mediante el análisis de datos reales.

Sensibility analysis of the proper structure in detecting collinearity

Keywords: Diagnóstico de Colinealidad, Factor de Inflación de la Varianza, Descomposición en Valores Singulares, Número de Condición, Descomposición de la Varianza de los Coeficientes en Proporciones.

*Estarelles, R. y Prieto, J. Metodología de las Ciencias del Comportamiento. Facultad de Psicología. Universidad de Valencia.

- Article rebut el gener de 1996.
- Acceptat el desembre de 1996.

1. INTRODUCCIÓN

La presencia de un elevado nivel de colinealidad es un hecho frecuente en gran número de investigaciones, y sus repercusiones afectan tanto a nivel estadístico como de cálculo en los resultados. Entre las repercusiones estadísticas más importantes se encuentran la imposibilidad para obtener estimaciones únicas para los coeficientes de regresión, incrementos en la variabilidad de los coeficientes estimados, errores de tipo II más elevados, sensibilidad ante pequeñas variaciones en la variable respuesta, dificultades en la interpretación de los resultados, entre otras. Dadas las implicaciones que suelen presentarse, el investigador deberá tener un claro conocimiento del nivel de incidencia y alteraciones que este fenómeno pudiera estar produciendo, número de relaciones colineales y su intensidad, su existencia de forma simultánea o independiente, su implicación en relaciones equivalentes o dominantes, así como si las alteraciones son lo suficientemente nocivas para los objetivos propuestos como para plantearse la utilización de estrategias alternativas. Por todo ello, el estudio de esta deficiencia y sus implicaciones ha sido el aspecto central de gran número de manuales y artículos sobre regresión desde hace años (Gunst y Mason, 1977; Belsley, Kuh y Welsch, 1980; Chatterjee y Price, 1977; Estarelles, 1989; Fox, 1991; Belsley, 1991, etc), habiéndose propuesto un amplio número de estrategias para su detección y tratamiento.

Desde un enfoque aplicado han ido ofreciéndose acercamientos basados fundamentalmente en la matriz de correlaciones, tal como el propio análisis de los elementos dicha matriz, su determinante, evaluación de correlaciones múltiples y parciales, o bien la consideración del *Factor de Inflación de la Varianza* «*VIF*» (Marquardt, 1970), así como otros índices derivados de este factor, como son el «Índice de Tolerancia» o el «Valor Promedio para VIF» (Berk, 1977). Sin embargo, desde la perspectiva del Análisis Numérico se han ido presentando estrategias, que aunque no pueden considerarse como independientes de las anteriores, han demostrado ser bastante más precisas (Berk, 1977; Mandel, 1982; Stewart, 1987, entre otros). Estos acercamientos, surgieron como un medio para disminuir los errores de cálculo derivados de la presencia de matrices próximas a la singularidad, y se han basado en el análisis de la «estructura propia» de la matriz de datos (Golub y Reinsch, 1970; Stewart, 1973; Lawson y Hanson, 1974; Mandel, 1982; etc). Esta línea de trabajo ha permitido profundizar en el conocimiento de las propiedades de la «estructura propia» de la matriz de predictores, ofreciendo una magnífica herramienta para la identificación y estudio de la dependencia entre variables, como es la descomposición de la matriz X en valores singulares, tal que $X = UDV'$, siendo U los «vectores propios» o «eigenvectores» de la matriz XX', V los «vectores propios» de la matriz $X'X$ y D los valores singulares, o lo que es lo mismo, las raíces cuadradas de los «eigenvalues» o «raíces características» para la matriz $X'X$.

1.2. «Estructura propia» de la matriz de predictores

El análisis de la «estructura propia» de la matriz X, a pesar de no ser un enfoque reciente para el diagnóstico de la colinealidad (Kendall, 1957; Silvey, 1969; Chatterjee y Price, 1977, etc), dista mucho de ser uno de los procedimientos más utilizados, y algunas de las estrategias de él derivadas, como es la «Descomposición de la Varianza de los Coeficientes en Proporciones», a pesar de haber demostrado su gran utilidad, han sido prácticamente ignoradas desde una perspectiva aplicada (Estarelles, 1989), siendo lamentable la falta de interés mostrada hacia enfoques que resultaron innovadores al inicio de la década de los 80 (Belsley, Kuh y Welsch, 1980), como es el análisis conjunto de dicha descomposición respecto al desglose de la matriz de predictores en valores singulares.

Posiblemente el procedimiento más utilizado desde esta perspectiva ha venido siendo la consideración de las «raíces características», y de los «vectores propios» para la matriz de correlaciones. Este criterio, a pesar de tener la ventaja de permitir analizar el número de relaciones colineales en un conjunto de datos (un valor de cero para los valores singulares estaría determinando la nulidad de X), no cuenta, sin embargo, con un punto de corte formal, a partir del cual pueda inferirse que se producirán graves alteraciones en los resultados. Asimismo, valores elevados para el vector propio suelen estar asociados con problemas de colinealidad, en cambio, si estos valores fuesen reducidos no podría desprenderse la conclusión opuesta (Belsley, 1984). Igualmente, una matriz bien condicionada podría incorporar raíces características con valores reducidos, por tanto, más que el análisis individual de raíces y vectores característicos, se precisará de otros acercamientos que permitan comparar la dispersión entre las «raíces características», intensidad con que se presenta el condicionamiento de los datos, nivel de perturbación en los resultados, así como análisis de las variables implicadas en cada relación colineal y tipo de dependencia establecido, independientemente de que, de forma complementaria, puedan incorporarse otras estrategias, tal como considerar el inverso de las «raíces características» (Johnston, 1984; Fox, 1991), la obtención de «Componentes Principales» (Kloeck y Mennes, 1960; Jollife, 1982), o el análisis de la estructura, incluyendo la variable respuesta, tal como si de un modelo de «Raíces Latentes» se tratara (Estarelles, Prieto, Castro y Aragón, 1995), entre otras.

Teniendo en cuenta que el hecho de que un valor singular sea reducido puede no ser indicativo de la presencia de colinealidad, el condicionamiento deficiente de una matriz podrá manifestarse a partir del producto de su máximo valor singular, es decir, de la norma espectral para aquella matriz, por el máximo valor singular de su inversa (Belsley, 1991). Este resultado conocido como «Índice de Condicionamiento» o «Número de Condición», fue introducido por Turing en 1948 para delimitar perturbaciones en la relación de sistemas lineales, y aplicado posteriormente para soluciones mínimo cuadráticas por Golub y Wilkinson (1966), con el fin de detectar

la sensibilidad de la ecuación ante modificaciones en los elementos del sistema. Teniendo en cuenta que puede demostrarse que $\|X\| = \mu_{máx}$, así como que si X es una matriz cuadrada $\|X^{-1}\| = 1/\mu_{mín}$ (Wilkinson, 1965), la relación entre ambos términos mostrará el condicionamiento deficiente de una matriz (para una exposición detallada ver Belsley, 1984):

$$IC = \frac{\mu_{máx}}{\mu_{mín}} = \sqrt{\frac{\lambda_{máx}}{\lambda_{mín}}} \geq 1 \tag{1}$$

siendo $\lambda_{máx}$ y $\lambda_{mín}$ las raíces características para la matriz de predictores de tamaño superior e inferior, respectivamente.

