Qüestiió

Quaderns d'Estadística i Investigació Operativa

Any 1999, volum 23, núm. 2
Segona època

Entitats patrocinadores:

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

Entitat col·laboradora:

International Biometric Society

Generalitat de Catalunya
Institut d'Estadística de Catalunya

Any 1999, volum 23, núm. 2

SUMARI

EDITORIAL

El segon número del volum 23 (1999) aplega set articles distribuïts en tres de les quatre seccions habituals. D'entre els treballs publicats cal ressaltar el fet que, per primera vegada en la nova etapa de *Qüestiió*, hi apareix la traducció al català en forma d'article d'una conferència que el 1997 va pronunciar Ivan P. Fellegi, director d'Statistics Canada (considerat com l'institut d'estadística nacional més prestigiós en els darrers tres anys). Al marge de l'interès intrínsec del seu contingut, la confiança que ens ha atorgat el seu autor en publicar-ho abans de la seva versió original en anglès, suposa un significatiu reconeixement a la trajectòria de *Qüestiió* i també un estímul per continuar publicant textos d'aquesta naturalesa.

El conjunt d'articles inclosos en aquest número s'afegeix als vuit originals publicats a l'anterior, de manera que, previsiblement, el nombre d'articles i pàgines publicades del volum d'enguany superarà la mitjana enregistrada per al conjunt de la segona època de *Qüestiió*. Juntament amb la millora qualitativa i quantitativa de la versió impresa, la revista també experimenta una projecció creixent en l'àmbit de la difusió electrònica: més de 6.700 consultes al web en el primer semestre del 1999, que han crescut ininterrompudament des de les 570 del mes de gener fins a les quasi 2.000 peticions d'accessos el mes de juny.

Paralel·lament, els recents desenvolupaments en l'àmbit institucional també han ajudat a potenciar la consulta d'articles, l'elaboració d'originals i la mateixa gestió administrativa mitjançant les pàgines del web que es va crear el maig de 1998; en aquest sentit, la configuració del web dedicat a *Qüestiió* és un bon exponent d'aquest dinamisme. Així, la informació accessible inicialment es limitava a la consulta dels articles per autors i per matèries, un resum de les normes de presentació d'originals, la descripció dels òrgans de govern de la revista i els avaluadors en el període 1987-97. Actualment, el domini de *Qüestiió* inclou, a més, una referenciació cronològica dels articles publicats fins el volum 22 (que es preveu actualitzar abans de final d'any amb el recull d'articles del volum 23), l'accés a les editorials-comentaris de cada número o volum, l'avanç del sumari del proper número, la consulta dels processos d'avaluació dels originals sotmesos a les seccions temàtiques, l'ampliació de les normes de presentació i l'enllaç amb la classificació AMS adoptada per *Qüestiió* en l'adscripció temàtica dels originals.

C.M. Cuadras i E. Ripoll, editors executius

Comentari de les seccions «Estadística» i «Biometria»

La secció «Estadística» agrupa dos articles. El primer d'ells, *Some results envolving the concepts of moments generating function and affinity between distribution functions. Extension for r k-dimensional normal distribution functions*, d'A. Dorival, presenta una extensió de l'afinitat entre funcions de densitat fent intervenir funcions generatrius de moments, la qual conté l'afinitat de Matusita i la funció generatriu com a casos particulars; d'altra banda, l'article presenta expressions concretes per a la distribució normal multivariant. El segon original, *Efecto del número de indicadores por factor sobre la identificación y estimación en modelos aditivos de AFC*, d'A. Olivé, J.M. Tomás i P.M. Hontangas, utilitza l'anàlisi factorial confirmatòria a fi d'estudiar les matrius «multirasgo-multimétodo» i evidencia que alguns dels mètodes clàssics en aquest àmbit mostren un comportament anòmal; el treball conclou amb la proposta de vies alternatives.

L'únic article que recull la secció «Biometria», *Estudios de supervivencia con datos no observados. Dificultades inherentes al enfoque paramétrico*, de G. Gómez i C. Serrat, presenta un enfocament paramètric dels models de supervivència amb dades faltants, fent una valoració de les dificultats pràctiques i del fonament d'aquests models que els autors il·lustren amb una aplicació.

Carles M. Cuadras, editor executiu

Comentari de la secció «Estadística Oficial» i altres apartats

En aquesta ocasió, la secció «Estadística Oficial» aplega un total de quatre articles força diferenciats. En primer lloc, tal com s'ha comentat anteriorment, s'ofereix la versió catalana de *Serveis Estadístics. Preparant-se per al futur*, ponència impartida per I. Fellegi a la «1997 UK Statistics Users Conference»; l'actual Director d'Statistics Canada fa una exhaustiva i profunda reflexió sobre les estratègies internes i externes que les oficines d'estadística han d'encarar a fi de preservar la seva pertinència. La perllongada experiència i la contrastada vàlua professional de l'autor en la producció i difusió d'informació estadística oficial, li permeten avaluar riscs i suggerir oportunitats en l'optimització dels actuals sistemes d'informació. En segon lloc, l'article de M.J. Bárcena i F. Tusell, *Enlace de encuestas: una propuesta metodológica y aplicación a la encuesta de presupuestos de tiempo*, incideix en la imputació de dades faltants mitjançant el recurs de la concatenació d'enquestes; la utilització combinada de les tècniques d'arbres de regressió i classificació es revela com una nova proposta efectiva i s'aplica amb èxit a una enquesta oficial de l'Institut Basc d'Estadística, marc d'activitat en què sovint es presenta aquesta problemàtica. Finalment, la sec-

ció inclou els dos primers articles que inauguren la publicació en successius números de *Qüestiió* d'una selecció de les ponències presentades a les jornades internacionals «Generació d'informació estadística: qualitat i limitacions 1998» organitzades per la Xarxa Temàtica «Enquestes i Qualitat de la Informació Estadística». En el treball de L. Lebart, *Qualité de l'information dans les enquêtes*, s'analitza la contribució de les variables de control numèriques, nominals o textuals, vinculades a l'execució d'enquestes a la millora de la seva concepció i de la mateixa operatòria; en aquest context, les tècniques de Data Mining i Text Mining ofereixen recursos adients per a grans volums de dades o de múltiples dimensions, resituant els conceptes de (mesura de la) qualitat de la informació recollida i processada. Pel que fa al *Diseño de encuestas de opinión: el barómetro CIS*, de V. Martínez, es ressalten els efectes de les baixes taxes de resposta o de la seva manca absoluta en l'àmbit de les enquestes d'opinió: el risc de biaix en les estimacions o l'atenció que requereix la selecció no probabilística de la mostra formen part d'una àmplia i precisa descripció del «baròmetro CIS», el sondeig més popular que mensualment segueixen estudiosos i públic en general al nostre país.

Seguidament, la «Secció docent i problemes» continua amb la presentació successiva d'enunciats de problemes i la resolució dels publicats en el número immediatament anterior.

Per últim, l'apartat «Ressenyes d'activitats institucionals» inclou, com ja és costum, una actualització d'activitats de la Sociedad Española de Biometría i dels cursos organitzats en col·laboració amb altres entitats. En segon lloc, una nova recensió del «TES Institute: Training of European Statisticians», amb la ressenya dels cursos del *Core Programme 1999-2000* que inclou el curs *Systems of Social Statistics* que acull l'Idescat del 4 al 8 d'octubre; com la majoria d'activitats del TES Institute que *Qüestiió* redifon des del volum 21 (1997), el seu programa s'adreça als membres d'instituts d'estadística europeus. Seguidament, es reprodueix l'anunci del «Tercer Congrés Europeu de Matemàtiques» –3ECM– (Barcelona, 10-14 de juliol 2000) que organitza la Societat Catalana de Matemàtiques-IEC, sota els auspicis de la European Mathematical Society, on hi col·laboren, entre d'altres organismes, la UB, la UPC i l'Idescat. En darrer lloc, s'anuncia una nova edició del «International Workshop on Statistical Modelling» –15th IWSM– (Bilbao, 17-21 de juliol 2000) que, com en altres ocasions que també ha anunciat *Qüestiió*, organitza la Universitat del País Basc.

Enric Ripoll, editor executiu

Estadística

QÜESTIIÓ, vol. 23, 2, p. 225-237, 1999

SOME RESULTS ENVOLVING THE CONCEPTS OF MOMENT GENERATING FUNCTION AND AFFINITY BETWEEN DISTRIBUTION FUNCTIONS. EXTENSION FOR r $k-$DIMENSIONAL NORMAL DISTRIBUTION FUNCTIONS

A. DORIVAL CAMPOS
Universidade de São Paulo

We present a function $\rho(F_1, F_2, t)$ which contains Matusita's affinity and express the «affinity» between moment generating functions. An interesting result is expressed through decomposition of this «affinity» $\rho(F_1, F_2, t)$ when the functions considered are $k-$dimensional normal distributions. The same decomposition remains true for others families of distribution functions. Generalizations of these results are also presented.

Keywords: Affinity, moment generating functions, distance, inner product, multivariate normal distributions, probability density functions, absolutely convergent.

AMS Classification: primary 62E99, secondary 62H20.

* Dpto. de Matemática Aplicada à Biologia. Faculdade de Medicina de Riberirão Preto. Universidade de São Paulo. Brasil.

- Received June 1998.
- Acepted March 1999.

1. INTRODUCTION AND PRELIMINARIES

Let F_1 and F_2 be two distribution functions defined on $\mathbb{R}$ and let us denote by $f_i(x)$ the probability density function of F_i with respect to a measure m in $\mathbb{R}$, for $i = 1, 2$.

We find in the literature several forms of defining distance between distributions of a same class. Matusita (1955) by making use of the distance function denoted by $d(F_1, F_2)$ and expressed by

$$d(F_1, F_2) = \left\{ \left(f_1^{1/2}(x) - f_2^{1/2}(x),\ f_1^{1/2}(x) - f_2^{1/2}(x) \right) \right\}^{1/2},$$

introduced in the statistical literature the concept of affinity between the distributions F_1 and F_2 denoted by $\rho_2(F_1, F_2)$ and defined by

$$\rho_2(F_1, F_2) = \left(f_1^{1/2}(x),\ f_2^{1/2}(x) \right)$$

which is related to $d(F_1, F_2)$ through the expression

$$d^2(F_1, F_2) = 2\{1 - \rho_2(F_1, F_2)\}$$

where (f, g) denotes the inner product of $f(x)$ and $g(x)$ defined by:

$$(f, g) = \int_{\mathbb{R}} f(x)\, g(x)\, dm$$

The importance and usefulness of the notions of distance and affinity between distributions, in statistics, were stressed in a series of papers by Matusita (1954, 1955, 1956, 1961, 1964, 1967b, 1973), Matusita & Motoo (1955), Matusita & Akaike (1956), Khan & Ali (1971) and others.

Concrete expressions for the affinity between two multivariate normal distributions were established by Matusita (1966). As a following step, Matusita (1967) extended the notion of affinity to cover the case where there are r distributions involved and established concrete expressions when the r distributions are $k-$dimensional normal.

Our work is characterized by the introduction of the concept of a function, denoted by $P(t)$, functionally expressed through the moment generating functions relative to the distributions considered and another expression denoted by $\rho(F_1, F_2, t)$ that contains as a particular case the affinity between the distribution functions F_1 and F_2, or in other words, express the «affinity» between the moment generating functions relative to F_1 and F_2. We also present a result that express the decomposition of $\rho(F_1, F_2, \underline{t})$ in a product of two factors identified as the affinity and the moment generating function when F_1 and F_2 are $k-$dimensional normal distributions. This result is extended to cover the case where there are r $k-$dimensional normal distributions.

In this same way, the concept of a more general function $D_r(s_1, \ldots, s_r; \underset{\sim}{t}_j)$ is introduced, and the results obtained through it contains those developed in this paper as those ones established by Matusita (1966, 1967a) and Campos (1978).

2. RESULTS

Definition 1. Let F_1 and F_2 be two distribution functions belonging to the same class and let $f_1(x)$ and $f_2(x)$ their respective probability density functions with respect to a measure m defined on $\mathbb{R}$. Let us suppose that there is a scalar t, $-h \leqslant t \leqslant h$ $(h > 0)$ such that the integral below, defined through the inner product, is absolutely convergent.

Now, we define:

(2.1)
$$P(t) = \left(\exp\{tx/2\}\left\{f_1^{1/2}(x) - f_2^{1/2}(x)\right\}, \exp\{tx/2\}\left\{f_1^{1/2}(x) - f_2^{1/2}(x)\right\}\right)$$

where (f, g) denotes the inner product of $f(x)$ and $g(x)$, defined by:

$$(f, g) = \int_{\mathbb{R}} f(x)\, g(x)\, dm$$

From (2.1), we obtain:

$$P(t) = M_1(t) + M_2(t) - 2\,\rho\,(F_1, F_2, t)$$

where:

$M_i(t)$ represent the moment generating function for the distribution F_i whose probability density function is $f_i(x)$, $i = 1, 2$;

and

(2.2) $$\rho\,(F_1, F_2, t) = \left(\exp\{tx/2\}\; f_1^{1/2}(x),\; \exp\{tx/2\}\; f_2^{1/2}(x)\right)$$

From (2.1) and (2.2) we verify that:

i) $P(0) = d^2\,(F_1, F_2)$,
ii) $F_1 = F_2$ implies $P(t) = 0$ for all $-h \leqslant t \leqslant h$
iii) $\rho\,(F_1, F_2, 0) = \rho_2\,(F_1, F_2)$
iv) $F_1 = F_2 = F$ implies $\rho(F, t) = M(t)$

where:

$$d^2\,(F_1, F_2) = \left(f_1^{1/2}(x) - f_2^{1/2}(x),\; f_1^{1/2}(x) - f_2^{1/2}(x)\right)$$

and $\rho_2(F_1, F_2)$ is the affinity between the distributions F_1 and F_2 as defined by Matusita (1966).

Teorema 1. Let F_1 and F_2 be $k-$dimensional nonsingular normal distributions, whose probability density functions are given by:

$$(2\pi)^{-k/2}\,|\underset{\sim}{A}|^{1/2}\,\exp\left\{-1/2(\underset{\sim}{A}^{-1}(\underset{\sim}{x}-\underset{\sim}{a}),\,\underset{\sim}{x}-\underset{\sim}{a})\right\}$$

and

$$(2\pi)^{-k/2}\,|\underset{\sim}{B}|^{-1/2}\,\exp\left\{-1/2(\underset{\sim}{B}^{-1}(\underset{\sim}{x}-\underset{\sim}{b}),\,\underset{\sim}{x}-\underset{\sim}{b})\right\},$$

respectively, where:

$\underset{\sim}{x}$ is a $k-$dimensional (column vector);
$\underset{\sim}{A}$ and $\underset{\sim}{B}$ are covariance matrices de degree K and
$\underset{\sim}{a}, \underset{\sim}{b}$ are $k-$dimensional mean (column) vectors.

In these conditions, we have:

$$\rho(F_1,\,F_2,\underset{\sim}{t}) = \rho_2(F_1,\,F_2)\cdot M_G(\underset{\sim}{t})$$

where:

$\underset{\sim}{t}$ is $k-$dimensional (column) vector and
$M_G(\underset{\sim}{t})$ is the moment generating function of a $k-$dimensional normal distribution with mean vector $\underset{\sim}{C}^{-1}(\underset{\sim}{A}^{-1}\underset{\sim}{a}+\underset{\sim}{B}^{-1}\underset{\sim}{b})$ and covariance matrix $2\,\underset{\sim}{C}^{-1}$, given by

$$M_G(\underset{\sim}{t}) = \exp\left\{(\underset{\sim}{C}^{-1}(\underset{\sim}{A}^{-1}\,\underset{\sim}{a}+\underset{\sim}{B}^{-1}\,\underset{\sim}{b}),\,\underset{\sim}{t})+(\underset{\sim}{C}^{-1}\,\underset{\sim}{t},\,\underset{\sim}{t})\right\} \tag{2.3}$$

with $\underset{\sim}{C} = \underset{\sim}{A}^{-1}+\underset{\sim}{B}^{-1}$.