Para conocer el número de dependencias, la lógica presentada podrá extenderse para el resto de valores singulares, pudiendo obtenerse tantos índices como predictores incorporados. Estos cocientes se conocen como «Índices de Colinealidad». No obstante, aún se desconocerá el nivel de distorsión que pueden presentar los resultados. Una estrategia que permite acometer con cierta garantía de éxito este objetivo es la «Descomposición de la Varianza de los Coeficientes de Regresión en Proporciones», introducida por Silvey en 1969.

Teniendo en cuenta que la matriz de Varianza-Covarianza para los coeficientes de regresión mínimo-cuadráticos se obtiene de: $V_{(b)} = \sigma^2(X'X)^{-1}$, esta matriz podrá expresarse equivalentemente a partir de la descomposición de X en valores singulares, tal que $V_{(b)} = \sigma^2 V D^{-2} V'$, siendo la varianza para cada coeficiente:

$$\text{var}_{(b_i)} = \sigma^2 \sum_j \frac{v_{ij}^2}{u_j^2} \tag{2}$$

siendo $V \equiv v_{ij}$ y u_j los valores singulares obtenidos. Por tanto, cuanto más reducidos sean estos valores, previsiblemente mayor será la variabilidad del coeficiente, aunque de la expresión anterior igualmente se deduce que una varianza del error reducida disminuirá los efectos de la colinealidad en la variabilidad de los coeficientes. No obstante, otras situaciones especiales también influirán en el mismo sentido, tal sería el caso de la presencia de colinealidad en dos conjuntos de variables que resultasen ser ortogonales entre sí (Belsley, 1991). La «Descomposición de la Varianza de los Coeficientes en Proporciones» se obtendrá a partir de:

$$\pi_{ji} = \frac{\omega_{ij}}{\omega_i}, \quad \text{siendo } \omega_i = \Sigma\, \omega_{ij} \quad \text{y} \quad \omega_{ij} = \frac{v_{ij}^2}{u_j^2} \tag{3}$$

La representación mediante una tabla conjunta de «Índices de Colinealidad» y «Descomposición de la Varianza» ofrece un medio para evaluar la degradación que

sufren los coeficientes estimados, informando del número y tipo de dependencias mostrado por las variables. Cuando esta identificación se torna compleja, como en el caso de que se presentasen relaciones equivalentes, el papel desempeñado por cada uno de los predictores podrá clarificarse mediante la utilización de «Regresiones Auxiliares» (Belsley, 1991; Estarelles, Castro y Prieto, 1995), aunque si bien estas regresiones solo deberán plantearse con fines descriptivos, no inferenciales, teniendo sentido su aplicación cuando tras el proceso de diagnóstico pudiera cuestionarse el papel desempeñado por las variables implicadas.

En el presente trabajo se considera el comportamiento de algunas de las estrategias más usuales en la detección de la colinealidad, destacándose las ventajas de aplicar estrategias basadas en la «estructura propia» respecto a otros acercamientos de uso más frecuente en investigación aplicada. Asimismo, se ha considerado de interés resaltar el comportamiento de los índices estudiados a partir de una base de datos reales, ya que éste es el punto de referencia más habitual en investigación aplicada, en donde difícilmente se encuentran aquellas situaciones teóricas ideales que pueden plantearse a partir de una simulación. Las estrategias seleccionadas en el presente estudio no representan todo el abanico de posibilidades que incorpora la literatura, ya que la inclusión de muchas de ellas ofrecería información redundante, no encontrándose entre los objetivos de este estudio el ofrecer una revisión exhaustiva de los numerosos índices que han ido desarrollándose. Para una exposición detallada ver Estarelles y Prieto (1995).

2. APLICACIÓN A PARTIR DE DATOS REALES

A partir de una muestra de 83 jóvenes, de edades comprendidas entre 20 y 23 años, compuesta por aproximadamente igual número de sujetos de ambos sexos, se pretendió explicar el nivel de *Auto-Satisfacción con el Comportamiento Personal*, a través de variables como: *Percepción de Afecto Familiar*, *Auto-Concepto Ético-Moral*, *Auto-Concepto Físico*, *Pensamientos Depresivos*, *Ausencia de Sentimientos de Culpabilidad* y *Problemas Sociales*.

Tras la realización de los análisis preliminares pertinentes, no se detectó la presencia de puntuaciones excesivamente extremas, en ninguna de las variables (ni aparentemente de forma conjunta), ni ninguna otra violación a los supuestos esenciales del modelo, pudiendo asumirse una relación lineal entre la variable criterio y los predictores. Sin embargo, podrían considerarse objeto de un análisis más detallado las correlaciones relativamente elevadas que se observan entre los predictores (Tabla 1), haciendo sospechar la posible presencia de relaciones colineales que pudieran estar afectando los resultados. No obstante, este aspecto, posiblemente por no tener connotaciones alarmantes, podría haber pasado erróneamente desapercibido

en una investigación aplicada que no tuviera unos intereses esencialmente metodológicos.

Tabla 1

Correlaciones entre las variables

VARIABLE	A.Sat.(Y)	A.Fam.(X_1)	A.Et.(X_2)	A.Fis.(X_3)	P.Dep.(X_4)	Pr.Soc.(X_5)	A.S.C(X_6)
A.Sat.(Y)	1.000	–	–	–	–	–	–
A.Fam.(X_1)	0.620	1.000	–	–	–	–	–
A.Et.(X_2)	0.650	0.580	1.000	–	–	–	–
A.Fis.(X_3)	0.703	0.699	0.633	1.000	–	–	–
P.Dep.(X_4)	−0.24	−0.12	−0.34	−0.02	1.000	–	–
P.Soc.(X_5)	0.804	0.752	0.726	−0.24	0.571	1.000	–
A.S.C.(X_6)	0.777	0.681	0.645	−0.26	0.500	0.707	1.000

Tras el estudio de cada uno de los predictores independientemente, así como tras el análisis de todos los subconjuntos posibles, se procedió a comparar el comportamiento de los coeficientes estimados y su variabilidad, observándose como los valores para los coeficientes iban modificándose a medida que se alteraba el conjunto de predictores, comprobándose igualmente la falta de significación de variables relevantes, así como cambios de signos no justificados en el valor de algunos coeficientes. Como puede observarse, los signos para los coeficientes de las variables X_1 y X_2 resultan modificados, y los valores para el resto de coeficientes se ven distorsionados. Sin embargo, tal como puede comprobarse en la tabla 2, no se observa un incremento notable en la variabilidad de los coeficientes. Tal como Myers (1990) demuestra, van a ser todos los coeficientes los que de alguna manera van a verse afectados por la colinealidad de los datos, lo cual no implica que esta deficiencia no esté alterando la estimación e interpretación del modelo, especialmente en casos en los que la colinealidad se da de forma simultánea (Estarelles, 1989), tal como veremos ocurre en el caso estudiado.

En este sentido, y como resumen de los sucesivos análisis realizados, a continuación se exponen los resultados para cada predictor respecto al criterio, y tras el análisis global del conjunto de predictores seleccionados.

Para la detección de la presencia de posibles relaciones colineales, es usual el considerar el «Factor de Inflación de la Varianza» *(VIF)* para cada uno de los predictores:

$$VIF_{(x_j)} = \frac{1}{1 - R_j^2} \tag{4}$$

siendo R_j^2 el *Coeficiente de Determinación* resultante de regresar la variable X_j respecto al resto de predictores incluidos en el modelo. También es común la utilización de otros índices derivados como es el nivel de «Tolerancia» $(1 - R_j^2)$. Estos valores suelen incluirse de forma automática por los principales paquetes de Software. En el presente estudio se ha incluido igualmente el valor promedio para el *Factor de Inflación*. La realización de los análisis sucesivos se ha efectuado a partir de una matriz de predictores homogeneizada, tal como se estima pertinente para este tipo de situaciones (Stewart, 1987; Myers, 1990; Belsley, 1991, etc).