Proof: From (2.2), we have:

$$\rho(F_1,\,F_2,\underset{\sim}{t}) = (2\pi)^{-k/2}|\underset{\sim}{A}\underset{\sim}{B}|^{-1/4}\int_{\mathbb{R}^k}\exp\{-1/4\,Q\}\,dx_1,\,\ldots,\,dx_k \tag{2.4}$$

where:

$$Q = (\underset{\sim}{A}^{-1}(\underset{\sim}{x}-\underset{\sim}{a}),\,\underset{\sim}{x}-\underset{\sim}{a})+(\underset{\sim}{B}^{-1}(\underset{\sim}{x}-\underset{\sim}{b}),\,\underset{\sim}{x}-\underset{\sim}{b})-4(\underset{\sim}{x},,\underset{\sim}{t})$$

By working with this algebraic sum of inner products, we obtain:

(2.5)
$$Q = ((\underline{A}^{-1} + \underline{B}^{-1})\,\underline{x}, \underline{x}) - 2(\underline{A}^{-1}\underline{a} + \underline{B}^{-1}\underline{b} + 2\,\underline{t}, \underline{x}) + (\underline{A}^{-1}\underline{a}, \underline{a}) + (\underline{B}^{-1}\,\underline{b}, \underline{b})$$

If we define the transformation:

$$\underline{y} = \underline{C}^{1/2}\,\underline{x}$$

with $\underline{C} = \underline{A}^{-1} + \underline{B}^{-1}$ and Jacobian equal to $\mod |\underline{C}|^{-1/2}$, (2.5) may be written as follows:

$$Q = (\underline{y}, \underline{y}) - 2(\underline{C}^{-1/2}(\underline{A}^{-1}\underline{a} + \underline{B}^{-1}\underline{b} + 2\underline{t}), \underline{y}) + (\underline{A}^{-1}\underline{a}, \underline{a}) + (\underline{B}^{-1}\underline{b}, \underline{b})$$

That is,

(2.6)
$$Q = Q_1 - (\underline{C}^{-1}(\underline{A}^{-1}\underline{a} + \underline{B}^{-1}\underline{b} + 2\underline{t}), \underline{A}^{-1}\underline{a} + \underline{B}^{-1}\underline{b} + 2\underline{t}) + (\underline{A}^{-1}\underline{a}, \underline{a}) + (B^{-1}\underline{b}, \underline{b})$$

where:

$$Q_1 = (\underline{y} - \underline{C}^{-1/2}(\underline{A}^{-1}\underline{a} + \underline{B}^{-1}\underline{b} + 2\underline{t}), \underline{y} - \underline{C}^{-1/2}(\underline{A}^{-1}\underline{a} + \underline{B}^{-1}\underline{b} + 2\underline{t}))$$

We have also:

$$(\underline{A}^{-1}\,\underline{a}, \underline{a}) = (\underline{A}^{-1}\underline{a},\, \underline{C}^{-1}\,\underline{A}^{-1}\,\underline{a}) + (\underline{A}^{-1}\,\underline{a},\, \underline{C}^{-1}\,\underline{B}^{-1}\,\underline{a})$$

and

$$(\underline{B}^{-1}\,\underline{b}, \underline{b}) = (\underline{B}^{-1}\underline{b},\, \underline{C}^{-1}\,\underline{A}^{-1}\,\underline{b}) + (\underline{B}^{-1}\,\underline{b},\, \underline{C}^{-1}\,\underline{B}^{-1}\,\underline{b})$$

By using these results in (2.6), we obtain, after same algebraic manipulations:

(2.7)
$$Q = Q_1 - (\underline{C}^{-1}\,\underline{B}^{-1}\,\underline{b},\, \underline{A}^{-1}\,\underline{a}) - (\underline{C}^{-1}\,\underline{A}^{-1}\,\underline{a},\, \underline{B}^{-1}\,\underline{b}) + (\underline{A}^{-1}\,\underline{a},\, \underline{C}^{-1}\,\underline{B}^{-1}\,\underline{a}) +$$
$$(\underline{B}^{-1}\underline{b}, \underline{C}^{-1}\underline{A}^{-1}\underline{b}) + (\underline{C}^{-1}(\underline{A}^{-1}a + \underline{B}^{-1}\underline{b} + 2\underline{t}), 2\underline{t}) - (2\,\underline{C}^{-1}\underline{t},\, \underline{A}^{-1}\underline{a} + \underline{B}^{-1}\underline{b})$$

From (2.7), it follows that:

$$Q = Q_1 + ((\underline{B}\,\underline{C}\,\underline{A})^{-1}\underline{b},\, \underline{b} - \underline{a}) - ((\underline{A}\,\underline{C}\,\underline{B})^{-1}\underline{a},\, \underline{b} - \underline{a}) -$$
$$- 4(\underline{C}^{-1}(\underline{A}^{-1}\underline{a} + \underline{B}^{-1}\,\underline{b}),\, \underline{t}) - 4(\underline{C}^{-1}\,\underline{t},\, \underline{t})$$

Since $\underset{\sim}{C} = \underset{\sim}{A}^{-1} + \underset{\sim}{B}^{-1}$, we have:

$$\underset{\sim}{A}\,\underset{\sim}{C}\,\underset{\sim}{B} = \underset{\sim}{B}\,\underset{\sim}{C}\,\underset{\sim}{A} = \underset{\sim}{A} + \underset{\sim}{B}$$

Or,

(2.8)
$$Q = Q_1 + ((\underset{\sim}{A} + \underset{\sim}{B})^{-1}(\underset{\sim}{b} - \underset{\sim}{a}),\, \underset{\sim}{b} - \underset{\sim}{a}) - 4(\underset{\sim}{C}^{-1}(\underset{\sim}{A}^{-1}\underset{\sim}{a} + \underset{\sim}{B}^{-1}\,\underset{\sim}{b}),\, \underset{\sim}{t}) - 4(\underset{\sim}{C}^{-1}\,\underset{\sim}{t},\, \underset{\sim}{t})$$

Using (2.8) in (2.4), we obtain:

$$\begin{aligned}\rho(F_1,\, F_2, \underset{\sim}{t}) \;=\; & (2\pi)^{-k/2}|\underset{\sim}{A}\underset{\sim}{B}|^{-1/4}\, \exp\left\{-1/4((\underset{\sim}{A} + \underset{\sim}{B})^{-1}\,(\underset{\sim}{b} - \underset{\sim}{a}),\, \underset{\sim}{b} - \underset{\sim}{a})\right\} \\ & \exp\left\{(\underset{\sim}{C}^{-1}(\underset{\sim}{A}^{-1}\,\underset{\sim}{a} + \underset{\sim}{B}^{-1}\,\underset{\sim}{b}),\, \underset{\sim}{t}) + (\underset{\sim}{C}^{-1}\,\underset{\sim}{t},\, \underset{\sim}{t})\right\} \\ & \cdot \int_{\mathbb{R}^k} \exp\left\{-\frac{1}{4}\,Q_1\right\} |\underset{\sim}{C}|^{-1/2}\, dy_1, \ldots, dy_k\end{aligned} \tag{2.9}$$

We easily verify that:

$$\int_{\mathbb{R}^k} \exp\left\{-\frac{1}{4}\,Q_1\right\} dy_1, \ldots, dy_k = 2^{k/2}\,(2\pi)^{k/2}$$

Or,

(2.10)
$$\begin{aligned}\rho(F_1,\, F_2,\, \underset{\sim}{t}) \;=\; & |\underset{\sim}{A}\underset{\sim}{B}|^{1/4} \left|\frac{1}{2}(\underset{\sim}{A} + \underset{\sim}{B})\right|^{-1/2} \exp\left\{-1/4((\underset{\sim}{A} + \underset{\sim}{B})^{-1}\,(\underset{\sim}{b} - \underset{\sim}{a}),\, \underset{\sim}{b} - \underset{\sim}{a})\right\} \cdot \\ & \cdot\; \exp\left\{(\underset{\sim}{C}^{-1}(\underset{\sim}{A}^{-1}\,\underset{\sim}{a} + \underset{\sim}{B}^{-1}\,\underset{\sim}{b}),\, \underset{\sim}{t}) + (\underset{\sim}{C}^{-1}\,\underset{\sim}{t},\, \underset{\sim}{t})\right\}\end{aligned}$$

since

$$|\underset{\sim}{C}|^{-1/2} = |\underset{\sim}{A}|^{1/2}\,|\underset{\sim}{A} + \underset{\sim}{B}|^{-1/2}\,|\underset{\sim}{B}|^{1/2}$$

It follows from theorem demonstrated by Matusita and (2.3) that (2.10) may be written as:

$$\rho(F_1,\, F_2,\, \underset{\sim}{t}) = \rho_2(F_1,\, F_2)\, M_G(\underset{\sim}{t})$$

Corolary 1. When $\underset{\sim}{A} = \underset{\sim}{B}$ it follows that

$$\rho(F_1,\, F_2, \underset{\sim}{t}) = \exp\left\{-1/8(\underset{\sim}{A}^{-1}(\underset{\sim}{b} - \underset{\sim}{a}), \underset{\sim}{b} - \underset{\sim}{a})\right\}\, M_G(\underset{\sim}{t})$$

where:

$$M_G(\underset{\sim}{t}) = \exp\left\{(1/2(\underset{\sim}{a}+\underset{\sim}{b}),\, \underset{\sim}{t}) + 1/2(\underset{\sim}{A}\,\underset{\sim}{t},\, \underset{\sim}{t})\right\}$$

Corolary 2. If $\underset{\sim}{a} = \underset{\sim}{b}$ it follows that:

$$\rho(F_1,\, F_2, \underset{\sim}{t}) = \left|\underset{\sim}{A}\,\underset{\sim}{B}\right|^{1/4} \left|\frac{1}{2}\left(\underset{\sim}{A}+\underset{\sim}{B}\right)\right|^{-1/2} \exp\left\{(\underset{\sim}{a}, \underset{\sim}{t}) + \left(\underset{\sim}{C}^{-1}\,\underset{\sim}{t}, \underset{\sim}{t}\right)\right\}$$

Corolary 3. For $F_1 = F_2 = F$, we have:

$$\rho(F,\, \underset{\sim}{t}) = \exp\left\{(\underset{\sim}{a},\, \underset{\sim}{t}) + 1/2(\underset{\sim}{A}\,\underset{\sim}{t},\, \underset{\sim}{t})\right\} = M_X(\underset{\sim}{t})$$

where $M_X(\underset{\sim}{t})$ is the moment generating function of F.

The conclusion (result) of theorem 1 is naturally generalized for r $k-$dimensional normal distributions. To accomplish this objective, we first generalize the concept of $\rho(F_1,\, F_2,\, \underset{\sim}{t})$, by considering r distributions $F_1, \ldots, F_r$, defined over the same space $\mathbb{R}$, with probability density functions $f_1(x), \ldots, f_r(x)$ with respect to a measure on $\mathbb{R}$, and let us suppose that the integral below be absolutely convergent. Then we define:

Definition 2

$$\rho(F_1, \ldots, F_r,\, t) = \int_{\mathbb{R}} \left\{\prod_{j=1}^{r} \exp(tx)\, f_j(x)\right\}^{1/r} dm$$

If $F_1, \ldots, F_r$ denote r $k-$ dimensional non singular normal distributions whose probability density functions are given for $j = 1, \ldots, r$ by

$$(2.11) \qquad (2\pi)^{-k/2}\, |\underset{\sim}{A}_j|^{-1/2} \exp\left\{-1/2(\underset{\sim}{A}_j^{-1}(\underset{\sim}{x} - \underset{\sim}{a}_j),\, \underset{\sim}{x} - \underset{\sim}{a}_j)\right\}$$

where $\underset{\sim}{A}_j$ is the covariance matrix and $\underset{\sim}{a}_j$ the mean vector of F_j, respectively, we have the following result whose proof we omit:

Theorem 2

$$\rho(F_1, \ldots, F_r,\, \underset{\sim}{t}) = \rho_r(F_1, \ldots, F_r)\, M_G(\underset{\sim}{t})$$

where

$$\rho_r(F_1, \ldots, F_r) = \left\{ \prod_{j=1}^{r} |\underset{\sim}{A}_j|^{-1/2r} \right\} \left| \frac{1}{r} \sum_{j=1}^{r} \underset{\sim}{A}_j^{-1} \right|^{-1/2}$$

$$\cdot \exp\left\{ -1/2r \left\{ \sum_{\substack{j=2 \\ l \leqslant i < j}}^{r} ((\underset{\sim}{A}_j \underset{\sim}{D} \underset{\sim}{A}_i)^{-1} \underset{\sim}{a}_j, \underset{\sim}{a}_j - \underset{\sim}{a}_i - \right.\right.$$

$$\left.\left. - \sum_{\substack{i=1 \\ i < j \leqslant r}}^{r-1} ((\underset{\sim}{A}_i \underset{\sim}{D} \underset{\sim}{A}_j)^{-1} \underset{\sim}{a}_i, \underset{\sim}{a}_j - \underset{\sim}{a}_i) \right\} \right\}$$

is the affinity between r $k-$dimensional normal (non singular) distributions obtained by Matusita (1967a) and expressed in another form by Campos (1978); and

$$M_G(\underset{\sim}{t}) = \exp\left\{ \left(\underset{\sim}{D}^{-1} \left(\sum_{j=1}^{r} \underset{\sim}{A}_j^{-1} \underset{\sim}{a}_j \right), \underset{\sim}{t} \right) + \frac{1}{2} \left(r \underset{\sim}{D}^{-1} \underset{\sim}{t}, \underset{\sim}{t} \right) \right\}$$

is the moment generating functions of a $k-$dimensional normal distribution with mean vector $\underset{\sim}{D}^{-1} \left(\sum_{j=1}^{r} \underset{\sim}{A}_j^{-1} \underset{\sim}{a}_j \right)$ and covariance matrix $r \underset{\sim}{D}^{-1}$ with $\underset{\sim}{D} = \sum_{j=1}^{r} \underset{\sim}{A}_j^{-1}$ and $\underset{\sim}{t}$ a $k-$dimensional (column) vector.

With the objective of generalizing these results, as those obtained by Matusita (1966, 1967a) or Campos (1978) we introduce the following definition.

Definition 3. Let $F_1, \ldots, F_r$ be multivariate distributions defined on the same space $\mathbb{R}$ and let $f_1(x), \ldots, f_r(x)$ be their respective probability density functions. Let us suppose that there are r scalars $s_1, \ldots, s_r$ such that:

$$\sum_{j=1}^{r} s_j = 1 \quad \text{and} \quad 0 \leqslant s_j \leqslant 1 \quad \text{for} \quad j = 1, \ldots, r.$$

In these conditions, and if the integral below is absolutely convergent, we define:

$$D_r(s_1, \ldots, s_r, \underset{\sim}{t}_j) = \int_{\mathbb{R}^k} \prod_{j=1}^{r} \exp\left\{ s_j(\underset{\sim}{x}, \underset{\sim}{t}_j) \right\} f_j^{s_j}(x)\, dx_1 \ldots dx_k$$

where $\underset{\sim}{t}_j$ is a $k-$dimensional (column) vector with components $t_{j_1}, \ldots, t_{j_k}$, $j = 1, \ldots, r$.