Tabla 2

Análisis de regresión para cada uno de los predictores y resultados globales

ANÁLISIS INDIVIDUAL				
Variables	Coef.	Err.Est.	t	p
Af.Fam.(X_1)	0.773	0.109	7.111	0.0000
A.Et.M.(X_2)	0.737	0.096	7.698	0.0000
A.Fís.(X_3)	1.260	0.142	8.898	0.0000
P.Depr.(X_4)	−0.622	0.273	−2.275	0.0255
Pr.Soc.(X_5)	1.347	0.111	12.167	0.0000
A.St.C.(X_6)	1.318	0.119	11.095	0.0000
ANÁLISIS CONJUNTO ($R^2 = 0.79$)				
Variables	Coef.	Err.Est.	t	p
Af.Fam.(X_1)	−0.375	0.111	−3.376	0.0012
A.Et.M.(X_2)	−0.153	0.096	−1.600	0.1137
A.Fís.(X_3)	0.737	0.144	5.136	0.0000
P.Depr.(X_4)	−0.197	0.143	−1.374	0.1736
Pr.Soc.(X_5)	0.856	0.158	5.418	0.0000
A.St.C.(X_6)	0.789	0.133	5.911	0.0000

Tabla 3

Inflación de la varianza «VIF»

Af.Fam.(X_1)	A.Et.(X_2)	A.Fís.(X_3)	P.Depr.(X_4)	Pr.Soc.(X_5)	A.St.C.(X_6)
3.10397	2.78129	2.51663	1.24895	3.48798	2.4248

Si se asume el punto de corte habitualmente establecido en la literatura de 10, aparentemente ninguna de las relaciones en el conjunto de datos parece mantener una relación excesivamente conflictiva para la estimación del modelo. Este resultado guarda relación con la falta de aparentes alteraciones en los errores estándar que se observaron (tabla 2), y que muestra la falta de fiabilidad de «*VIF*» para determinados conjuntos de datos, especialmente cuando las dependencias se dan de forma simultánea. Por otro lado, el valor promedio para este índice es igual a 2.59. La superioridad de este resultado respecto a la unidad no es excesivamente elevada, sin embargo, las alteraciones anteriormente observadas aconsejan un diagnóstico más riguroso, tal como el que se desprende del análisis de la «estructura propia». En la tabla siguiente se incluyen las raíces características para la matriz de predictores.

Tabla 4

Raíces características de la matriz de predictores

λ_0	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
0.0047	0.0044	0.0063	0.0075	0.0021	0.0444	6.9348

Tal como puede observarse, existen cinco raíces con valores muy próximos a cero, lo cual permite identificar el condicionamiento deficiente de los datos, especialmente si tenemos en cuenta que, aún más importante que la proximidad de este valor a cero, es la gran disparidad existente entre el valor máximo respecto al resto de valores para las raíces características inferiores. Los «Índices de Colinealidad» para cada uno de los predictores se incluyen en la tabla siguiente. Se ha estimado conveniente la inclusión de la constante, con el fin de comprobar si existe algún tipo de relación colineal con este término.

Tabla 5

Índices de colinealidad

IC_0	IC_1	IC_2	IC_3	IC_4	IC_5	IC_6
38.56	39.78	33.33	30.44	56.88	12.50	1.00

En la literatura se han sugerido puntos de corte aproximados para este índice (Belsley, Kuh y Welsch, 1980; Fox, 1991, etc), habiéndose asumido que valores superiores a 30 pondrán de manifiesto la presencia de distorsiones importantes. El número de «Índices de Colinealidad» que resulten superiores al valor de corte asumido indicará el número de dependencias existentes. Como puede observarse, la presencia de cinco «Índices de Colinealidad» por encima de 30 ratifican el problema de dependencia ya apuntado.

Como puede comprobarse, el análisis de la «estructura propia» permite analizar con mayor precisión que el «Índice de Inflación de la Varianza» la presencia de relaciones colineales, así como la intensidad de esta deficiencia. No obstante, un aspecto de interés, aún no claramente determinado, será la identificación y vinculación entre aquellas variables principalmente implicadas. Para ello se procederá a la «Descomposición de la Varianza» de cada uno de los coeficientes en proporciones, analizándose la incidencia de estos valores respecto a los «Índices de Colinealidad» previamente obtenidos.

Tabla 6

Diagnóstico de la colinealidad mediante la descomposición de la varianza

IC_j	Descomposición de la Varianza en Proporciones						
	b_0	b_1	b_2	b_3	b_4	b_5	b_6
1	0.001	0.003	0.009	0.012	0.003	0.000	0.007
12.50	0.072	0.049	0.045	0.026	0.028	0.051	0.092
30.44	0.206	0.278	0.312	0.123	0.017	0.093	0.108
33.33	0.280	0.214	0.170	0.274	0.039	0.201	0.093
38.56	0.176	0.149	0.163	0.192	0.056	0.082	0.068
39.78	0.168	0.210	0.159	0.235	0.073	0.102	0.113
56.88	0.097	0.097	0.142	0.138	0.784	0.471	0.519

Teniendo en cuenta que son cinco los «Índices de Colinealidad» superiores a 30, el análisis de la tabla se centrará especialmente en estos valores. El estudio de los resultados obtenidos permite identificar que cuatro de las relaciones mostradas son equivalentes, lo cual dificulta la lectura e interpretación del papel desempeñado por cada predictor en las mismas. En estos casos la distribución de la variabilidad, a lo largo de los índices con valores similares, suele presentarse de forma arbitraria, por lo que el papel de cada una de las variables en dichas dependencias queda obscurecido. En este caso, la variabilidad de los coeficientes podría quedar distribuida a lo largo de los «Índices de Colinealidad» con relaciones equivalentes, sugiriéndose en tal caso, como punto de corte aproximado para pensar en la degradación de los coeficientes, que el sumatorio de dichas proporciones de varianza ascienda a .50 (Belsley, 1991), siendo evidente la alteración que sufre la variabilidad del término constante.

Por otro lado, partiendo de que el «Número de Condición» asciende a 56.88, podría considerarse que en cierta medida este índice está mostrando una depedencia dominante respecto a los anteriores, quedando claramente implicados el cuarto, quinto y sexto predictor del modelo. Este hecho puede dificultar la ubicación de aquellos

predictores respecto a su posible implicación en otras dependencias. En este caso, por reducido que fuese el porcentaje de varianza que aportasen respecto a otras relaciones, no podría descartarse sin más su papel en aquellas dependencias. Evidentemente ninguno de los predictores está exento de formar parte de las dependencias que se presentan, pudiendo inferirse claramente la simultaneidad en la colinealidad del modelo. El estudio descriptivo que se llevó a cabo posteriormente se resume en la tabla siguiente.