If $F_1, \ldots, F_r$ denote $k-$dimensional normal (non singular) distributions defined as (2.11) we establish the following result:

Theorem 3

$$D_r(s_1, \ldots, s_r; \underset{\sim}{t}_j) = D_r(s_1, \ldots, s_r)\, M_G\left(\sum_{j=1}^{r} s_j \underset{\sim}{t}_j\right) \tag{2.12}$$

where:

(2.13)

$$D_r(s_1, \ldots, s_r) = \left\{\prod_{j=1}^{r} |\underset{\sim}{A}_j|^{-s_j/2}\right\} \left|\sum_{j=1}^{r} s_j \underset{\sim}{A}_j^{-1}\right|^{-1/2}$$

$$\cdot \exp\left\{-1/2\left\{\sum_{\substack{j=2 \\ l \leqslant i < j}}^{r} (s_i s_j (\underset{\sim}{A}_j \underset{\sim}{C} \underset{\sim}{A}_i)^{-1} \underset{\sim}{a}_j, \underset{\sim}{a}_j - \underset{\sim}{a}_i) - \right.\right.$$

$$\left.\left. - \sum_{\substack{i=1 \\ i < j \leqslant r}}^{r} (s_i s_j (\underset{\sim}{A}_i \underset{\sim}{C} \underset{\sim}{A}_j)^{-1} \underset{\sim}{a}_i, \underset{\sim}{a}_j - \underset{\sim}{a}_i)\right\}\right\}$$

and $M_G\left(\sum_{j=1}^{r} s_j \underset{\sim}{t}_j\right)$ is the moment generating function for a $k-$dimensional normal distribution with mean vector $\underset{\sim}{C}^{-1}\left(\sum_{j=1}^{r} s_j \underset{\sim}{A}_j^{-1} \underset{\sim}{a}_j\right)$ and covariance matrix $\underset{\sim}{C}^{-1}$, expressed by:

$$M_G\left(\sum_{j=1}^{r} s_j \underset{\sim}{t}_j\right) = \exp\left\{\left(\underset{\sim}{C}^{-1}\left(\sum_{j=1}^{r} s_j \underset{\sim}{A}_j^{-1} \underset{\sim}{a}_j\right), \sum_{j=1}^{r} s_j \underset{\sim}{t}_j\right) + \right.$$

$$\left. + \left(1/2\, \underset{\sim}{C}^{-1}\left(\sum_{j=1}^{r} s_j \underset{\sim}{t}_j\right), \sum_{j=1}^{r} s_j \underset{\sim}{t}_j\right)\right\} \tag{2.14}$$

with $\underset{\sim}{C} = \sum_{j=1}^{r} s_j \underset{\sim}{A}_j^{-1}$.

Proof: By the definition 3, we have:

(2.15)

$$D_r(s_1, \ldots, s_r; \underset{\sim}{t}_j) = (2\pi)^{-k/2} \prod_{j=1}^{r} |\underset{\sim}{A}_j|^{-s_j/2} \int_{\mathbb{R}^k} \exp\left\{-\frac{1}{2} Q\right\} dx_1 \ldots dx_k$$

where

$$Q = \sum_{j=1}^{r} s_j \left(\underset{\sim}{A}_j^{-1}(\underset{\sim}{x} - \underset{\sim}{a}_j), \underset{\sim}{x} - \underset{\sim}{a}_j\right) - 2\sum_{j=1}^{r} s_j(\underset{\sim}{x}, \underset{\sim}{t}_j)$$

That is,

(2.16) $$Q = (\underset{\sim}{C}\,\underset{\sim}{x}, \underset{\sim}{x}) - 2(\underset{\sim}{b}, \underset{\sim}{x}) - 2(\underset{\sim}{t}, \underset{\sim}{x}) + \sum_{j=1}^{r} \left(s_j \underset{\sim}{A}_j^{-1} \underset{\sim}{a}_j, \underset{\sim}{a}_j\right)$$

with

$$\underset{\sim}{b} = \sum_{j=1}^{r} s_j \underset{\sim}{A}_j^{-1} \underset{\sim}{a}_j$$

and

$$\underset{\sim}{t} = \sum_{j=1}^{r} s_j \underset{\sim}{t}_j$$

After the transformation

$$\underset{\sim}{y} = \underset{\sim}{C}^{1/2} \underset{\sim}{x}$$

whose Jacobian is mod $|\underset{\sim}{C}|^{-1/2}$ and same intermediate steps, (2.16) may be written

(2.17)

$$Q = \left(\underset{\sim}{y} - \underset{\sim}{C}^{-1/2}(\underset{\sim}{b} + \underset{\sim}{t}), \underset{\sim}{y} - \underset{\sim}{C}^{-1/2}(\underset{\sim}{b} + \underset{\sim}{t})\right) - \left(\underset{\sim}{C}^{-1}\underset{\sim}{b}, \underset{\sim}{t}\right) -$$
$$- \left(\underset{\sim}{C}^{-1}\underset{\sim}{t}, \underset{\sim}{b}\right) - \left(\underset{\sim}{C}^{-1}\underset{\sim}{t}, \underset{\sim}{t}\right) + \sum_{j=1}^{r} \left(s_j \underset{\sim}{A}_j^{-1} \underset{\sim}{a}_j, \underset{\sim}{a}_j\right) - \left(\underset{\sim}{C}^{-1}\underset{\sim}{b}, \underset{\sim}{b}\right)$$

One may also prove that:

$$\sum_{j=1}^{r}\left(s_j\,\underset{\sim}{A}_j^{-1}\,\underset{\sim}{a}_j,\,\underset{\sim}{a}_j\right)-\left(\underset{\sim}{C}^{-1}\,\underset{\sim}{b},\,\underset{\sim}{b}\right)=\sum_{\substack{j=2\\ l\leqslant i<j}}^{r}\left(s_i\,s_j(\underset{\sim}{A}_j\,\underset{\sim}{C}\,\underset{\sim}{A}_i)^{-1}\,\underset{\sim}{a}_j,\,\underset{\sim}{a}_j-\underset{\sim}{a}_i\right)-$$

$$-\sum_{\substack{i=1\\ i<j\leqslant r}}^{r-1}\left(s_i\,s_j(\underset{\sim}{A}_i\,\underset{\sim}{C}\,\underset{\sim}{A}_j)^{-1}\,\underset{\sim}{a}_i,\,\underset{\sim}{a}_j-\underset{\sim}{a}_i\right)$$

On applying this result, (2.17) is expressed as:

$$Q=Q_3+Q_1-Q_2-2\left(\underset{\sim}{C}^{-1}\,\underset{\sim}{b},\,\underset{\sim}{t}\right)-\left(\underset{\sim}{C}^{-1}\,\underset{\sim}{t},\,\underset{\sim}{t}\right) \tag{2.18}$$

where:

$$Q_3=\left(\underset{\sim}{y}-\underset{\sim}{C}^{-1/2}(\underset{\sim}{b}+\underset{\sim}{t}),\,\underset{\sim}{y}-\underset{\sim}{C}^{-1/2}(\underset{\sim}{b}+\underset{\sim}{t})\right)$$

$$Q_1=\sum_{\substack{j=2\\ l\leqslant i<j}}^{r}\left(s_i\,s_j(\underset{\sim}{A}_j\,\underset{\sim}{C}\,\underset{\sim}{A}_i)^{-1}\,\underset{\sim}{a}_j,\,\underset{\sim}{a}_j-\underset{\sim}{a}_i\right)\quad\text{and}$$

$$Q_2=\sum_{\substack{i=1\\ i<j\leqslant r}}^{r-1}\left(s_i\,s_j(\underset{\sim}{A}_i\,\underset{\sim}{C}\,\underset{\sim}{A}_j)^{-1}\,\underset{\sim}{a}_i,\,\underset{\sim}{a}_j-\underset{\sim}{a}_i\right)$$

By substitution of (2.18) in (2.15), we obtain:

(2.19)

$$D_r(s_1,\ldots,s_r;\,\underset{\sim}{t}_j)=(2\pi)^{-k/2}\prod_{j=1}^{r}|\underset{\sim}{A}_j|^{-s_j/2}\cdot$$

$$\cdot\exp\left\{-\frac{1}{2}(Q_1-Q_2)\right\}\exp\left\{\left(\underset{\sim}{C}^{-1}\,\underset{\sim}{b},\,\underset{\sim}{t}\right)+\left(\frac{1}{2}\,\underset{\sim}{C}^{-1}\,\underset{\sim}{t},\,\underset{\sim}{t}\right)\right\}\cdot$$

$$\cdot\int_{\mathbb{R}^k}\exp\left\{-\frac{1}{2}\,Q_3\right\}|\underset{\sim}{C}|^{-1/2}\,dy_1\ldots dy_k$$

The integral of (2.19) after the transformation

$$\underset{\sim}{z}=\underset{\sim}{y}-\underset{\sim}{C}^{-1/2}(\underset{\sim}{b}+\underset{\sim}{t})$$

becomes equal

$$|\underset{\sim}{C}|^{-1/2}(2\pi)^{k/2}$$

By using this result in (2.19), we obtain, in accord with (2.13) and (2.14) the result that we have established through theorem 3, that is,

$$D_r(s_1, \ldots, s_r; \underset{\sim}{t}_j) = D_r(s_1, \ldots, s_r) \cdot M_G \left(\sum_{j=1}^{r} s_j \underset{\sim}{t}_j \right)$$

This result is a generalization in the sense of summarize the results established through theorems 1 and 2 as those obtained by Matusita (1966, 1967a) or Campos (1978).

REFERENCES

Ahmad, I.A. (1980). «Nonparametric estimation of an affinity measure between two absolutely continuous distributions with hypotheses testing applications». *Annals of the Institute of Statistical Mathematics*, 32, 223-240.

Ahmad, I.A. (1980). «Nonparametric estimation of Matusita's measure of affinity between absolutely continuous distributions». *Annals of the Institute of Statistical Mathematics*, 32, 241-246.

Campos, A.D. (1978). «Afinidade entre distribuições normais multivariadas». *Atas do 3.º Simpósio Nacional de Probabilidade e Estatística* (ed. IME-USP), 75-79.

Cuadras, C.M. (1988). «Statistical distances». *Estadística Española*, 119, 295-378.

Cuadras, C.M. (1989). «Distance analysis in discrimination and classification using both continous and categorical variables». In *Statistical Data Analysis and Inference*, (Y. Dodge, ed.), Elsevier, Amsterdam, 459-473.

Cuadras, C.M. and Fortiana, J. (1993). «Applying distances in statistics». *Qüestiió*, 17, 39-74.

Khan, A.H. & Ali, S.M. (1971). «A new coefficient of association». *Annals of the Institute of Statistical Mathematics*, 23, 41-50.

Krzanowski, W.J. (1983). «Distance between populations using mixed continuous and categorical variables». *Biometrika*, 70, 235-243.

Matusita, K. (1954). «On the estimation by the minimum distance method». *Annals of the Institute of Statistical Mathematics*, 5, 59-65.

Matusita, K. (1955). «Decision rules based on the distance for problems of fit, two samples and estimation». *Annals of the Institute of Statistical Mathematics*, 26, 631-640.

Matusita, K. (1956). «Decision rule based on the distance for the classification problem». *Annals of the Institute of Statistical Mathematics*, 8, 67-77.

Matusita, K. (1961). «Interval estimation based on the notion of affinity». *Bulletin of the International Statistical Institute*, 38, 4, 241-244.

Matusita, K. (1964). «Distance and decision rules». *Annals of the Institute of Statistical Mathematics*, 16, 305-315.

Matusita, K. (1966). «A distance and related statistics in multivariate analysis». *Multivariate Analysis - Proceedings of an International Symposium* (ed. P.R. Krishnaiah). Academic Press New York, 187-200.

Matusita, K. (1967a). «On the notion of affinity of several distributions and some of its applications». *Annals of the Institute of Statistical Mathematics*, 19, 181-192.

Matusita, K. (1967b). «Classification based on distance in multivariate Gaussian cases». *Proceedings of the Fifth Berkeley Simposium on Mathematical Statistics and Probability*. University California Press, 1, 299-304.

Matusita, K. (1973). «Correlation and affinity in Gaussian cases». *Multivariate Analysis - III - Proceedings of the Third International - Symposium* (ed. P.R. Krishnaiah), Academic Press, New Yor, 345-349.

Matusita, K. (1973). «Discrimination and the affinity of distributions». In: *Discriminant Analysis and Applications*. (T. Cacoullos, ed.), Academic Press, N.Y., 213-223.

Matusita, K. (1977). «Cluster analysis and affinity of distributions. Recent Developments in Statistics». *Proceedings of the 1976 European Meeting of Statisticians*, 537-544.

Matusita, K. & Motoo, M. (1955). «On the fundamental theorem for the decision rule based distance II.»*Annals of the Institute of Statistical Mathematics*, 7, 137-142.

Matusita, K. & Akaike, H. (1956). «Decision rules based on the distance for the problems of independence, invariance and two samples». *Annals of the Institute of Statistical Mathematics*, 7, 67-80.

QÜESTIIÓ, vol. 23, 2, p. 239-259, 1999

EFECTO DEL NÚMERO DE INDICADORES POR FACTOR SOBRE LA IDENTIFICACIÓN Y ESTIMACIÓN EN MODELOS ADITIVOS DE ANÁLISIS FACTORIAL CONFIRMATORIO

A. OLIVER
J. M. TOMÁS
P. M. HONTANGAS
Universitat de València*

La matriz multirrasgo-multimétodo (MRMM) es un diseño de investigación de larga tradición en Psicología. Las técnicas de análisis de datos adecuadas para una correcta extracción de conclusiones han estado sujetas a controversia. Parece, no obstante, que diversos modelos de análisis factorial confirmatorio resultan muy adecuados. De entre los diversos modelos, dos de ellos han recibido gran atención, el modelo completo, que apareció primero en la literatura, y el de unicidades correlacionadas, que parece una alternativa razonable a los problemas que aparecen en el primero. Los resultados de ambos modelos en la literatura se refieren a situaciones con un solo indicador por combinación rasgo-método. La presente investigación simula datos de matrices MRMM para múltiples indicadores por combinación rasgo-método y somete a prueba la adecuación de las estimaciones de ambos modelos. Los resultados apuntan a un mejor comportamiento del modelo completo, si bien los sesgos, aunque triviales en cuantía, aumentan conforme aumenta la correlación entre métodos.

Effects of number of indicators per factor on identification and estimation of confirmatory factor analysis models.

Palabras clave: Estimación, sesgo, análisis factorial confirmatorio (AFC), matriz MRMM, número de indicadores, correlación entre métodos.

Clasificación AMS: 62H25, 62P15

La presente investigación se enmarca en el proyecto PB96-0791 del Programa Sectorial de Promoción General del Conocimiento subvencionado por la DGES, y en sendos proyectos concedidos a los dos primeros autores por el subprograma SEUID (MEC)-Royal Society de Londres. Los autores desean agradecer los comentarios de un revisor, que mejoraron sustancialmente la versión final de este trabajo. Una versión preliminar de este trabajo se presentó en el *V Congreso de Metodología de las Ciencias Humanas y Sociales*, Sevilla, 23-26 de Septiembre, 1997.

* Àrea de Metodologia de les Ciències del Comportament. Facultat de Psicologia. Universitat de València. Av. Blasco Ibáñez, 21. 46010 València. España. Correo electrónico: oliver@uv.es.

– Recibido en noviembre 1998.
– Aceptado en febrero de 1999.

1. INTRODUCCIÓN

El diseño de investigación mediante la matriz multirrasgo-multimétodo fue presentado por Campbell y Fiske (1959). En su aplicación paradigmática se miden distintas variables mediante diversos métodos. Así, puede medirse extraversión, neuroticismo y depresión por medio de tres métodos de medida (autoinforme, registro clínico y observación externa, por ejemplo). Este ejemplo de diseño sería una matriz multirrasgo-múltimétodo 3×3, con tres rasgos y tres métodos. El número de rasgos y de métodos utilizados en la literatura es muy variable, y pueden encontrase diseños de 3×2, 4×4, 9×9, etc. El objetivo de la matriz multirrasgo-multimétodo era ofrecer un diseño de investigación que permitiera analizar la validez convergente-discriminante de las medidas incluidas.

Este diseño de investigación sigue plenamente vigente, aunque el análisis de datos propuesto por estos autores ha sido ampliamente criticado (entre otros, Bagozzi, 1993; Schmitt y Stults, 1986; Widaman, 1985). Como respuesta a estas críticas, se han planteado aplicaciones de diferentes tipos de técnicas para mejorar la extracción de conclusiones. De entre éstas, las soluciones más prometedoras parecen ser las basadas en distintos modelos de análisis factorial confirmatorio (Jöreskog, 1971, 1972; Marsh, 1988, 1989, Rindskopf, 1983).

Dentro del análisis factorial confirmatorio se proponen diversas parametrizaciones o modelos, que varían en términos de supuestos, condiciones de aplicación y adecuación de las estimaciones. El primero en aparecer es el modelo completo, o modelo de rasgos correlacionados y métodos correlacionados (Jöreskog, 1971). Una representación de éste puede verse en la figura 1 (a) para tres rasgos y tres métodos. Como una alternativa a este modelo aparece el modelo de unicidades correlacionadas presentado por Marsh (1988, 1989), basándose en trabajos de Kenny (1976, 1979). Esta parametrización extrae las conclusiones sobre efectos de método a partir de la correlación entre unicidades de las medidas de distintos rasgos que tienen en común un mismo método. Una representación de este modelo, también para tres rasgos y tres métodos, aparece en la figura 1 (b).

El modelo completo ha sufrido fuertes críticas ya que, aunque cuenta con ventajas frente a otras técnicas, presenta muchos problemas, especialmente de identificación y estimación (ver, entre otros, Brannick y Spector, 1990; Kenny y Kashy, 1992; Kumar y Dillon, 1992; Marsh y Bailey, 1991; Wothke, 1996). Por su parte, el modelo de unicidades correlacionadas ha aparecido como una alternativa razonablemente robusta a violaciones de sus supuestos, que ofrece estimaciones adecuadas de la validez convergente-discriminante y de los efectos de método, y que presenta problemas de identificación y/o estimación con escasa probabilidad (Marsh y Bailey, 1991). Los resultados obtenidos se derivan de estudios analíticos y empíricos, con datos reales y de simulación, y resultan altamente consistentes entre sí, lo que ha permitido a un número de autores re-

comendar la elección del modelo de unicidades correlacionadas en estudios de matrices multirrasgo-multimétodo (Brannick y Spector, 1990; Marsh y Bailey, 1991). Los resultados anteriores se han establecido a partir de datos de matrices multirrasgo-multimétodo en su definición más prototípica. Esto es, cuando se utiliza una sola medida, una variable, como indicador de cada combinación rasgo-método. Si volvemos a la figura 1 (a), la primera variable observable es una medida del primer rasgo tomada mediante el primer método de medida, y es la única medida de ese rasgo tomada con ese método.

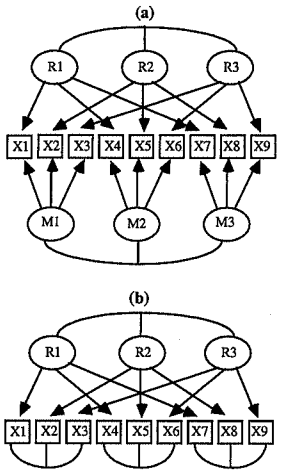


Figura 1. Modelo completo (a) y modelo de unicidades correlacionadas (b).

Notas: Las unicidades se eliminan por simplicidad; las líneas curvas representan covarianzas

Es el único indicador de la combinación primer rasgo por primer método, y de la misma forma ocurre para cada combinación rasgo-método. Lo mismo sucede cuando el modelo es el de unicidades correlacionadas, que supone una parametrización distinta, pero comparte con el anterior modelo las variables observables o medidas. Aún siendo lo más usual en la literatura, ninguna característica del diseño MRMM hace pensar que escoger una única variable observable por combinación rasgo-método sea la única posibilidad, ni siquiera la más recomendable. El uso de más de un indicador o variable observable por combinación de rasgo-método, bien corresponda a distintas medidas de un mismo constructo o a la misma variable medida repetidamente en el tiempo, puede conllevar importantes ventajas. Estas ventajas son comparables a aquéllas que en general se producen al aumentar el número de indicadores en cualquier modelo de análisis factorial confirmatorio. De hecho, «la aproximación de un solo indicador debe verse como un caso especial de la aproximación más general de múltiples indicadores, en la

cual se imponen sobre la estructura de los indicadores múltiples restricciones implícitas, a veces no probadas, y quizá arbitrarias» (Marsh, 1993, p. 78). Saris y Andrews (1991) proponen la formulación como modelo de análisis factorial de segundo orden con un solo indicador por combinación rasgo-método. Posteriormente, Marsh (1993) propuso el uso de múltiples indicadores para representar cada combinación rasgo-método como un factor latente, usando estos factores latentes de primer orden como indicadores de los factores de rasgo y de método, que se modelarían, por tanto, como factores de segundo orden. Una representación de un modelo de este tipo puede verse en la figura 2. Además, esta aproximación puede usarse para aplicar las directrices de Campbell y Fiske (1959), eliminando algunos de sus problemas. Por ejemplo, se utilizan variables latentes libres de error, en lugar de variables observables (Marsh, 1993).

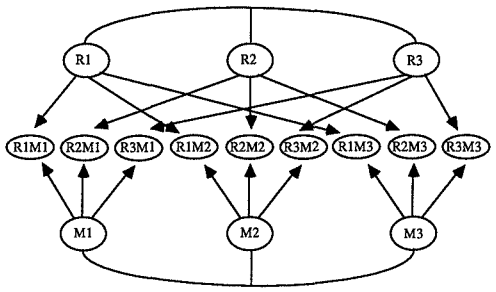


Figura 2. Modelo de matriz multirrasgo-multimétodo con indicadores múltiples analizado mediante análisis factorial confirmatorio de segundo orden.

Notas: Se omiten las variables observables y las unicidades por simplicidad; cada factor de primer orden se infiere a partir de dos o más variables observables; las unicidades, residuales, se eliminan por simplicidad; las líneas curvas representan covarianzas.

Aún sin considerar factores de segundo orden, el análisis factorial confirmatorio de matrices multirrasgo-multimétodo se puede aplicar, y se ha aplicado con éxito, a situaciones en que se tiene más de una variable (indicador) por combinación de rasgo-método. Este es un caso típico cuando se analizan efectos de método en el análisis de ítems de una escala e incluso con combinaciones o paquetes de ítems. En estos casos, los efectos de rasgo se modelan como factores de primer orden, y los efectos de método bien como factores de primer orden, bien mediante unicidades correlacionadas, en función de que se utilice el modelo completo o el de unicidades correlacionadas. Un modelo de estas características puede verse en la figura 3, donde las variables observables pueden ser ítems de una escala, o simplemente distintas medidas observables de un mismo constructo.

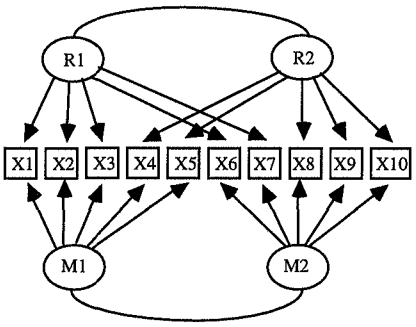


Figura 3. Modelo multirrasgo-multimétodo con más de un indicador por combinación de rasgo-método parametrizado como un modelo completo.

Notas: Se omiten las unicidades por simplicidad; las líneas curvas representan covarianzas.

Lo que se plantea en estos modelos son efectos de método asociados a los ítems y/o variables observables bajo consideración. Se asume que éstos pueden formularse y puntuarse de formas muy diferentes y que estas características formales, al margen del contenido, del rasgo que se intente medir, pueden alterar las relaciones entre variables (Saris y Meurs, 1990). Estos efectos pueden considerarse de método, y por tanto son susceptibles de estudio bajo la lógica del diseño de matrices multirrasgo-multimétodo. De entre las características más estudiadas en la literatura pueden citarse el número de categorías, los ítems comparativos frente a los absolutos, longitud de las baterías, ítems formulados positivamente frente a ítems en negativo, etc. (por ejemplo, Andrews, 1986; Molenaar, 1986). En el estudio de estas características se han utilizado tanto el modelo de rasgos correlacionados y métodos correlacionados como el modelo de unicidades correlacionadas. Sin embargo, existe una laguna en la literatura al respecto del funcionamiento diferencial de ambos modelos en situaciones donde se emplean múltiples indicadores. La justificación del uso de uno u otro modelo se ha basado en la literatura que analiza la situación paradigmática de un indicador por combinación de rasgo-método. Esta situación puede ser muy diferente, desde un punto de vista estadístico, a un diseño que estudie los efectos de método cuando se presentan múltiples indicadores. Una aplicación concreta puede arrojar luz sobre las diferencias. Un diseño de matriz multirrasgo-multimétodo 3×3 presenta nueve variables observables y seis variables latentes, lo que ofrece una razón de variables a factores de 1.5. El uso de múltiples indicadores se puede ilustrar con el trabajo de Marsh (1996), ampliado posteriormente por Tomás y Oliver (1999), donde se aplicó análisis factorial confirmatorio para estudiar efectos de método en los ítems de la escala de autoestima de Rosenberg. Se asumía que la escala era unifactorial (un factor de rasgo), pero el formular los ítems en forma

positiva frente a negativa introducía varianza de método (dos factores de método). Dado que la escala de Rosenberg está compuesta de diez ítems, esto ofrecía una ratio de variables a factores de 3.33. Otro ejemplo de aplicación de análisis factorial confirmatorio al estudio de efectos de método con múltiples indicadores se ofrece en Bagozzi y Heatherton (1994). Estos autores estudian los efectos de dos factores de método en una medida multifactorial de autoestima-estado de 20 ítems que presentaba una ratio de variables observables a factores también de 3.33. En trabajos realizados sobre la versión en castellano del STAI-estado (Hontangas, Oliver y Tomás, 1997), se comparan distintas estructuras factoriales con factores de método, que ofrecían ratios de variables a factores de 5, 6.66 y hasta 10. En resumen, aplicar el análisis factorial confirmatorio de matrices multirrasgo-multimétodo generalmente llevará asociada una mayor razón de variables a factores, lo que puede alterar las virtudes y problemas conocidos hasta el momento con estos modelos.

El objetivo de este trabajo es comparar la capacidad del modelo de rasgos correlacionados y métodos correlacionados frente al modelo de unicidades correlacionadas, para estimar de forma adecuada parámetros poblacionales conocidos allá donde se evalúen efectos de método cuando exista más de un indicador por combinación de rasgo-método (ver figuras 3 y 4). Se estudian ambos modelos a través de diversos niveles de correlación entre factores de método, y variando también el número de indicadores por combinación de rasgo-método. Las hipótesis principales que se pretenden contrastar son:

- Al respecto del modelo completo:
  a) presentará menores problemas de identificación y estimará mejor a medida que aumente el número de indicadores por combinación rasgo-método, y
  b) estimará de forma adecuada a todos los niveles de correlación entre métodos.

- Al respecto del modelo de unicidades correlacionadas:
  a) presentará problemas de estimación mayores a medida que aumente el número de indicadores, y
  b) tendrá mayores problemas de estimación a medida que aumente la correlación entre métodos.

MÉTODO

Modelos hipotéticos y generación de las muestras

Por tratarse de un experimento Monte Carlo, se extraen muestras de modelos poblacionales conocidos, en este caso de matrices multirrasgo-multimétodo. Los modelos poblacionales se corresponden en genérico con el modelo de la figura 3.

Tanto la generación de datos como la estimación de los modelos de ecuaciones estructurales se realizó mediante el paquete estadístico EQS 5.1 (Bentler, 1995). Las capacidades de simulación del programa permitieron calcular los dieciséis conjuntos de matrices de varianzas-covarianzas necesarias para el diseño, con cien replicaciones por conjunto. Las variables se generaron para que cumplieran el supuesto de normalidad multivariada. El número de iteraciones máximo para alcanzar la convergencia en el proceso de estimación se fijó a 100.

Diseño

El diseño es factorial mixto 4×4×2 con medidas repetidas en el último factor (tipo de modelo de AFC). Se manipularon las siguientes variables:

- *Número de indicadores por combinación de rasgo-método*, con cuatro niveles, 2, 3, 4 y 5 indicadores.
- *Magnitud de la correlación entre factores de método*, con cuatro niveles, desde factores ortogonales, a valores de .2, .4 y .6. La correlación nula entre métodos es asumida por el modelo de unicidades correlacionadas. Por el contrario, el nivel más alto de correlación, .6, puede considerarse una correlación entre métodos elevada. El estudio más completo en este contexto consideró correlaciones entre factores de método desde .25 a .49 (Marsh y Bailey, 1991).
- Modelo estimado, con dos niveles, modelo de rasgos correlacionados y métodos correlacionados (que se corresponde en genérico con el modelo de la figura 3) y modelo de unicidades correlacionadas (que se corresponde en genérico con el modelo de la figura 4).

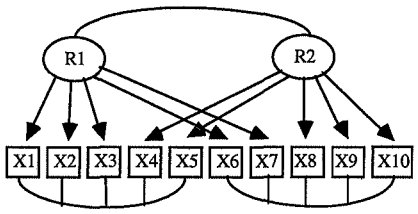


Figura 4. Modelo multirrasgo-multimétodo con más de un indicador por combinación de rasgo-método parametrizado como un modelo de unicidades correlacionadas.

Notas: Se omiten las unicidades por simplicidad; las líneas curvas representan covarianzas.

Las variables dependientes pueden incluirse en tres bloques:

a) Problemas de identificación-estimación, que incluye la variable convergencia (si el modelo converge o no); la variable soluciones impropias (si el modelo ha presentado o no valores fuera de rango, etc.), y soluciones bien definidas (si el proceso de estimación converge y no se presenta ninguna solución impropia, entonces la solución se considera bien definida).
b) Correlación entre rasgos. Para cada modelo ha de estimarse la correlación entre los dos factores de rasgo planteados en la generación de datos. Esta correlación presentaba un valor poblacional de .4. Por lo tanto, la variable dependiente necesaria para evaluar la adecuación de la estimación de este parámetro es la correlación estimada para cada una de las replicaciones.
c) Saturaciones factoriales. A diferencia del caso de las correlaciones entre rasgos, en que el valor estimado es suficiente como variable dependiente, en el caso de las saturaciones se necesitan dos variables dependientes para la evaluación de la adecuación de la estimación, que respectivamente miden el signo y cuantía en valor absoluto de la discrepancia entre las estimaciones y sus valores poblacionales. Estas dos variables dependientes, las mismas utilizadas por Marsh y Bailey (1991) son, respectivamente:
   - Sesgo, medido como el promedio para cada modelo del alejamiento de la saturación estimada frente a su valor poblacional.
   - Precisión, operacionalizada como la raíz cuadrada de la media de la diferencia entre las saturaciones estimadas y su valor poblacional al cuadrado. Esto es, la raíz del error cuadrático medio (RMSE, root mean square error).

El tamaño de la muestra se controló por constancia. A través de todos los modelos se empleó un tamaño muestral de 500. Este tamaño de la muestra se considera muy adecuado en estudios de simulación en modelos de ecuaciones estructurales (Anderson & Gerbing, 1984; Boomsma, 1982). El método de estimación también se controló por constancia, se empleó máxima verosimilitud dado que en la generación de los datos se asumió la normalidad multivariada.

RESULTADOS

Los resultados se ofrecen separados para los distintos grupos de variables dependientes consideradas: a) problemas de estimación, considerando las convergencias, soluciones impropias y soluciones mal definidas; b) la estimación de la correlación entre rasgos (validez discriminante), y c) la estimación de las saturaciones de rasgo (validez convergente).