Tabla 7

Resumen regresiones auxiliares

PREDICTOR	VARIABLES CON PESO REGRESION AUXILIARES	R^2
X_1	2, 3, 5 y 6	0.57
X_2	1, 3, 5 y 6	0.52
X_3	1 y 2	0.39
X_4	5 y 6	0.31
X_5	1, 2, 4 y 6	0.63
X_6	1, 2, 4 y 5	0.59

Como puede observarse, el análisis de datos reales se aleja considerablemente de aquellas situaciones ideales que se ofrecen como prototipo en los manuales, y en donde, sin excesivo problema, es factible seleccionar las variables más representativas en cada dependencia para su posterior estudio mediante análisis adicionales. Sin embargo, el enfoque utilizado permite analizar el número y tipo de dependencias así como la degradación sufrida por los coeficientes, poniéndose en evidencia el ciclo de estrechas interacciones que de forma tan usual se presentan en gran número de investigaciones, con la consiguiente alteración en los resultados, pero no siempre con incrementos notables en la variabilidad de los coeficientes.

De los resultados expuestos se desprende la importancia de la relación entre la variable «Pensamientos Depresivos» y las variables «Problemas Sociales» y «Culpabilidad», y cómo estas dos variables, a su vez, están relacionadas con «Percepción de Afecto Parental» y «Auto-Concepto Ético-Moral», y por tanto, la relación parcial de estas variables con «Pensamientos Depresivos». De igual manera, la interacción entre la variable «Auto-Concepto Físico» con las variables «Percepción de Afecto Parental» y «Auto-Concepto Ético- Moral» es destacada, de ahí la relación de esta variable con «Problemas Sociales» y «Culpabilidad». De todo ello se deduce, por tanto, la simultaneidad de las interacciones observadas, así como las alteraciones observadas

como consecuencia del condicionamiento de los datos, y que podría sugerir su estudio a través de la consideración de efectos directos e indirectos.

3. CONCLUSIONES Y DISCUSIÓN

El tipo de datos que se han considerado en el presente estudio, y las relaciones que entre las variables se presentan, son un aspecto habitual en gran número de investigaciones sociales. A pesar de ello, no es usual que el investigador se detenga a analizar los posibles efectos de las relaciones entre los predictores, y menos aún si el nivel de alteración ocasionado pudiera cuestionar los resultados y conclusiones buscados.

Ahora bien, cuando en los resultados se hace referencia al empleo de un índice de diagnóstico para detectar la presencia de colinealidad, el más habitual es sin duda el *Factor de Inflación de la Varianza*, a pesar de contar con gran número de limitaciones, como es la ausencia de un punto de corte formal, su falta de precisión cuando la matriz de predictores es casi singular, su imposibilidad para facilitar la detección del número de conjuntos de variables implicados en dicho fenómeno, así como el nivel de alteración que estuviera afectando la estimación del modelo. La consideración de las raíces características de una matriz es un criterio más adecuado en la medida en que permite analizar con mayor precisión el número de relaciones colineales. Sin embargo, tampoco esta estrategia cuenta con un punto de corte formal para la identificación de dependencias, siendo de mayor utilidad la obtención de los valores singulares para la matriz de predictores, y a partir de aquí calcular otros instrumentos de diagnóstico más precisos como son los «Índices de Colinealidad» y la «Descomposición de la Varianza de los Coeficientes de Regresión en Proporciones». Los «Índices de Colinealidad» permiten analizar el número de relaciones colineales y su importancia, sin embargo, tienen el inconveniente de ser muy susceptibles ante variaciones en el tipo de escala, así como ante ligeras modificaciones en las observaciones que componen el conjunto muestral. Por tanto, de forma complementaria será conveniente proceder al análisis conjunto de estos índices respecto a la «Descomposición de la Varianza» de los coeficientes, con el fin de identificar el tipo y número de dependencias, implicación de los predictores en cada relación colineal, así como nivel de alteración sufrido por las estimaciones. Con este fin podrán incluirse análisis de regresión adicionales, así como la evalución de efectos directos e indirectos, como en la situación analizada en el presente estudio.

Dependiendo del tipo de interés del analista este proceso de diagnóstico podría mejorarse mediante la utilización de estrategias que no sólo detecten la perturbación que presentan los coeficientes estimados, sino las nefastas consecuencias que se derivan. Ello requiere de estudios más sofisticados, como es la consideración de «elipsoides

de isodensidad γ» (Belsley, 1991; Estarelles, Castro y Prieto, 1995), pudiendo complementarse esta información con un análisis pormenorizado de la influencia ejercida en el condicionamiento de los datos por las observaciones que componen el conjunto muestral (Estarelles, 1995). Estos criterios permitirán decidir con mayor precisión la posibilidad de utilizar estrategias de análisis alternativas. Las ventajas y aplicación de estos criterios de forma accesible para el usuario serán objeto de atención en futuros trabajos.

REFERENCIAS

[1] **Belsley, D.A.** (1984). «Eigenvector Weaknesess and Other Topics for Assessing Conditioning Diagnostics. Letters». *Technometrics*, **26**, 297–299.

[2] **Belsley, D.A.** (1991). *Conditioning Diagnosis*. New York. John Wiley & Sons.

[3] **Belsley, D.A., Kuh, E.** y **Welsch, R.E.** (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York. John Wiley & Sons.

[4] **Berk, K.N.** (1977). «Tolerance and Condition in Regression Computations». *J. of the American Statistical Association*, **72**, 863–866.

[5] **Chatterjee, S.** y **Price, B.** (1977). *Regression Analysis by Example*. New York. Wiley.

[6] **Estarelles, R.** (1989). *Colinealidad en el Análisis de Datos*. I Symposium de Metodología de las Ciencias del Comportamiento. Salamanca.

[7] **Estarelles, R.** (1995). *Identificación de Observaciones con Influencia en Datos Debilmente Condicionados*. 5ª Conferencia de Biometría. Valencia.

[8] **Estarelles, R., Castro, J.** y **Prieto, J.** (1995). *Criterios formales y gráficos en la detección de colinealidad*. 5ª Conferencia de Biometría. Valencia.

[9] **Estarelles, R.** y **Prieto, J.** (1995). *Colinealidad: Efectos, Detección y Tratamiento*. Ed. Orenga. Valencia.

[10] **Estarelles, R., Prieto, J., Castro, J.** y **Aragón, J.L.** (1995). *Componentes Principales* vs *Raíces Latentes como Técnicas Alternativas a OLS*. 5ª Conferencia de Biometría. Valencia.

[11] **Fox, J.** (1991). *Regression Diagnostics*. Newbury Park. Sage Publications.

[12] **Golub, G.H.** y **Reinsch, C.** (1970). «Singular Value Decomposition and Least-Squares Solutions». *Numerische Mathematik*, **14**, 403–420.

[13] **Golub, G.H.** y **Wilkinson, J.H.** (1966). «Note on the Iterative Refinement of Least Squares Solution». *Numer. Math.*, **9**, 139–148.

[14] **Gunst, R.F.** y **Mason, R.L.** (1977). «Advantages of Examining Collinearities in Regression Analysis». *Biometrics*, **33**, 249–260.

[15] **Johnston, J.** (1984). *Econometric Methods*. 3ª Ed. New York. MacGraw-Hill.

[16] **Jolliffe, I.T.** (1982). «A note on the Use of Principal Components in Regression». *Appl. Statist.*, **31**, 300–303.

[17] **Kendall, M.G.** (1957). *A Course in Multivariate Analysis*. Griffin. London.

[18] **Kloeck, T.** y **Mennes, L.B.M.** (1960). «Simultaneous Equations Estimation Based on Principal Components of Predeter mined Variables». *Econometrika*, **28**, 45–61.