No convergencias, soluciones impropias y soluciones mal definidas

Las frecuencias de no convergencias, las soluciones impropias y las soluciones mal definidas pueden consultarse en la tabla 1.

Tabla 1. Frecuencias de no convergencias, soluciones impropias y soluciones mal definidas para los dos modelos. Se incluyen las frecuencias por condición y a través de los niveles de las variables independientes (V. I.) en cada modelo.

		Modelo completo			**Modelo de unicidades correlacionadas**		
Tipo de matriz	Correlación métodos	No convergencias	Soluciones impropias	Sol. mal definidas	No convergencias	Soluciones impropias	Sol. mal definidas
	.0	7	18	25	0	100	100
Dos	.2	6	56	61	0	100	100
indicadores	.4	2	52	54	0	100	100
	.6	4	56	58	0	100	100
	.0	2	4	6	0	100	100
Tres	.2	3	24	27	0	100	100
indicadores	.4	5	28	33	0	100	100
	.6	6	40	45	0	100	100
	.0	2	4	6	0	100	100
Cuatro	.2	2	20	22	0	100	100
indicadores	.4	4	24	27	0	100	100
	.6	6	30	34	0	100	100
	.0	0	0	0	0	100	100
Cinco	.2	5	5	10	0	100	100
indicadores	.4	1	5	6	0	100	100
	.6	1	10	11	0	100	100
V. I.							
	2	19	182	198	0	400	400
	3	16	96	111	0	400	400
Indicadores	4	14	78	89	0	400	400
	5	7	20	27	0	400	400
	.0	11	26	37	0	400	400
Corr. entre	.2	16	105	120	0	400	400
métodos	.4	12	109	120	0	400	400
	.6	17	146	148	0	400	400

Con respecto al modelo completo, el número de indicadores presenta un efecto claro sobre la convergencia, las soluciones impropias y, por tanto, las soluciones mal defini-

das. A medida que aumentan el número de indicadores disminuyen los problemas de identificación y estimación habituales en este modelo. Las soluciones no convergentes aparecen, en cualquier caso con baja frecuencia, mientras que las soluciones impropias son bastante más comunes, especialmente los casos Heywood. Por su parte, la presencia de correlación entre factores de método no afecta de forma apreciable a las no convergencias, que ocurren en una pequeña proporción. Sin embargo sí afecta, y ostensiblemente, a la frecuencia de aparición de soluciones impropias y mal definidas. El número de éstas aumenta drásticamente cuando aparece esta correlación, y tanto más cuanto mayor es ésta.

Por su parte, para el modelo de unicidades correlacionadas no se producen problemas de convergencia. Sin importar el número de indicadores ni la correlación entre métodos, todos los modelos convergen. Sin embargo, las soluciones impropias (y por tanto las mal definidas) ocurren siempre. Estos problemas de estimación suelen ser valores fuera de rango, correlaciones mayores de uno, varianzas de error negativas, etc. Así pues, el modelo de unicidades correlacionadas presentó problemas de estimación graves que afectaban a todos los modelos. Cuando en cualquier modelo de ecuaciones estructurales aparecen soluciones mal definidas, está totalmente contraindicado interpretar los resultados. Esto se debe a la dificultad genérica de establecer la identificación del modelo y, por tanto, a la cautela que debe tenerse cuando cualquier resultado ofrece algún indicio de problemas de identificación. Por lo tanto, y ante la posibilidad de severos problemas de identificación, a partir de aquí se realizarán análisis por separado para el modelo completo y el de unicidades correlacionadas, con especial hincapié en el primero. En los análisis del modelo completo se utilizaron únicamente las soluciones bien definidas. Los análisis del modelo de unicidades correlacionadas se quedarán en el estudio descriptivo de los índices.

Correlación entre factores de rasgo

En la tabla 2 puede consultarse la correlación promedio por condición del estudio, así como las correlaciones promedio para los distintos niveles de las dos variables independientes. Hay que recordar que el valor poblacional se fijó a .4, y es frente a este valor que se deben evaluar las estimaciones. Por tanto, resulta sencillo apreciar las desviaciones frente a este valor. Aunque el modelo de unicidades correlacionadas no ofreció soluciones bien definidas, y por tanto los resultados deben interpretarse con cautela, se ofrecen también en la tabla 2 los promedios de la correlación entre los rasgos a través de niveles y condiciones.

Tabla 2. Número de soluciones bien definidas por celda, correlación media entre rasgos estimada, sesgo de la estimación (Notas: En el caso del modelo de unicidades correlacionadas, no hay ninguna solución bien definida, por lo que se ofrecen los resultados para las soluciones mal definidas, ME = sesgo, RMSR = raíz del error cuadrático medio, D.T. = desviación típica, las medias por nivel son las medias de la variable dependiente para cada nivel de las variables explicativas o independientes).

			Correlación entre rasgos			Saturaciones de rasgo			
Modelo	Matriz	ρ	N	r	D.T.	ME	D.T.	RMSR	D.T.
Completo			1175	.3863	.0970	−.0103	.0484	.0687	.0406
		.0	75	.4067	.0565	.0020	.0201	.0589	.0166
	Dos	.2	39	.4156	.1086	−.0266	.0688	.0942	.0568
	indicadores	.4	46	.3363	.1450	−.0413	.0831	.1168	.0770
		.6	42	.3712	.1372	−.0170	.0624	.1032	.0549
		.0	94	.4021	.0538	−.0001	.0199	.0515	.0131
	Tres	.2	73	.3885	.0766	−.0095	.0425	.0657	.0271
	indicadores	.4	67	.3627	.1310	−.0207	.0605	.0815	.0458
		.6	55	.3697	.1326	−.0192	.0680	.0924	.0525
		.0	94	.4055	.0499	−.0005	.0170	.0507	.0102
	Cuatro	.2	78	.3713	.1054	−.0156	.0499	.0693	.0322
	indicadores	.4	73	.3779	.1293	−.0168	.0603	.0779	.0410
		.6	66	.3664	.1227	−.0243	.0681	.0906	.0598
		.0	100	.3988	.0457	−.0007	.0174	.0448	.0079
	Cinco	.2	90	.4010	.0568	.0019	.0277	.0534	.0128
	indicadores	.4	94	.3857	.0875	−.0044	.0384	.0582	.0219
		.6	89	.3857	.1065	−.0119	.0533	.0685	.0370
		2	202	.3850	.1135	−.0173	.0606	.0881	.0567
	Número de	3	289	.3834	.0994	−.0109	.0486	.0698	.0385
	indicadores	4	311	.3821	.1039	−.0131	.0510	.0702	.0406
Medias		5	373	.3929	.0773	−.0037	.0365	.0559	.0240
por nivel		.0	363	.4030	.0511	.0000	.0185	.0509	.0129
	Correlación	.2	280	.3915	.0862	−.0099	.0462	.0667	.0336
	de métodos	.4	280	.3700	.1209	−.0176	.0596	.0785	.0492
		.6	252	.3747	.1217	−.0176	.0620	.0852	.0516
Unicidades			1600	.4358	.0571	.0226	.0249	.0674	.0220
		.0	100	.4015	.0573	.0056	.0214	.0704	.0291
	Dos	.2	100	.4334	.0582	.0165	.0222	.0713	.0260
	indicadores	.4	100	.4512	.0435	.0327	.0230	.0740	.0245
		.6	100	.4834	.0548	.0433	.0229	.0781	.0239
		.0	100	.3966	.0558	.0028	.0195	.0691	.0268
	Tres	.2	100	.4237	.0532	.0166	.0197	.0611	.0176
	indicadores	.4	100	.4432	.0550	.0281	.0210	.0653	.0174
		.6	100	.4657	.0526	.0422	.0210	.0716	.0180
		.0	100	.4003	.0523	.0030	.0177	.0711	.0300
	Cuatro	.2	100	.4257	.0482	.0144	.0184	.0629	.0175
	indicadores	.4	100	.4495	.0463	.0283	.0184	.0645	.0179
		.6	100	.4719	.0446	.0418	.0184	.0680	.0147
		.0	100	.3959	.0474	.0012	.0179	.0634	.0228
	Cinco	.2	100	.4190	.0402	.0146	.0201	.0593	.0175
	indicadores	.4	100	.4449	.0471	.0302	.0193	.0625	.0164
		.6	100	.4666	.0436	.0408	.0201	.0658	.0154
		2	400	.4424	.0613	.0245	.0266	.9735	.0269
	Número de	3	400	.4323	.0596	.0224	.0249	.0668	.0268
	indicadores	4	400	.4369	.0547	.0219	.0233	.0666	.0211
Medias		5	400	.4316	.0519	.0217	.0245	.0628	.0183
por nivel		.0	400	.3986	.0532	.0031	.0191	.0685	.0273
	Correlación	.2	400	.4255	.0504	.0150	.0201	.0636	.0204
	de métodos	.4	400	.4472	.0480	.0290	.0205	.0665	.0197
		.6	400	.4719	.0494	.0420	.0206	.0708	.0188

Se realizó un ANOVA entre-sujetos 4×4 para contrastar los efectos principales y de interacción sobre la variable dependiente correlación estimada entre factores de rasgo. Para este análisis el tipo de modelo se mantuvo constante; sólo se analizaron las soluciones bien definidas del modelo completo. Únicamente el efecto principal de la variable correlación entre métodos resultó estadísticamente significativo ($F_{3,1159} = 9.79, p < .001$). Dado el elevado tamaño de la muestra, se ofrecen las eta-cuadrado (tabla 3). Cuando los métodos son independientes o su correlación baja, el modelo es capaz de estimar de forma adecuada la correlación entre rasgos (la estimación de la validez discriminante es adecuada). Cuando la correlación entre los factores de método aumenta (.4 y .6) se produce una infraestimación de la correlación entre rasgos. La infraestimación que se produce, incluso a los niveles altos de correlación entre métodos, puede calificarse de ligera.

Tabla 3. Eta-cuadrado para los ANOVAS realizados sobre las diversas variables dependientes consideradas. Se ofrecen tan sólo los resultados para las soluciones bien definidas, y por lo tanto sobre el modelo completo, ya que el modelo de unicidades correlacionadas resultó siempre en soluciones mal definidas.

	Correlaciones entres rasgos	Saturaciones de rasgo	
Factores	*r*	ME	RMSR
M	.003	.015	.104
C	.025	.031	.142
M × C	.012	.015	.027

Notas: M = número de indicadores; C = correlación entre factores de método; r = correlación estimada entre métodos; ME = sesgo; RMSE = raíz del error cuadrático medio. Se redondeó al tercer decimal. Se analizaron sólo los 1175 modelos con soluciones bien definidas.

Saturaciones factoriales de rasgo

Para estudiar el efecto principal del número de indicadores y de la correlación entre métodos y su interacción sobre la estimación de las saturaciones de rasgo, se consideraron dos variables dependientes sesgo y precisión (RMSE) de la estimación. Se realizó un ANOVA entre-sujetos 4×4 para cada variable dependiente. Como en el caso del ANOVA realizado sobre la variable correlación entre rasgo, se utilizaron sólo las soluciones bien definidas, y por tanto sólo se incluyeron en el análisis las estimaciones mediante el modelo completo. Las medias de los efectos principales y de la interacción, para las dos variables dependientes, pueden consultarse en la tabla 2. Con propósitos descriptivos, también se presentan en esta tabla los promedios de estos dos índices para las soluciones estimadas mediante el modelo de unicidades correlacionadas.

El sesgo promedio para todas las condiciones fue de $-.01$, que puede considerarse poco apreciable. La cuantía del sesgo viene afectada por el número de indicadores ($F_{3,1159} = 6.08, p < .001$), por la correlación entre factores de método ($F_{3,1159} = 12, p < .001$), y en menor cuantía por la interacción de ambas variables ($F_{9,1159} = 1.99, p = .037$). El porcentaje de varianza explicada en su conjunto fue del 5.2%. El mayor efecto corresponde, al igual que en el caso de la correlación entre rasgos, al de la variable correlación entre métodos ($\eta^2 = .031$). Como puede verse en la tabla 2, el modelo completo estima las saturaciones de rasgo (validez convergente) prácticamente sin sesgos cuando los factores de método son independientes. Cuando los factores de método son oblicuos, y a medida que su correlación aumenta hasta moderada-alta, aumenta el sesgo que, como promedio, se sitúa en una diferencia con respecto al valor poblacional de $-.02$. Este sesgo puede considerarse pequeño, incluso en el nivel de mayor correlación entre rasgos, cuando su cuantía resultó mayor. Por su parte, el efecto de interacción y el efecto principal restantes son muy modestos. A destacar, no obstante, que para la condición de cinco indicadores, y prácticamente para cualquier correlación entre métodos, el sesgo es insignificante.

También se realizó un ANOVA entre-sujetos 4×4 sobre la precisión (RMSE), el segundo de los indicadores considerados. El efecto principal del número de indicadores resultó estadísticamente significativo ($F_{3,1159} = 44.92, p < .001$), así como el de la correlación entre factores de método ($F_{3,1159} = 64.03, p < .001$), y la interacción ($F_{9,1159} = 3.59, p < .001$). El porcentaje de varianza explicado por los tres efectos alcanzó un 21.5%. En la tabla 3 pueden consultarse los valores de eta-cuadrado para cada variable independiente y su interacción. A la vista de las medias ofrecidas en la tabla 2, el efecto principal del número de indicadores resulta muy claro, aumentando la precisión de la estimación a medida que se introduce mayor número de indicadores por combinación rasgo-método. El efecto principal de la variable correlación entre métodos también resulta claro, el modelo completo presenta una mayor precisión conforme disminuye la correlación entre los factores de método. En cuanto al efecto de interacción, que resultó apreciablemente más débil que los efectos principales, la estimación es muy precisa cuando el número de indicadores es cinco, aún cuando los factores de método presenten una correlación elevada. No obstante, el sesgo que se produce en conjunto puede, una vez más, considerarse pequeño, incluso para las condiciones en que la precisión es menor.

CONCLUSIONES Y DISCUSIÓN

A la luz de los resultados, el comportamiento de ambos modelos, el completo y el de unicidades correlacionadas, ha sido muy diferente y contrario a parte de la evidencia analítica y de simulación divulgada hasta la fecha. Una primera conclusión general es que el modelo completo ha mostrado mejor comportamiento que el de Marsh para dos o más indicadores por combinación rasgo-método en diseños multirrasgo-

multimétodo, las condiciones bajo estudio. Esta conclusión debe complementarse con otras de carácter específico que se desglosan a continuación para cada uno de los modelos.

Los problemas encontrados en el modelo de unicidades correlacionadas son severos y están asociados con las soluciones impropias que parecen indicar problemas de identificación y/o estimación. En estas condiciones los resultados no deberían interpretarse, recomendación razonable dada la tendencia del diseño multirrasgo-multimétodo a generar este tipo de problemas. Resultados empíricos previos no hacían predecible este comportamiento tan inadecuado del modelo de unicidades correlacionadas, que venía siendo utilizado, precisamente, cuando el modelo completo presentaba este tipo de problemas. Analíticamente, no obstante, este comportamiento es comprensible. Tomemos como ejemplo el caso de dos factores de rasgo y dos de método con cinco indicadores por combinación. En este caso, para cuantificar los efectos de método en el modelo completo es necesario estimar veinte parámetros, saturaciones factoriales en este caso, diez por cada factor de método. Si se quiere valorar estos mismos efectos de método desde el modelo de unicidades correlacionadas se deben estimar noventa parámetros, cuarenta y cinco covarianzas de error por método. En estos casos, puede esperarse que aparezcan problemas de estimación, sobre todo en las varianzas de error, que pueden ofrecer fácilmente soluciones negativas, como así ha ocurrido. Como un revisor acertadamente comentó, pueden adoptarse distintas medidas para disminuir la sobreparametrización del modelo de unicidades correlacionadas. Por ejemplo, las covarianzas entre residuales pueden asumirse iguales dentro de cada combinación rasgo-método, etc. Implicando una disminución en el número de covarianzas entre residuales libres a estimar, el número de restricciones varía en función de los indicadores disponibles por combinación rasgo-método. Estas restricciones (fijar a cero las varianzas de error negativas, por ejemplo) son también una práctica habitual en el caso del modelo completo y conseguirían aumentar el porcentaje de soluciones bien definidas. Como solución a posteriori, es sólo recomendable tras descartarse otras alternativas como el empleo de otros modelos, mayor número de indicadores, mayor tamaño muestral, etc. El objetivo de este trabajo era evaluar estos modelos tal y como se definen a priori, sin restricciones, siendo un problema empírico determinar la adecuación de las restricciones en cada aplicación. Para los datos simulados en el presente trabajo no existía tal problema, puesto que los modelos se estimaron con distintos valores para las covarianzas entre unicidades de una misma combinación rasgo-método. Aunque no es aplicable al presente estudio de simulación, el uso de esta constricción en modelos como el de unicidades correlacionadas puede ser una buena alternativa en situaciones aplicadas cuando ambos modelos, completo y de unicidades, ofrezcan soluciones mal definidas. En cualquier caso, una minuciosa evaluación del ajuste ofrecerá datos sobre la plausibilidad de las restricciones empleadas.

Aunque al producirse soluciones mal definidas no deberían interpretarse los resultados, puede ser de interés evaluar el sesgo y precisión de la estimación también para

el modelo de unicidades correlacionadas. Un primer resultado llamativo es que con el modelo de unicidades correlacionadas se produce una sobreestimación, tanto en las correlaciones entre rasgos como en las saturaciones factoriales. Por tanto, la evaluación de la validez convergente de las medidas se ve incrementada, apareciendo los rasgos más solapados de lo que en realidad están. Si se analiza en mayor profundidad lo que ocurre en las estimaciones al manipular las dos variables independientes, se halla un patrón de comportamiento claro del modelo. A la vista de los promedios de la tabla 2, puede decirse que el modelo de unicidades correlacionadas sobreestima por igual sin importar el número de indicadores por combinación rasgo-método, o en cualquier caso las diferencias en sesgo son mínimas. Este no es el caso para los diversos niveles de correlación entre métodos. Los sesgos en la estimación aumentan a medida que aumenta la correlación entre métodos. Este resultado era esperado ya que el modelo de unicidades correlacionadas se basa en el supuesto de métodos no correlacionados.

En cuanto al modelo completo cabe destacar varios resultados. Los problemas de identificación y/o estimación han ocurrido, pero con menor frecuencia de la esperada en función de los resultados de la literatura. Se encuentran en ésta porcentajes de soluciones mal definidas que superan el 77%, mientras en el presente caso de dos o más indicadores por combinación rasgo-método, éstas, en su conjunto, disminuyen hasta un 26.56%. Este resultado es tanto más interesante cuanto que baja el porcentaje de soluciones mal definidas conforme aumentan el número de indicadores, situándose en tan sólo un 6.75% cuando éstos son cinco. Un segundo aspecto a resaltar es la infraestimación que en general produce el modelo a través de las distintas condiciones exploradas. Esta infraestimación es, no obstante, pequeña, incluso en las condiciones que han resultado más desfavorables para el modelo. El sesgo y precisión de la estimación ha disminuido, en general, conforme aumentaba el número de indicadores por combinación. Este resultado es acorde a la hipótesis planteada y esperado, al menos, por dos motivos: a) en cualquier tipo de análisis factorial confirmatorio aumentar el número de indicadores por factor facilita la identificación, y b) al introducirse mayor número de indicadores se consigue elevar la razón de variables observables a factores. El efecto de la variable correlación entre factores de método ha resultado contrario al esperado por hipótesis. A mayor correlación, la estimación resultaba más sesgada y menos precisa. No obstante, cabría esperar ausencia de efecto de la variable correlación entre factores de método, ya que el modelo permite la correlación entre rasgos y entre métodos. Aparece, pues, la necesidad de explicar este resultado en términos analíticos. Como resultado más aplicable de este estudio, incidir en que el modelo completo ha mostrado un buen comportamiento bajo las condiciones de este estudio, y se recomienda su uso frente al modelo de unicidades correlacionadas.

Los resultados de este experimento han resultado claros y, en algunos aspectos, relativamente sorprendentes. A nuestro entender abren una nueva vía de investigación poco representada en la literatura, a la vez que introducen un interrogante sobre las principales conclusiones establecidas sobre análisis factorial confirmatorio de matri-

ces multirrasgo-multimétodo. El corpus de conocimiento acumulado sobre los diversos modelos de análisis factorial confirmatorio de matrices multirrasgo-multimétodo ofrecía un veredicto claro: de entre los modelos aditivos disponibles, el completo con sus variantes y el de unicidades correlacionadas, era este último el que había que escoger. A la multitud de problemas de identificación y estimación del modelo completo se había unido el buen comportamiento del modelo de Marsh, junto con su robustez frente a incumplimientos de sus supuestos, lo que lo hacían muy versátil para enfrentar una amplia gama de situaciones. Diversos autores concluyeron, por tanto, que el modelo de Marsh resolvía de forma más adecuada los diseños multirrasgo-multimétodo y era el modelo a elegir. Esta evidencia sólo se mantiene, claro está, mientras se mantengan aproximadamente las mismas condiciones en que se realizaron los estudios. Un aspecto común a todos ellos era la presencia de un único indicador por combinación rasgo-método. La evidencia del comportamiento cuando se presentan dos o más indicadores provenía de dos fuentes. Por un lado la indirecta, ya que en la técnica de análisis factorial confirmatorio, en su sentido más amplio, era conocido que a mayor número de indicadores era tanto menos probable encontrar problemas de identificación y estimación. Y en segundo lugar, los resultados de estudios aplicados en que no se apreciaban mayores problemas de identificación y/o estimación al aplicar el modelo completo. Toda esta evidencia resultaba, no obstante, inconexa y parcial, sugiriendo la necesidad de estudios de simulación, con una selección precisa de variables y un contexto controlado. Aunque deba complementarse con nuevos estudios de simulación que exploren otras condiciones, este trabajo ha intentado contestar de forma precisa a las hipótesis planteadas, abriendo nuevos interrogantes en un campo en que aparentemente las respuestas estaban bien establecidas.

REFERENCIAS

Anderson, J.C. & Gerbing, D.W. (1984). «The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis». *Psychometrika*, 49, 155-172.

Andrews, F.M. (1984). «Construct validity and error components of survey measures: A structural modeling approach». *Public Opinion Quarterly*, 48, 409-442.

Bagozzi, R.P. (1993). «Assessing construct validity in personality research: Applications to measures of self-esteem». *Journal of Research in Personality*, 27, 49-87.

Bagozzi, R.P. & Heatherton, T.F. (1994). «A general approach to representing multifaceted personality constructs: application to state self-esteem». *Structural Equation Modeling*, 1, 1, 35-67.

Bentler, P.M. (1995). *EQS structural equations program manual.* Encino, C.A.: Multivariate Software, Inc.

Boomsma, A. (1982). «The robustness of LISREL against small sample sizes in factor analysis models». In K.G. Jöreskog and H. Wold (Eds.), *Systems under indirect observation: Causality, structure, prediction (Part I)*. Amsterdam, The Netherlands: North-Holland.

Brannick, M.T. & Spector, P.E. (1990). «Estimation problems in the block-diagonal model of the multitrait-multimethod matrix». *Applied Psychological Measurement*, 14, 325-339.

Campbell, D.T. & Fiske, D.W. (1959). «Convergent and discriminant validation by multitrait-multimethod matrix». *Psychological Bulletin*, 56, 81-105.

Hontangas, P., Oliver, A. y Tomás, J.M. (1997). «Estudio de efectos de método en la validez factorial del STAI-estado». *VII Conferencia Española de Biometría*. Córdoba, 21-24 de septiembre.

Jöreskog, K.G. (1971). «Statistical analysis of sets of congeneric tests». *Psychometrika*, 52, 99-111.

Jöreskog, K.G. (1974). «Analyzing psychological data by structural analysis of covariance matrices». In R.C. Atkinson, D.V. Krantz, R.D. Luce & P. Suppes (Eds.), *Contemporary developments in mathematical psychology* (Vol. 2, pp. 1-56). San Francisco CA: W.U. Freeman.

Kenny, D.A. (1976). «An empirical application of confirmatory factor analysis to the multitrait-multimethod matrix». *Journal of Experimental Social Psychology*, 12, 247-252.

Kenny, D.A. (1979). *Correlation and causality*. New York, Wiley.

Kenny, D.A. & Kashy, D.A. (1992). «Analysis of the multitrait-multimethod matrix by confirmatory factor analysis». *Psychological Bulletin*, 112, 165-172.

Kumar, A. & Dillon, W.R. (1992). «An integrative look at the use of additive and multiplicative covariance structure models in the analysis of multitrait-multimethod data». *Journal of Marketing Research*, 29, 51-64.

Marsh, H.W. (1988). «Multitrait-multimethod analyses». In J.P. Keeves (Ed.), *Educational research methodology, measurement and evaluation: An international handbook*. Oxford, Pergamon Press.

Marsh, H.W. (1989). «Confirmatory factor analysis of multitrait-multimethod data: Many problems and a few solutions». *Applied Psychological Measurement*, 13, 335-361.

Marsh, H.W. (1993). «Multitrait-multimethod analyses: inferring each trait-method combination with multiple indicators». *Applied Measurement in Education*, 6, 49-81.

Marsh, H.W. (1996). «Positive and negative global self-esteem: A sustantively meaningful distinction or artifactor?». *Journal of Personality and Social Psychology*, 70, 810-819.

Marsh, H.W. & Bailey, M. (1991). «Confirmatory factor analysis of multitrait-multimethod data: A comparison of alternative models». *Applied Psychological Measurement*, 15, 47-70.

Molenaar, N.J. (1986). *Formuleringseffecten in survey-interviews.* Amsterdam, VU-Uitgeverij.

Saris, W.E. & Andrews, F.M. (1991). «Evaluation of measurement instruments using a structural equation modeling approach». In P.P. Biener, R.M. Groves, L.E. Lyberg, N. Mathiowtez & S. Sudman (Eds.), *Measurement errors in surveys* (pp. 575-599). New York: John Wiley & Sons.

Saris, W.E. & van Meurs, A. (1990). *Evaluation of measurement instruments by meta-analysis of multitrait-multimethod studies.* Amsterdam: North Holland.

Schmitt, N. & Stults, D.N. (1986). «Methodology review: Analysis of multitrait-multimethod matrices». *Applied Psychological Measurement*, 10, 1-22.

Tomás, J.M. & Oliver, A. (1999). «Rosenberg´s self-esteem scale: two factors or method effects». *Structural Equation Modeling*, 6, (1), 84-98.

Wothke, W. (1996). «Models for multitrait-multimethod matrix analysis». In G.A. Marcoulides & R.E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques.* New Jersey, LEA.

Widaman, K.F. (1985). «Hierarchically nested covariance structure models for multitrait-multimethod data». *Applied Psychological Measurement*, 9, 1-26.

ENGLISH SUMMARY

EFFECTS OF NUMBER OF INDICATORS PER FACTOR ON IDENTIFICATION AND ESTIMATION OF CONFIRMATORY FACTOR ANALYSIS MODELS

A. OLIVER
J. M. TOMÁS
P. M. HONTANGAS
Universitat de València*

The multitrait-multimethod (MTMM) matrix is a research design with a long tradition in Psychology. However, data analyses of such design have been under controversy. It seems that confirmatory factor analysis (CFA) provides a solid base for the analysis of the MTMM matrix. Among the different CFA approaches to this design, the correlated traits correlated methods (CTCM) model, that was first presented, and the correlated traits correlated uniqueness (CTCU) model, which provides an alternative to the main problems of the first model, have received attention in the literature. Results for these two models have been based on a classic MTMM design, with a single indicator per trait-method combination. The present investigation simulates MTMM data with multiple indicators per trait-method combination, and tests for bias in the estimates of the CTCM and CTCU models. The CTCM model performed better than the CTCU model. The amount of underestimation was trivial. Bias slightly increased when method intercorrelations increased.

Keywords: Estimation, bias, confirmatory factor analysis (CFA), multitrait-multimethod (MTMM) matrix, number of indicators, correlation between methods.

AMS Classification: 62H25, 62P15

This research was partially funded by project PB96-0791 from Programa Sectorial de Promoción General del Conocimiento., DGES (MEC) and by SEUID-98 (MEC)-Royal Society of London. We thank a reviewer for helpful suggestions. A preliminary version of this paper was presented at *V Congreso de Metodología de las Ciencias Humanas y Sociales* (5th Congress of Methodology for the Behavioural and Social Sciences) held in Sevilla, Spain, 23-26 September 1997.

* Àrea de Metodologia de les Ciències del Comportament. Facultat de Psicologia. Universitat de València. Av. Blasco Ibáñez, 21. 46010 València. España. E-mail: oliver@uv.es.

– Received November 1998.
– Accepted February 1999.

Campbell & Fiske (1959) proposed the multitrait-multimethod (MTMM) design to analyze the convergent and discriminant validity of psychological measures. Altogether with the design, they also proposed data analysis tools. Although the design, as formulated by the authors, is still in use, the data analysis has been critisised (for example, Bagozzi, 1993; Schmidt & Stults, 1986; Widaman, 1985). Confirmatory factor analysis has been defended as the most valuable statistical technique for this design (Jöreskog, 1971, 72; Marsh, 1988, 89; Rindskopf, 1983). However, several confirmatory models exist, based on different assumptions, conditions, etc. Among these, the two most popular are the complete model, or correlated traits-correlated methods (CTCM) model, in which method effects are posited as latent variables, and the correlated traits-correlated uniqueness (CTCU) model, in which method effects are posited as covariances among the uniquenesses. Most analytical, empirical and simulation evidence on the adequacy of these models to extract conclusions of the MTMM design is based on matrices with a single indicator (observed variable) per trait-method combination. The use of multiple indicators has been proposed, however, to overcome major problems in the use of confirmatory factor analysis for the MTMM matrices (Marsh, 1993). In general, the use of multiple indicators per trait-method combination is associated to a higher ratio of observed variables to factors, compared to the typical MTMM design when one indicator per combination exist. The main aim of this paper is to compare the ability of the CTCM and CTCU models to adequately estimate key parameters in the MTMM design. Monte Carlo simulation is used to compare several levels of correlation among methods, varying the number of indicators per trait-method combination.

METHOD

EQS 5.1 (Bentler, 1995) has been used to simulate the data. The simulation design was a $4 \times 4 \times (2)$. The first independent variable was the number of indicators per trait-method combination: 2, 3, 4 and 5 indicators. The second independent variable was the correlation among methods: 0, .2, .4 and .6. Finally, the CTCM and CTCU models were used to model the data. The effect of these three independent variables on identification, correlation among traits and trait factor loadings has been studied.

RESULTS

Regarding the nonconvergence and improper solutions, on one hand the CTCM model had less non-convergent and ill-defined solutions as the correlation among methods decreases. The identification problems also decreased as the number of indicators increased. On the other hand, the CTCU model had no convergence problems, although all the solutions found by this model were ill-defined. In respect to correlation among traits the CTCM models had no bias when methods were independent. There was slight negative bias when methods were correlated. The amount of bias increased as the

correlation increased. The CTCU model solutions were not included in the analysis, because there were 100% of ill-defined solutions.

The CTCM correctly estimated the trait factor loadings (convergent validity) for the condition of independent methods. In the presence of oblique methods, there were slight biases. These biases increased as the correlation among methods increased. The amount of bias (although slight at the different levels) decreased as the number of indicators increased. The accuracy of estimates also increased as the number of indicators increased. The accuracy of estimates also increased as the correlation among methods decreased.