[19] **Lawson, C.R.** y **Hanson, R.J.** (1974). *Solving Least-Squares Problems*. Prentice Hall. Englewood Cliffs, N.J.

[20] **Mandel, J.** (1982). «Use of Singular Decomposition in Regression Analysis». *Amer. Stat.*, **36**, 15–24.

[21] **Marquardt, D.W.** (1970). «Generalized Inverses, Ridge Regression, Biased Linear Estimation and Non-linear Estimation». *Technometrics*, **12**, 591–612.

[22] **Myers, R.H.** (1990). *Classical and Modern Regression with Applications*. Belmont. Duxbury Press.

[23] **Silvey, S.D.** (1969). «Multicollinearity and Imprecise Estimation». *J. of the Royal Stat. Soc.*, **Series B 31**, 539–552.

[24] **Stewart, G.W.** (1973). *Introduction to Matrix Computations*. Academic Press. New York.

[25] **Stewart, G.W.** (1987). «Collinearity and Least Squares Regression». *Statistical Science*, **1**, 68–100.

[26] **Turing, A.M.** (1948). «Rounding-off Errors in Matrix Processes». *Quart. J. Mech. Appl. Math.*, **1**, 287–308.

[27] **Wilkinson, J.H.** (1965). *The Algebraic Eigenvalue Problem*. Oxford University Press.

ENGLISH SUMMARY

SENSIBILITY ANALYSIS OF THE PROPER STRUCTURE IN DETECTING COLLINEARITY

ESTARELLES, R.*

PRIETO, J.*

Universidad de Valencia

The collinearity diagnosis has already been widely treated and due to this many strategies have been suggested. However, many proposals have been critized because of their drawbacks, given that they can neither determine statistical harm, nor identify properly the variables that cause the conditioning, among other things. The aim of this paper is to point out the advantages of using indexes based on the eigen-system of X'X, instead of some of the strategies more usually applied. This objective has been verified by means of an empirical research.

Keywords: Collinearity Diagnostic, The Variance Inflation Factor, Singular Decomposition, The Condition Number, Regression-Coefficient Variance Decomposition.

*Estarelles, R. y Prieto, J. Metodología de las Ciencias del Comportamiento. Facultad de Psicología. Universidad de Valencia.

-Received january 1996.

-Accepted december 1996.

Secció docent i problemes

SECCIÓ DOCENT I PROBLEMES

La "SECCIÓ DOCENT I PROBLEMES" a la revista QÜESTIIÓ té l'objectiu d'incloure una secció on es publiquen articles de caire docent, difícilment publicables en revistes de recerca. A cada número de QÜESTIIÓ s'inclourà d'un a tres problemes i les solucions es donaran en el número següent.

Els lectors poden, si ho volen, proposar problemes amb les solucions pertinents i enviar-los a QÜESTIIÓ, que farà una selecció i en publicarà els més adequats, fent la corresponent referència a l'autor.

També seran ben rebudes solucions alternatives a les propostes fetes per l'autor dels problemes; l'editorial es reservarà, però, el dret a publicar-les.

PROBLEMES PROPOSATS

PROBLEMA N° 62

Siguin X, Y dues variables aleatòries $N(0,1)$ independents. Considera el canvi de variables

$$X = R \cdot \cos\theta \qquad Y = R \cdot \sin\theta$$

1) Demostra que R i θ són estocàsticament independents.

2) Dóna una demostració, exclusivament geomètrica, de que la variable suma $X + Y$ és també normal.

C.M. Cuadras

Universitat de Barcelona

PROBLEMA N° 63

Considerem una mostra aleatòria simple $x_1, x_2, \ldots, x_n$ d'una variable aleatòria amb densitat $f(x,\theta)$. Suposem que volem contrastar

$$H_0: \quad \theta \in \omega \qquad H_1: \quad \theta \in \Omega - \omega$$

on $\omega \subset \Omega$ són regions paramètriques. Sigui

$$\lambda = \max_{\theta \in \omega} L(x_1, \ldots, x_n; \theta) / \max_{\theta \in \Omega} L(x_1, \ldots, x_n; \theta)$$

la raó de versemblança i suposem que es compleixen les condicions de regularitat de manera que la distribució asimptòtica de $V = -2\log\lambda$ és χ^2_{2k} sota H_0, on $2k > 0$.

1) Demostrar que, asimptòticament, la distribució de λ és la mateixa que la del producte

$$U = \prod_{i=1}^{k} U_i$$

on cada U_i és uniforme (0,1) i les U_i són independents.

2) Suposem $k = 1$. Especifica les regions d'acceptació de H_0 i de H_1, expressades com $\lambda \geqslant c$ i $\lambda < c$, respectivament.

C.M. Cuadras

Universitat de Barcelona

SOLUCIONS ALS PROBLEMES PROPOSATS AL VOLUM 20. N° 2

PROBLEMA N° 60

El enunciado del problema aparece en el libro de Barnett (1982, p. 139). Para resolverlo, reescribimos el estimador r_u del modo

$$r_u = \frac{1}{n}\sum_{i=1}^{N}\frac{x_i}{y_i}e_i,$$

donde e_i se distribuye binomial $B(n, y_i/Y)$, y por tanto

$$E(e_i) = ny_i/Y, \quad \text{y} \quad V(e_i) = (ny_i/Y)(1 - y_i/Y).$$

a) $$E(r_u) = \frac{1}{n}\sum_{i=1}^{N}\frac{x_i}{y_i}E(e_i) = \frac{X}{Y} = R,$$

y por ello, r_u es insesgado para R.

b) $$V(r_u) = \frac{1}{n^2}\left[\sum_{i=1}^{N}\frac{x_i^2}{y_i^2}V(e_i) + \sum_{i=1}^{N}\sum_{j\neq i}^{N}\frac{x_i x_j}{y_i y_j}\operatorname{Cov}(e_i, e_j)\right] =$$

$$= \frac{1}{n^2}\left[\sum_{i=1}^{N}\frac{x_i^2}{y_i^2}\frac{ny_i}{Y}\left(1 - \frac{y_i}{Y}\right) + \sum_{i=1}^{N}\sum_{j\neq i}^{N}\frac{x_i x_j}{y_i y_j}\left(-n\frac{y_i}{Y}\frac{y_j}{Y}\right)\right] =$$

$$= \frac{1}{n}\left[\sum_{i=1}^{N}\frac{x_i^2}{y_i}\frac{1}{Y} - \sum_{i=1}^{N}\frac{x_i^2}{Y^2} - \sum_{i=1}^{N}\sum_{j\neq i}^{N}\frac{x_i x_j}{Y^2}\right] =$$

$$= \frac{1}{n}\left[\frac{1}{n}\sum_{i=1}^{N}\frac{x_i^2}{y_i^2}\frac{ny_i}{Y} - \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{x_i x_j}{Y^2}\right] =$$

$$= \frac{1}{n}\left\{\left[\frac{1}{n}\sum_{i=1}^{N} r_i^2 E(e_i)\right] - \left[\frac{1}{n}\sum_{i=1}^{N} r_i E(e_i)\right]^2\right\} =$$

$$= \frac{1}{n}\left[\frac{1}{n}\sum_{i=1}^{N}(r_i - R)^2 E(e_i)\right] = \frac{1}{n}\sum_{i=1}^{N}(r_i - R)^2\frac{y_i}{Y}$$

$$c)\ \mathrm{E}[s^2(r_u)] = \frac{1}{n(n-1)}\,\mathrm{E}\left[\sum_{i=1}^{n}(r_i - r_u)^2\right] =$$

$$= \frac{1}{n(n-1)}\,\mathrm{E}\left[\sum_{i=1}^{N}(r_i^2 - 2r_i r_u + r_u^2)\right] =$$

$$= \frac{1}{n(n-1)}\,\mathrm{E}\left[\sum_{i=1}^{N} r_i^2 e_i - 2\sum_{i=1}^{n} r_i\left(\frac{1}{n}\sum_{j=1}^{N} r_j e_j\right) + n r_u^2\right] =$$

$$= \frac{1}{n(n-1)}\left[\sum_{i=1}^{N} r_i^2 \frac{n y_i}{Y} - 2\mathrm{E}\left(\sum_{i=1}^{n} r_i r_u\right) + n\mathrm{E}(r_u^2)\right] = (*)$$