CONCLUSIONS

Both models' behaviour (CTCM and CTCU) has been different to that found in the previous literature. A general conclusion is that the complete model shown a better behaviour when multiple indicators per trait-method combination exist. The CTCU model presented severe problems in the estimation process. 100% of the solutions were ill-defined.

These results were counterintuitive, given that the CTCU model appeared in the literature to overcome major estimation problems of the CTCM model, when a single indicator per trait-method combination exists. The CTCM model dramatically decreased the number of estimation problems when compared to previous studies when there were a single indicator per combination.

In summary, for the simulated conditions in this work, the CTCM model seems an adequate analytical tool for the analysis of MTMM matrices if there are several indicators per trait-method combination.

Estadística Oficial

QÜESTIIÓ, vol. 23, 2, p. 263-296, 1999

SERVEIS ESTADÍSTICS. PREPARANT-SE PER AL FUTUR*

I. P. FELLEGI
Statistics Canada**

L'objectiu d'aquesta anàlisi és mirar endavant: cap als reptes als quals s'enfronten les agències estadístiques i com els han d'encarar. S'analitza el context dins el qual han d'evolucionar els instituts d'estadística: les principals forces que modifiquen l'economia i la societat, les mesures polí-tiques emergents que aparentment requereixen informació estadística nova o addicional i els canvis en la naturalesa dels governs que tenen repercussions significatives per als instituts d'estadística.

El gruix de l'article proposa una «estratègia interna» per als instituts d'estadística. Això inclou el desenvolupament de nous sistemes d'informació estadística que vagin més enllà de la tradicional funció d'«observació» dels fenòmens i intentin aportar informació sobre les seves causes. Un altre component clau d'aquesta estratègia ha de ser l'assoliment d'un alt nivell d'adaptabilitat, la qual exigeix iniciatives específiques.

Finalment, l'article explora els elements d'una «estratègia externa»: l'assoliment i el manteniment d'un alt nivell de pertinència; l'objectivitat política i la seva percepció, i la col·laboració internacional.

Statistical Services. Preparing for the Future

Paraules clau: Estadística oficial, reptes polítics, sistemes d'informació, flexibilitat, pertinència, col·laboració internacional.

Classificació AMS: 62P25, 92G99.

* Aquest article és una traducció autoritzada de la conferència d'Ivan P. Fellegi a la «1997 UK Statistics Users Conference» (Londres, 11 de novembre de 1997). La versió original en anglès serà publicada properament al «Journal of Official Statistics»; conseqüentment, en aquesta ocasió no procedeix el resum ampli en anglès, llevat de l'extracte breu que apareix al final de l'article.

** Dr. Ivan P. Fellegi, Chief Statistician of Canada, Statistics Canada, 26-A, R.H. Coats Building, Ottawa, Canada K1A 0T6. E-mail: fellegi@statcan.ca.

Traducció d'Alfons Garín i Llobregat. Supervisió d'Enric Ripoll i Font (Institut d'Estadística de Catalunya).

1. INTRODUCCIÓ: ELS REPTES POLÍTICS

La tasca que m'heu encomanat no és gens fàcil. Implica fer una ullada al futur i mirar de distingir algunes tendències destacades. Per fer-ho més difícil, aquesta anàlisi ha de servir per derivar-ne una estratègia per als instituts d'estadística. Us adonareu que parlo d'«instituts d'estadística» en lloc de «sistema del Regne Unit». Si ja em sembla prou valent abordar la qüestió genèrica, penso que seria una temeritat esperar entendre les vostres circumstàncies nacionals.

Segons la dita, «fer prediccions és molt difícil, especialment sobre el futur». Per això és una sort per a mi poder fer servir el treball d'altres, encara que sigui per coincidència, per una part del meu propi treball. Recentment, el cap del servei públic del Canadà (el «Clerk of the Privy Council») va encarregar a un grup d'importants analistes polítics dels principals departaments[1] (als quals em referiré com a «Policy Group») la realització d'un informe «sobre els aspectes conflictius amb més probabilitat d'aparèixer en la societat canadenca a l'any 2005 a causa de les tendències econòmiques, demogràfiques i socials». Van presentar l'informe a l'octubre del 1996, i sembla adient començar amb un resum de les seves conclusions. Encara que l'informe es va escriure des d'una perspectiva canadenca, crec que les seves troballes tenen rellevància per a la majoria dels països desenvolupats.

Després tractaré algunes tendències de l'entorn polític i de govern que tenen importància per a l'estadística. El gruix de l'informe està dedicat a fer un esbós del que jo crec que podria ser una sòlida estratègia per als instituts d'estadística.

2. PRINCIPALS REPTES POLÍTICS[2]

2.1. Factors principals

El Comitè va identificar cinc factors amb impactes importants en un ampli ventall d'àmbits polítics.

[1] Statistics Canada va jugar un paper important en els fets que van precedir l'encàrrec d'aquest informe i també en la seva preparació. Al Cap del Servei Estadístic li van demanar que presidís un comitè interdepartamental d'alts càrrecs per informar sobre la capacitat d'anàlisi política del govern. Una de les seves recomanacions clau va ser que mantinguessin aquesta capacitat d'acord amb la necessitat de fer anàlisi política seriosa per part dels ministeris i el «Clerk of the Privy Council.» A continuació, es va encarregar al «Policy Group» l'esmentat informe. Aquest es va inspirar extensament en una anàlisi profunda sobre les tendències més importants a llarg termini preparada per Statistics Canada.

[2] Aquesta part de l'article es basa en l'informe «Growth, Human Development, Social Cohesion» (N. del T: «Creixement, desenvolupament humà, cohesió social») preparat per un comitè interdepartamental per a l'oficina del «Clerk of the Privy Council», Canadà, a l'octubre del 1996.

(1) Globalització

Per globalització s'entén la integració de la producció i de la distribució a través de les fronteres nacionals. Gràcies a l'espectacular caiguda del cost del transport de mercaderies i de duanes, i a la sensacional evolució de les comunicacions per ordinador, les companyies multinacionals i empreses associades ara poden operar com a entitats integrades, tot i que les seves operacions s'estenen per tot el món. Això els permet explotar els avantatges relatius de cada lloc sense perdre els tradicionals beneficis d'una forta centralització. El fenomen que s'està donant va més enllà d'un simple augment de les exportacions; s'estan produint acords de producció i de distribució completament nous que fan que les fronteres nacionals tinguin cada vegada menys importància. La globalització no afecta només l'economia, sinó tots els aspectes de la societat i la cultura. De fet, un dels reptes més importants serà obtenir els beneficis de la integració econòmica mantenint la independència en altres camps com la cultura, el medi ambient i els programes socials.

(2) La revolució de la informació

La velocitat de canvi de l'anomenada revolució de la informació va ser molt ben descrita per The Economist[3]: «Si els cotxes s'haguessin desenvolupat al mateix ritme que els microprocessadors en les dues últimes dècades, avui un cotxe normal costaria menys de 5$ i podrien fer 250.000 milles amb un galó de combustible (N. del T.: aproximadament 88.000 km. amb 1 litre).» Aquest desenvolupament al cor de la globalització de la producció econòmica afecta la prestació de serveis d'una manera possiblement més fonamental que a la producció de mercaderies. Tots els sectors han fet enormes inversions en tecnologia de la informació, però de moment (si confiem en les nostres mesures de productivitat) la resposta de la societat envers aquestes inversions ha estat limitada. De tota manera, i de cara al futur, no es perd l'esperança d'assolir, com a fruit de l'esmentada revolució de la informació, importants millores en la productivitat, amb la conseqüent millora del rendiment econòmic.

No obstant això, la tecnologia afectarà de manera diferent cadascú. Tot això comporta un risc important, com, per exemple, la possible aparició d'una nova i dominant línia de desigualtat social, que separaria aquells amb un gran domini i facilitat d'accés a la tecnologia dels que no tenen els recursos necessaris per beneficiar-se'n.

La tecnologia també afecta la cultura i la cohesió social, i promou la formació de grups d'interès que transcendeixen les fronteres nacionals. La identificació amb entitats tradicionalment unificadores com la nació o la ciutat és cada cop més feble.

[3]Article aparegut al número del 28 de setembre del 1996 de The Economist, titulat «The hitchhiker's guide to cybernomics», a la pàgina 8.

(3) Problemes mediambientals

Tots som conscients d'una sèrie de problemes del medi ambient, com ara el forat a la capa d'ozó, l'escalfament del planeta o el col·lapse de la pesca en determinades àrees tradicionals. Però la nostra consciència no va acompanyada d'un coneixement profund ni dels riscs reals ni de les connexions entre els canvis en l'activitat socioeconòmica i el medi ambient. Això fa que resulti particularment difícil trobar un equilibri entre els objectius econòmics, socials i mediambientals.

(4) Problemes demogràfics diversos

La natalitat a la majoria de països desenvolupats està per sota de la taxa de reemplaçament generacional. Aquest fet, combinat amb l'increment continuat de la longevitat, produeix societats en les quals el pes de la tercera edat es nota cada cop més. Aquesta tendència s'accelerarà, al menys temporalment, quan la generació del «baby boom» de la postguerra arribi a l'edat de jubilació.

La contínua evolució de la família com a unitat social bàsica suposa una sèrie de reptes encara poc coneguts. No és només que el rol de les dones hagi canviat, ara que la família amb dos sous s'ha convertit en la norma, sinó que altres formes de família comencen a ser freqüents (pares i mares solteres, parelles de fet, parelles del mateix sexe). Els impactes d'aquests canvis no són del tot coneguts però es noten, sens dubte, en àmbits tan importants com la pobresa, el mercat laboral, el desenvolupament dels infants, els plans de pensions i la cura dels malalts crònics de la tercera edat.

La immigració s'ha convertit, no només a Canadà sinó a quasi tota l'àrea de l'OCDE, en la principal font d'increment de la població. Tot i que això, generalment, es considera un factor positiu, la integració de persones procedents de diferents cultures té repercussions de cara a les normes socials, la política i la identitat col·lectiva. La immigració il·legal, que a Canadà no es pot considerar un problema, és, en aquests moments, una font important de problemes per a molts països desenvolupats i una font potencial de futures divisions socials.

Aquests desenvolupaments, i d'altres, causen una sèrie d'impactes acumulatius que s'estenen en el temps fins molt més enllà del present. L'evolució d'un nen fins que es converteix en un ciutadà feliç i productiu, el desenvolupament d'un càncer o d'una malaltia del cor, o la disponibilitat d'una pensió en el moment de la jubilació, són exemples de processos que duren una pila d'anys i impliquen interaccions molt complexes. El repte polític suprem és assolir una millor comprensió dels factors causals, particularment d'aquells susceptibles d'ésser modificats mitjançant programes socials. Una millora significativa de la informació pot suposar una ajuda incommensurable a l'hora d'assolir aquests coneixements necessaris.

(5) Context fiscal

Tots els problemes descrits fins ara, i l'espai polític disponible per tractar-los, interactuen amb el context fiscal; per exemple, l'equilibri necessari entre la reducció del deute de l'estat i els recursos que cal esmerçar per solucionar els problemes existents o emergents.

2.2. Reptes polítics

El «Policy Group» va identificar un nombre de reptes polítics específics, molts d'ells estretament relacionats amb els problemes descrits anteriorment. Van separar aquests reptes en tres grans grups, dels quals em concentraré en dos: creixement i desenvolupament humà. Aquí només puc fer un breu resum del que ells consideraven àrees especialment preocupants per a Canadà.

Creixement

- Creixement econòmic. El creixement del PNB real per càpita i de la mitjana real d'ingressos per família s'ha reduït considerablement en els últims vint anys respecte del que havia estat en dècades passades.
- Productivitat. Malgrat la introducció, durant els últims vint-i-cinc anys, de tecnologia informàtica a gran escala, les taxes de creixement de la productivitat han declinat en alguns dels països més industrialitzats. Els motius d'aquesta caiguda no es coneixen en profunditat, malgrat la importància crucial que té la productivitat de cara al creixement econòmic.
- Ajust a l'anomenada economia basada en els coneixements. Som conscients de que el centre de gravetat de l'economia s'està inclinant envers les indústries que ofereixen serveis i coneixements (i tecnologia). Tant en termes d'ingressos com d'accessibilitat al mercat laboral, l'educació, i possiblement encara més l'alfabetisme, proporciona uns guanys considerables a l'individu, guanys similars als que obtenen les empreses innovadores i amb capacitat d'adaptació. Però, un cop més, és molt limitada la nostra comprensió de les relacions entre diferents factors, com ara l'adquisició d'habilitats per part de l'individu, les pràctiques empresarials en els negocis, la creixent aparició de noves formes d'ocupació, l'ús de la tecnologia, la innovació i la productivitat.
- Factors mediambientals. Aquests interactuen amb l'economia i amb molts altres camps. Hi ha qüestions bàsiques, no gaire ben enteses, que es descriuen amb certa ambigüetat sota l'encapçalament de desenvolupament sostenible.

Desenvolupament humà

- Desequilibris en l'ús del temps. El «Policy Group» va cridar l'atenció sobre el que ells van anomenar desequilibris en l'ús del temps. Ara es treballa menys temps en

comparació amb el temps que s'està jubilat. En trenta anys, la proporció entre els dos ha baixat amb un factor de dos, ja que l'edat d'entrada a la força laboral ha anat augmentant mentre l'edat de jubilació baixava. Aquest desenvolupament, que sembla part d'una tendència a llarg termini més que una simple conseqüència del cicle econòmic, causa problemes importants en els sistemes de pensions públics i privats.

La distribució del treball disponible s'està polaritzant cada cop més: en els últims vint anys el nombre de persones que treballen tant a temps parcial com a jornada completa ha anat augmentant regularment. L'atur a gran escala coexisteix amb l'absoluta manca de temps lliure que pateixen principalment parelles joves de treballadors amb fills, fet que pot tenir conseqüències importants però molt poc conegudes sobre la qualitat de l'atenció que reben els seus fills i els seus pares quan són vells.

- Problemes del mercat laboral. Hi ha signes visibles de canvis importants en els mercats laborals. En la majoria de països industrialitzats, l'atur ha augmentat entre el cicle econòmic de la postguerra i el següent; hi ha més feines no estandarditzades (a temps parcial, temporals o autònoms); els individus canvien més freqüentment d'estat (estar contractat, a l'atur o en períodes aïllats de formació); hi ha una polarització tant dels ingressos com de les hores treballades, i l'educació i, possiblement encara més, les habilitats concretes estan adquirint una importància creixent. Tot i que es fa un seguiment d'aquests canvis, els factors subjacents que els provoquen són força desconeguts.

- Transicions. A part de les transicions que es donen dins el mercat laboral, hi ha altres canvis clau a la vida com els de casa a l'escola, de l'escola al treball, de la infantesa a la paternitat o maternitat, i del treball a la jubilació. Tenim molt pocs coneixements sobre els factors dels quals depèn que la transició tingui èxit o de la interacció entre les diverses forces en joc.

- Creixent polarització dels ingressos i el paper del sistema de redistribució de la renda. Una troballa important és que, tant als Estats Units com al Canadà, la polarització dels ingressos bruts familiars s'ha accentuat durant les últimes dues dècades: aquells que ja gaudien d'uns ingressos alts han incrementat la seva proporció del total, mentre que els que estaven al capdavall del repartiment han vist com es reduïa la seva. Al contrari que als Estats Units, al Canadà aquesta tendència va ser totalment compensada pel caràcter progressiu del sistema redistributiu de la renda mitjançant els impostos (almenys fins ara). Aquestes tendències posen de relleu qüestions importants en relació amb el paper de les xarxes de seguretat social, particularment en un període en el qual la lluita contra el deute té la màxima prioritat.