Pero,

$$\mathrm{E}(r_i r_u) = \mathrm{Cov}(r_i, r_u) + \mathrm{E}(r_i)\mathrm{E}(r_u) = \frac{1}{n}\mathrm{V}(r_i) + R^2 = \mathrm{V}(r_u) + R^2,$$

luego,

$$\mathrm{E}\left(\sum_{i=1}^{n} r_i\, r_u\right) = n\left[\mathrm{V}(r_u) + R^2\right],$$

de donde,

$$(*) = \frac{1}{n-1}\left[\sum_{i=1}^{N} r_i^2 \frac{y_i}{Y} - 2R^2 - 2\mathrm{V}(r_u) + \mathrm{V}(r_u) + R^2\right] =$$

$$= \frac{1}{n-1}\left[\frac{1}{n}\sum_{i=1}^{N} r_i^2\,\mathrm{E}(e_i) - R^2 - \mathrm{V}(r_u)\right] =$$

$$= \frac{1}{n-1}\left[\frac{1}{n}\sum_{i=1}^{N}(r_i - R)^2\,\mathrm{E}(e_i) - \mathrm{V}(r_u)\right] =$$

$$= \frac{1}{n-1}\left[n\mathrm{V}(r_u) - \mathrm{V}(r_u)\right] = \frac{n-1}{n-1}\mathrm{V}(r_u) = \mathrm{V}(r_u)$$

Referencia

Barnett, V. (1982). *Elements of Sampling Theory*. Hodder and Stoughton. Londres.

M. Ruiz Espejo
U.N.E.D.

PROBLEMA N° 61

a) Llamamos $p_{i|j}$ a la probabilidad de seleccionar en la segunda etapa a la unidad i, condicionado a que en la primera etapa se seleccionó la unidad j. Por ser esquema IPPS $(n = 2)$,

$$\pi_i = \sum_{j \neq i}^{N} \left(p_i p_{j|i} + p_j p_{i|j}\right) + p_i p_{i|i} =$$

$$= p_i \sum_{j=1}^{N} p_{j|i} + \sum_{j \neq i}^{N} p_j p_{i|j} = p_i + \sum_{j \neq i}^{N} p_j p_{i|j} \tag{1}$$

Imponiendo que el esquema es IPPS $(n = 2)$, tenemos también que

$$\pi_i = 2 p_i$$

De (1) y (2) deducimos que

$$\sum_{j \neq i}^{N} p_j p_{i|j} = p_i \qquad (i = 1,2,\ldots,N) \tag{2}$$

por lo que sumando en i, tenemos

$$\sum_{i=1}^{N} \sum_{j \neq i}^{N} p_j p_{i|j} = \sum_{i=1}^{N} p_i = 1, \tag{3}$$

por ser los p_i normalizados, lo que implica que

$$\sum_{i=1}^{N} p_i p_{i|i} = 0,$$

y al ser $p_i > 0$ para todo $i = 1,2,\ldots,N$; deducimos que

$$p_{i|i} = 0 \qquad \text{para todo} \quad i = 1,2,\ldots,N.$$

Es decir, el esquema IPPS ($n = 2$) es sin reemplazamiento. Además, se dará

$$\sum_{i=1}^{N} p_{i|j} = 1 \qquad (j = 1,2,\ldots,N). \tag{4}$$

b) Como consecuencia, nos queda ahora el sistema simplificado, de (3) y (4),

$$\sum_{j \neq i}^{N} p_j p_{i|j} = p_i \qquad (i = 1,2,\ldots,N) \tag{5}$$

y

$$\sum_{i \neq j}^{N} p_{i|j} = 1 \qquad (j = 1,2,\ldots,N) \tag{6}$$

Si los p_i están fijados, el sistema (5) y (6) tiene $2N$ ecuaciones con $N(N-1)$ incógnitas. Como $N(N-1) \geqslant 2N$ (si $N > 2$), con igualdad si y sólo si $N = 3$, deducimos que sólo en el caso $N = 3$ habrá solución única para el esquema muestral, aunque para $N \geqslant 4$ puede haber más de una solución, como lo prueban los esquemas de Rao (1965) y Durbin (1967).

El esquema único IPPS ($n = 2$) para $N = 3$, se obtiene reescribiendo las ecuaciones (5) y (6). El sistema resultante es

$$\left.\begin{aligned} p_2 p_{1|2} + p_3 p_{1|3} &= p_1 \\ p_1 p_{2|1} + p_3 p_{2|3} &= p_2 \\ p_1 p_{3|1} + p_2 p_{3|2} &= p_3 \\ p_{2|1} + p_{3|1} &= 1 \\ p_{1|2} + p_{3|2} &= 1 \\ p_{1|3} + p_{2|3} &= 1 \end{aligned}\right\} \tag{7}$$

que resolviéndolo tenemos por solución única,

(8) $$p_{2|1} = (p_2 - p_3)/p_1 \quad (\geqslant 0)$$

(9) $$p_{1|2} = p_1/p_2 \quad (\leqslant 1)$$

(10) $$p_{3|1} = (p_1 - p_2 + p_3)/p_1 \quad (\geqslant 0)$$

$$p_{1|3} = 0$$
$$p_{2|3} = 1$$
$$p_{3|2} = 1 - p_1/p_2.$$

De (8), (9) y (10) podemos asegurar que las condiciones necesarias y suficientes para que exista solución única posible del sistema (7) con $N = 3$, y tengan sentido probabilístico son:

(1.a) $$p_1 \leqslant p_2.$$

(1.b) $$p_3 \leqslant p_2 \qquad \text{y}$$

(1.c) $$p_2 \leqslant p_1 + p_3.$$

La demostración es inmediata por lo expuesto previamente. Ahora queda comprobar que $\pi_i = 2p_i \quad (i = 1,2,3)$. En efecto,

$$\pi_1 = p_1 + p_2 p_{1|2} + p_3 p_{1|3} = p_1 + p_2 \frac{p_1}{p_2} + p_3 \cdot 0 = 2p_1,$$

$$\pi_2 = p_2 + p_1 p_{2|1} + p_3 p_{2|3} = p_2 + p_1 \frac{p_2 - p_3}{p_1} + p_3 = 2p_2, \qquad \text{y}$$

$$\pi_3 = p_3 + p_1 p_{3|1} + p_2 p_{3|2} = p_3 + p_1 \frac{p_1 - p_2 + p_3}{p_1} + p_2 \frac{p_2 - p_1}{p_2} = 2p_3,$$

luego efectivamente, estamos ante un esquema IPPS ($n = 2$). Además es fácil comprobar que para este esquema,

$$\pi_{12} = p_1 + p_2 - p_3$$
$$\pi_{13} = p_1 - p_2 + p_3 \qquad \text{y}$$
$$\pi_{23} = -p_1 + p_2 + p_3.$$

Si $N \geqslant 4$, las soluciones $\{p_{i|j}\}_{1 \leqslant i \neq j \leqslant N}$ pueden no ser únicas y admitir, como ya hemos indicado, otros esquemas muestrales IPPS ($n = 2$). La mera comprobación de que el esquema muestral más simple estudiado aquí no está contenido en los esquemas de Rao (1965) o Durbin (1967), se dejan como ejercicio para el lector.

Referencia

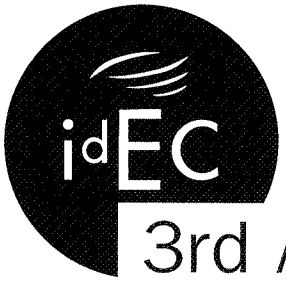
Durbin, J. (1967). «Design of multistage surveys for the estimating of sampling errors». *Appl. Statist.*, **16**, 152–164.

Rao, J.N.K. (1965). «On two simple schemes of unequal probability sampling without replacement». *J. Indian Statist. Assoc.*, **3**, 173–180.

M. Ruiz Espejo

U.N.E.D.

Cursos i congressos d'estadística



3rd Applied Statistics Week

Short Courses

Design and Analysis of Survey Data

Barcelona, 27 June to 3 July 1997

INSTITUT D'EDUCACIÓ CONTÍNUA

IN COOPERATION WITH: Institut d'Estadística de Catalunya

COURSE FEES

The course fees include course notes, a Diploma of Institut d'Educació Contínua, meals and refreshments during the courses.

Course 1: 60.000 ptas **(40.000 ptas. if paid before the 16th of May 1997)**
Course 2: 75.000 ptas **(50.000 ptas. if paid before the 16th of May 1997)**
Course 3: 75.000 ptas **(50.000 ptas. if paid before the 16th of May 1997)**

For those who register for 2 courses the fee has a reduction of 10%
For those who register for 3 courses the fee has a reduction of 20%

SPECIAL FEES for doctoral students and participants in previous years of the APPLIED STATISTIC WEEK:

Course 1: 30.000 ptas
Course 2: 37.500 ptas
Course 3: 37.500 ptas

Registration

The registration form must be sent by post or by fax to the IDEC. Payment can be made by bank transfer (a copy of the bank transfer should be sent with the registration form), by credit card (VISA) or by bank cheque to the Institut d'Educació Contínua. In the case of foreign payments, the cheque should be in convertible pesetas.

Accommodation facilities

The Continuing Education Institute has arranged a special price of 8.500 ptas. with some hotels in Barcelona for the accommodation of participants:

Astoria ***, address: Paris, 203 (300m from IDEC)
Balmes ***, address: Mallorca, 216 (200m from IDEC)

In all cases, prices are for a double room for either one or two persons (VAT included).

Information and enrolment

IDEC (Continuing Education Institute)
Mrs. Maria Torras
Balmes, 132
08008 Barcelona - Spain
Telephone: (34-3) 542 18 00/03
Fax: (34-3) 542 18 08
E-Mail: idec@upf.es

COURSE 1.

DATA COLLECTION IN MARKET AND OPINION RESEARCH

Willem Saris

In market and opinion research data collection is often seen as a simple task since everybody can formulate questions. In this course we will illustrate by very simple examples that this task is not so easy. Large differences in results can be obtained depending on the wording of the questions, the way the data are collected (face-to-face, telephone or computer-assisted). Also the response burden has a large effect on the results. In this course procedures will be discussed: (i) to design research instruments in a scientific way, (ii) to evaluate the quality of instruments, and (iii) to take advantage of the recent trends towards automation of data collection through computer-assisted interviewing.

Willem Saris is Professor in the Department of Methods and Techniques for Political Science at the University of Amsterdam. He is author of the book «Computer-Assisted Interviewing» in the Sage University Series, and his present research interests include structural equation modelling and the improvement of measurement in social science research.

Date: 27 June, from 9.30 to 13.30 and from 15:00 to 18.00, and 28 June, from 9.30 to 13.30 hours

COURSE 2.

DESIGN AND ANALYSIS OF SAMPLE SURVEYS

Kirk Wolter

The purpose of this course is to provide training in both the design and the analysis of sample surveys. All training will focus on the probability-sampling approach to survey design, which is the most rigorous and scientifically accepted approach for surveys that demand high precision, receive close scrutiny, and form the basis for important business or policy decisions. The material covered will include basic methods of sample selection (simple random sampling, systematic sampling, and probability-proportional-to-size sampling), stratification, multiple-stage sampling, basic methods of estimation, handling of nonresponse or missing data, and sampling for telephone surveys. The course will also cover statistical modelling based upon survey data and variance estimation for complex surveys, including methods for means, proportions, ratios, regression coefficients, and statistics arising in the analysis of two-way contingency tables. The use of the program PC CARP for computing statistics from complex sample designs will be demonstrated, and instruction will be given in the use of the program.

Kirk Wolter is Professor in the Department of Statistics at the University of Chicago and is Senior Vice President, Statistics and Methodology at the National Opinion Research

Center, a company that specializes in the design, conduct and analysis of large, complex surveys. He is author of the book «Introduction to Variance Estimation» and numerous journal articles.

Date: 30 June and 1 July, from 9.30 to 13.30 and from 15.00 to 18.00 hours

COURSE 3.

MODELLING AND INTERPRETATION OF SURVEY DATA
Jacques Hagenaars

The analysis of survey data goes further than simple summaries of frequencies and crosstabulations. There is a wealth of models available for exploring causal and non-causal relationships among sets of intercorrelated variables as well as testing specific hypotheses about relationships and trends in the surveyed population. In this course specific attention will be given to the application of loglinear models and latent class analysis to survey data. Latent class models enable the survey researcher to take into account many kinds of response errors, such as unreliability, test-retest effects, item non-response, and different survey conditions. Longitudinal surveys, where a panel of respondents is followed over time, will also be discussed from the modelling perspective, as well as aspects of causal modelling. Participants will be introduced to the use of loglinear and latent class software and the interpretation of results. Emphasis will be on the critical use of these techniques, rather than on technical details.

Jacques Hagenaars is Professor of Social Science Methodology and Chair of the Department of Methodology at the Faculty of Social Sciences, Tilburg University. He is the author of the Sage books «Categorical Longitudinal Data: Loglinear Panel, Trend and Cohort Analysis» and «Loglinear Models with Latent Variables».

Date: 2-3 July, from 9.30 to 13.30 and from 15.00 to 18.00 hours

Place

The classes will be held at IDEC (Continuing Education Institute, Pompeu Fabra University) - Balmes, 132. 08008 Barcelona

Language

All the courses will be taught in English

Programme directors

Michael J. Greenacre
Albert Satorra
Pompeu Fabra University, Barcelona

5Pa

NGUS '97

IV International Meeting of Multidimensional Data Analysis

New Spad Users Group

IV Congrès International D'Analyses Multidimensionnelles des Données

Nouveau Group des Utilisateurs de Spad

Bilbao (SPAIN)
September 10-11-12, 1997

Dpto de Economía Aplicada III
(Econometría y Estadística)
Facultad de CCEE y Empresariales
Universidad del País Vasco /
Euskal Herriko Unibertsitatea

eman ta zabal zazu

U.P.V. E.H.U.

Location / *Lieu:*

Facultad de CCEE y Empresariales
Avda. Lehendakari Aguirre, 83
48015 Bilbao - Spain

Scientific programme committee / *Comité de programme scientifique:*

Tomás Aluja (Spain)
Mónica Bécue (Spain)
Carles Cuadras (Spain)
Karmele Fernández (Spain)
Angeles Iztueta (Spain)
Carlo Lauro (Italy)
Ludovic Lebart (France)
Francesco Mola (Italy)
Alain Morineau (Président) (France)
Jean Pierre Nakache (France)
Jérôme Pagès (France)
Robin Sibson (UK)
Fernando Tusell (Spain)
Amaya Zárraga (Spain)

Organizing commitee / *Comité d´organisation:*

Karmele Fernández (Presidente) (Spain)
Miguel A. García (Spain)
Angeles Iztueta (Spain)
M. Isabel Landaluce (Spain)
M. José Mijangos (Spain)
Amaya Zárraga (Spain)

Purpose / *Objectif:*

To provide an opportunity for researchers involved in statistical analysis of multidimensional data to meet and present their work to colleagues/*Fournir un forum pour que les chercheurs dans le domaine de l´analyse de données multivariées puissent présenter leurs travaux*

5-82

Conferenciers / *Lecteurs:*

Tomás Aluja (Spain)
Mónica Bécue (Spain)
Carles Cuadras (Spain)
Carlo Lauro (Italy)
Ludovic Lebart (France)
Francesco Mola (Italy)
Alain Morineau (France)
Jean Pierre Nakache (France)
Jérôme Pagès (France)
Robin Sibson (UK)

Call for papers / *Présentation des originaux:*

All official languages of the EC may be used for discussion, but no simultaneous translation will be provided / *Toutes les langues de la CE sont admises, mais il n´y aura pas de traduction simultanée*

A Summary of the work, 4 pages maximun (bibliography, graphics, tables... included), should be sent to Karmele Fernández before 15 April 1997. We have made a model in LATEX for the later reprography and we can supply it at the Internet address below or you can request it / *Un résumé de la communication comportant 1 à 4 pages, bibliographie, cadres et graphiques compri, doit être expédié à Karmele Fernández au plus tard le 15 Avril 1997. Un exemple de résumé utilisant le format LATEX a été preparé, pour la reprographie, et on peut se le procurer à l´adresse Internet où vous pouvez le demander:*

http://www.et.bs.ehu.es/ngus97.html
Fax: 34 4 4797554
email: NGUS@bs.ehu.es

Collaborator organizations / *Organismes collaborateurs:*

UPV/EHU (Universidad del País Vasco)
EUSTAT (Euskal Estatistika Erakundea)
BBK (Bilbao Bizkaia Kutxa)
CIDIR (Centro de Informática para la Docencia, Investigación, Red...)
CISIA (Centre International de Statistique et d'Informatique Appliquées)

Transport and accomodation / *Transport et logement:*

The participants are invited to contact directly to our official travel agency/ *Les participants sont invités a contacter directment notre agence de voyages:*

Viajes Ecuador (*)
Sarriena Auzoa s/n
Campus de Leioa
Tfn:34 4 4633194 - 34 4 4633165
Fax: 34 4 4630194
48940 Leioa (Vizcaya)- SPAIN

(*) Reduced tariff for NGUS'97/ *Tarif réduit pour NGUS'97*

VI CONFERENCIA ESPAÑOLA DE BIOMETRÍA

Córdoba, 21-24 setiembre 1997

Organizada por la Región Española de la International Biometric Society, la VI Conferencia Española de Biometría se celebrará en la Escuela Técnica Superior de Ingenieros Agrónomos y Montes de la Universidad de Córdoba, los días 21, 22, 23 y 24 de setiembre de 1997, en sesiones de comunicaciones, pósters, conferencias, mesas redondas y demostraciones de logicial estadístico.

Las normas de redacción de las comunicaciones, así como plantillas en LaTeXy Word para su realización, se encuentran en la página web de Internet: http://www.uco.es/~biometry/

Calendario (plazos)

- *Inscripción:* 15 de mayo (después de esta fecha se aplicará un recargo del 20%).
- *Recepción de trabajos:* 1 de junio.
- *Resúmenes de pósters y demostraciones de programas:* 15 de julio.
- *Comunicación de la aceptación:* 30 de julio.
- *Recepción de los congresistas:* domingo 21 de setiembre.
- *Sesiones:* 22, 23 y 24 de setiembre.
- *Clausura:* 24 de setiembre.

Cuotas de inscripción y formas de pago

Ordinaria	22.000 Pta
Miembros de la Sociedad de Biometría	16.000 Pta
Estudiantes	5.000 Pta

- Cheque bancario al Departamento de Estadística, Universidad de Córdoba, Apdo. de Correos 3048, 14080 Córdoba.
- Transferencia a la cta. 0030-4001-93-0105145-271 de Banesto, Avda. Gran Capitán, 20, 14001 Córdoba.

Reserva de hoteles

Viajes Intercontinental «Congreso de Biometría»
Manuel de Sandoval, 1
14008 Córdoba
Fax +34 (9) 57-470144
Tel. +34 (9) 57-470694

Información

VI Conferencia Española de Biometría
Departamento de Estadística. ETSIAM
Apdo. de Correos 3048
14080 Córdoba

Correo electrónico: biometry@lucano.uco.es
Web Internet: http://www.uco.es/~biometry/
Fax +34 (9) 57-218481
Tel. +34 (9) 57-218568

Butlleta de subscripció a la revista **Qüestiió**

Nom i cognoms _____

_____

Empresa/Institució _____

_____

Adreça _____

_____

Codi postal _____ ciutat _____

Tel. _____ Fax _____

Data d'expedició _____

Signatura: DNI o NIF _____

Desitjo subscriure'm a **Qüestiió** per a l'any 1997.
El preu de la subscripció és de 3.000 PTA.

Forma de pagament

☐ Transferència al compte de la Caixa de Catalunya número 6985/77,
Agència 100, Comte d'Urgell 162, 08036 Barcelona

☐ Domiciliació bancària

☐ Taló nominatiu a l'Institut d'Estadística de Catalunya

☐ Gir postal

☐ En efectiu

Retornar aquesta butlleta (o una fotocòpia) a:

Qüestiió:
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona

Autorització de domiciliació bancària per al pagament de les subscripcions anuals de la revista **Qüestiió**

El sotasignat _____

autoritza el Banc/Caixa _____

Agència núm. _____ adreça _____

Codi postal _____ ciutat _____

a abonar a la revista **Qüestiió** amb càrrec al meu compte número _____

_____, les subscripcions a **Qüestiió**.

_____ , a _____ d _____ de 19 _____

(Signatura)

El sotasignat _____

autoritza la revista **Qüestiió** a carregar al meu compte número _____

_____ , al Banc/Caixa _____

_____

Agència núm. _____ adreça _____

Codi postal _____ ciutat _____

l'import de les tarifes vigents de les subscripcions a la revista **Qüestiió.**

_____ , a _____ d _____ de 19 _____

(Signatura)