- Infància i joventut. Un dels temes més complicats és el relacionat amb l'atur entre el jovent. La taxa d'atur entre la gent jove és insistentment alta. I el que és pitjor, aquesta tendència ha persistit tot i que cada cop més joves s'estan a l'escola i cada cop més temps, fet que alleugera la pressió sobre el mercat laboral. Fins i tot quan

tenen feina, els ingressos reals dels treballadors joves han baixat en els últims vint anys, en contrast amb els dels treballadors de més edat. El nostre coneixement de les causes i els possibles remeis per a aquesta situació és molt incomplet. Hi ha qüestions relacionades amb els nens que són encara més desconegudes: per exemple, l'impacte en els fills de l'estatus socioeconòmic dels pares, la mena d'educació que aquests els donen, el sistema educatiu, els professors, el barri, l'entorn o l'herència.

- Salut. Els retalls en els serveis mèdics i hospitalaris, motivats pel ferm control sobre les despeses de l'estat, han concentrat una bona part del debat públic. Malgrat això, hi ha cada cop més evidències de que el sistema sanitari només és un dels factors que afecta la salut de la població, i potser ni tan sols sigui el més important. Alguns dels altres factors estan relacionats amb l'estil de vida, l'estatus socioeconòmic o el medi ambient. Fins i tot dins del sistema sanitari, hi ha casos ben documentats de pràctiques molt divergents (per exemple, en la propensió a practicar operacions de «bypass» coronari o de cesària), amb repercussions econòmiques importants, els impactes de les quals sobre la salut a llarg termini són gairebé desconeguts.

- Pobles aborígens. Hi ha tota una sèrie de qüestions polítiques relacionades amb la població aborigen, que és un tema de primer ordre al Canadà. En gairebé qualsevol escala socioeconòmica els aborígens estan per sota de la resta de la població.

- Llei i ordre. Les enquestes d'opinió demostren que existeix una preocupació creixent entre la població respecte de la delinqüència. Possiblement com a resposta envers aquesta inquietud popular, la taxa d'empresonament està creixent amb molta més velocitat que qualsevol mesura de la criminalitat, tot i que no hi ha cap evidència de que l'augment d'empresonaments faci baixar el nivell de delinqüència.

Aquesta llista no és exhaustiva, ni cal que ho sigui. Espero que sigui suficient per indicar l'ampli espectre de preocupacions que dominen la nostra agenda política. Crec que també demostra que, mentre que el nostre sistema estadístic actual aconsegueix *descriure* a grans trets els fenòmens que ens preocupen, a l'hora d'entendre'ls normalment no ens ajuda prou.

3. TENDÈNCIES DE GOVERN

El paper principal dels instituts d'estadística és, sense cap mena de dubte, d'assistència al procés de l'acció política. En conseqüència, el paper de l'estat, i les expectatives populars que aquest paper pugui generar, tenen una importància crucial per a nosaltres. Per això val la pena revisar breument les tendències més recents i la seva evolució més probable.

Durant gran part del període de la postguerra, les preocupacions econòmiques de gairebé tots els països de l'OCDE eren de caire macroeconòmic, i la seva actitud era in-

tervencionista. La resposta estadística a l'intervencionisme en qüestions fiscals, monetàries, industrials i de comerç va ser el desenvolupament d'un sistema exhaustiu de comptabilitat econòmica amb el suport d'un sistema igualment exhaustiu d'enquestes sobre llars i empreses.

En el camp social, durant els anys de la postguerra es van establir programes universals d'assistència sanitària, d'accés a l'ensenyament secundari, de subsidis per als aturats, de pensions, etcètera. Aquests programes eren relativament menys exigents amb el sistema estadístic. Ens demanaven (eventualment) que en calculéssim el cost, altres despeses importants (per exemple, personal sanitari i docent), relacions clau (nombre d'estudiants per professor, nombre d'habitants per metge), mesures «pures» de rendiment (nombre d'estudiants graduats, nombre d'altes als hospitals), etcètera.

La situació ha canviat gradualment en els vuitanta, i molt més de pressa en els noranta. Hi ha diversos factors responsables d'aquests canvis: un creixement econòmic més lent; unes polítiques macroeconòmiques aparentment ineficaces, i el cost dels programes socials, que quasi invariablement superaven les estimacions inicials, van donar com a resultat uns dèficits creixents. Al mateix temps, la globalització dels mercats financers va incrementar la pressió sobre els estats per «ocupar-se del dèficit.» La qüestió, doncs, es va convertir en prioritat nacional, la qual cosa va comportar retalls substancials en els programes establerts, fet que va conduir inevitablement a preguntes com «què funciona?» i «per a qui és essencial el programa?» El sistema estadístic no es trobava preparat per contestar aquestes difícils preguntes. La reducció de programes estatals sempre és una tasca políticament difícil. Ho va ser encara més enmig d'un període de creixement lent amb una taxa d'atur persistentment alta. La combinació va exacerbar les preocupacions de la població, que ja d'entrada es mostrava escèptica respecte del paper de l'estat i, fins i tot, respecte de l'estat mateix.

Tots aquests processos van posar en marxa una recerca d'evidències dels impactes reals dels programes estatals sobre la societat i l'economia, liderada pels mateixos estats, que volien i necessitaven informació per poder fer la seva pròpia anàlisi abans de prendre decisions polèmiques respecte l'eliminació o modificació dels programes públics establerts. Però també necessitaven informació nova i detallada com a suport objectiu en debats públics. De fet, va aparèixer un nou moviment que demanava que el govern identifiqués i adoptés oficialment indicadors de rendiment, concentrant-se en els resultats en lloc dels processos. La combinació entre un creixent rigor intel·lectual i uns mitjans financers limitats per iniciar nous programes, ha estimulat una visió segons la qual fer les preguntes adequades i assegurar-se de la disponibilitat de la informació necessària per contestar-les és una funció bàsica dels governs.

Aquesta evolució té clares i importants repercussions per als instituts d'estadística. En l'àmbit econòmic, el seguiment de processos macroeconòmics continua sent prioritat. Tot i mantenir el seu interès macroeconòmic, els governs canadencs més recents han parat molta més atenció a consideracions microeconòmiques com, per exemple, com-

prendre els factors que condicionen l'èxit de les empreses i mirar d'apuntalar la intervenció macroeconòmica amb polítiques coherents ideades per ajudar les empreses a nivell microeconòmic.

Si aquesta tasca ja sembla prou descoratjadora, el terreny social encara és més exigent; i ho és per diverses raons. Per començar, els resultats són habitualment el producte d'efectes a llarg termini, com per exemple en els camps de la salut i de l'educació, i que no es poden atribuir amb certesa a una causa única. Com a màxim, podem aspirar a identificar els factors que presentin una tendència a fer variar els resultats en una direcció o una altra. A més, l'interès polític ha canviat: quan abans se centrava en consideracions sobre els impactes socials a grans trets, ara es concentra en els impactes sobre grups particulars (una tasca intrínsecament complicada). I per fer-ho encara més complicat, la ponderació de diferents alternatives polítiques implica inevitablement examinar els seus possibles impactes futurs (una forma de simulació). Efectivament, se'ns demana que proporcionem una base de dades estadístiques prou rica per identificar els instruments que, amb un cost acceptable, puguin donar els resultats desitjats [5].

Els instituts d'estadística oficial estan destinats a tenir molta més relació amb la política, però ho han de fer de manera que es preservi tant la seva independència política com la seva reputació de competència professional. En realitat, això ara és més important que mai, degut precisament a l'augment de l'impacte de l'estadística oficial, combinat amb un escenari en el qual el govern està mirant de convèncer una opinió pública escèptica i insegura de la bondat de les seves accions, fent servir evidències que siguin acceptades com a «objectives» [1].

4. REPERCUSSIONS PER ALS INSTITUTS D'ESTADÍSTICA: UNA VISIÓ GENERAL

Es pot discutir si el «Policy Group» va identificar correctament els factors amb més probabilitat de produir un impacte en la nostra societat canviant, i també si van deduir correctament les àrees concretes de la política a les quals s'haurà de parar màxima atenció. Cal reconèixer que això no seria gaire productiu. Al cap i a la fi, tradicionalment, les societats no han estat gaire encertades a l'hora de preveure els reptes amb els quals s'haurien d'enfrontar al cap de pocs anys. Tot i això, existeixen possibles conseqüències, importants per nosaltres, que ens arrisquem a ignorar.

Tot i que no proposo que s'accepti sense crítica la seva llista específica de reptes polítics esperats, crec que és possible identificar terrenys en els quals es poden anticipar els reptes més importants (el tema de l'atur, el caràcter evolutiu de l'ocupació, la productivitat, les causes d'una bona o mala salut, el paper de l'educació i la formació, etc.). Després de tot, si estem desenvolupant importants sistemes estadístics és per il·luminar un espai ample de la política, i no per donar suport a cap iniciativa política específica. Els

camps a tractar, junt amb la naturalesa de la informació que probablement es necessitarà, ens haurien de servir de guia respecte dels desenvolupaments estadístics que seran necessaris.

El que es desprèn clarament del treball del «Policy Group» és que moltes de les àrees que ells van identificar es caracteritzen per buits importants en la seva comprensió. Per exemple, en Garnet Picot [7] va analitzar diverses mesures governamentals ideades per combatre les altes taxes d'atur, i va demostrar que la seva efectivitat depèn de quina causa, d'entre vàries possibles, considerem responsable del problema. Hi ha unes quantes causes plausibles que hi contribueixen: el pobre creixement econòmic, la manca d'incentius per acceptar llocs de treball poc remunerats (degut al caràcter redistributiu dels sistemes socials i al subsidi d'atur), les altes quotes de seguretat social que fan que resulti molt car contractar nous treballadors, el desequilibri entre les habilitats necessàries i les disponibles, salaris mínims alts, etc. És imprescindible clarificar la seva importància relativa si es volen aplicar solucions polítiques sòlides. Per la seva part, [7] descriu els sistemes d'informació estadística que són necessaris per assolir aquest objectiu, alguns dels quals són força innovadors.

També crec que es pot prendre com a una bona base de treball la suposició de que l'entorn polític evolucionarà seguint, a grans trets, les línies comentades a la secció sobre tendències de govern, amb repercussions clares sobre el caràcter general del servei estadístic que serà necessari.

En resum, per tant, podem preveure els reptes principals. En alguns casos haurem de millorar els sistemes d'informació estadística i en altres haurem de començar a controlar alguns fenòmens que s'han convertit en prioritaris, com el medi ambient. Crec que el repte més gran serà dissenyar nous sistemes d'informació que, més enllà de fer un simple seguiment dels fets, estiguin orientats explícitament a mirar d'esbrinar-ne les causes subjacents. Tot i que segurament podem identificar els terrenys en els quals és necessària aquesta informació, és difícil saber quins temes concrets de la política haurem de dilucidar (amb una gran dosi de perspicàcia però també d'objectivitat). Davant d'aquests reptes la nostra estratègia ha d'ésser audaç, però també sòlida (i, per tant, evolutiva). Els tres ideals que hem de seguir són: pertinència, adaptabilitat i objectivitat.

Aquests tres temes ocuparan la resta de l'article, que es dividirà en les següents seccions: prerequisits i orientacions dels nous sistemes d'informació, essencials per dilucidar qüestions de política o com a font d'informació en els debats polítics; estratègies referents als nostres programes bàsics i la flexibilitat organitzativa necessària per desenvolupar-los, i, per acabar, una secció dedicada al que podríem anomenar «estratègia externa», per assegurar i, a l'hora, protegir la pertinència del programa i la seva objectivitat apolítica.

5. REPERCUSSIONS PER ALS INSTITUTS D'ESTADÍSTICA: NOUS SISTEMES D'INFORMACIÓ

5.1. Elements d'una estratègia

El nostre repte central és intentar desenvolupar nous models de sistemes estadístics, necessaris com a font d'informació per debats polítics en temes clau. El desenvolupament d'una política concreta és un procés complex. Implica els polítics, la gent, els grups amb interessos especials i també els investigadors i analistes polítics de dins i de fora del govern. Les experiències, els punts de vista i la ideologia política de tots aquests grups es combinen per arribar a propostes polítiques. Tanmateix, és essencial que hi hagi evidències, tant empíriques com teòriques, per enriquir i moderar els punts de vista de tots els participants. És necessari que hi hagi un esforç coordinat per mirar d'entendre els factors influents, per poder preveure amb anticipació el resultat més probable de l'aplicació de polítiques alternatives. Això no suposa usurpar el paper del procés polític, sinó ajudar-lo proporcionant informació [5].

No obstant això, cal deixar clar que ajudar a comprendre els fenòmens socials i econòmics és una tasca d'un ordre de magnitud més difícil que limitar-se a controlar els seus impactes. En aquesta secció donaré una visió general dels elements d'una estratègia que hem intentat seguir a Statistics Canada i després hi afegiré alguns exemples il·lustratius de sistemes d'informació nous que hem posat en pràctica.

(i) El desenvolupament s'ha de recolzar en mesures adients de resultats.

Un objectiu clau és ajudar els responsables del procés polític a distingir els ressorts polítics pertinents amb més probabilitat de resultar efectius a l'hora d'apropar-nos a la consecució d'objectius econòmics i socials desitjables. Lògicament, per tant, el desenvolupament de nous sistemes d'informació hauria de començar amb un intent d'entendre quins són els resultats clau que s'intenten assolir en un determinat terreny polític. Això no implica polititzar el sistema estadístic: pel bé de tots, convé que el govern miri d'assolir els seus objectius d'acord amb la millor informació disponible, i que aquesta mateixa informació estigui igualment a l'abast de tots aquells que desitgin promoure polítiques diferents.

Una de les principals dificultats és que en moltes àrees, particularment en el terreny social, no hi ha indicadors àmpliament acceptats que permetin avaluar els resultats. Per exemple, tots estem d'acord en que volem que millori l'educació, però no hi ha un consens generalitzat sobre el què això significa ni sobre com s'haurien d'avaluar els possibles progressos. ¿Volem que el resultat d'una bona educació sigui que la gent tingui èxit en el mercat laboral, que siguin individus madurs, amb la capacitat per continuar formant-se de manera adaptable, que abracin els valors de la bona ciutadania, o alguna combinació de les anteriors? Per molt difícils que siguin aquestes qüestions, són

clarament essencials per al desenvolupament d'un sistema de l'estadística de l'educació realment útil i pertinent.

(ii) Un objectiu principal és connectar les polítiques amb els resultats

El primer pas a l'hora de desenvolupar sistemes d'informació útils és aconseguir indicadors de resultats mesurables i d'alta qualitat. Aquests són fonamentals per controlar el progrés que es fa envers un objectiu, ja que ens diuen si les millores són detectables. Però els bons indicadors per ells sols són insuficients, perquè els objectius fonamentals de la societat, si més no en una democràcia, rarament són assolibles mitjançant la intervenció directa: per exemple, els llocs de treball no es poden crear directament, ni es pot millorar la salut de la població a través de mesures directes. Per tant, els sistemes estadístics realment útils han de permetre als observadors i analistes esbrinar les relacions entre els resultats i les intervencions polítiques: els anomenats ressorts polítics.

No podem donar per suposat que els ressorts polítics tradicionals siguin necessàriament els més importants. Per exemple, cada cop és més acceptat el fet que la dimensió de la classe o la relació d'alumnes per professor no són molt determinants en els resultats dels alumnes. De manera similar, els ingressos familiars poden tenir tant a veure amb la longevitat com les intervencions mèdiques del sistema sanitari. La conseqüència és que els sistemes d'informació d'importància per a la política han de basar-se en visions àmplies dels factors causals relacionats i en com es relacionen els uns amb els altres; en altres paraules, s'haurien de basar en un marc conceptual.

(iii) Els marcs conceptuals són indispensables

No hi ha gaire acord respecte del que constitueix un marc conceptual adequat per a un sistema estadístic. Un marc conceptual no és, per ell mateix, una teoria. És, més aviat, una reflexió, construïda amb molta cura, de la nostra comprensió actual dels factors que tenen un impacte potencialment significant sobre els indicadors que hem seleccionat. Crec que una de les seves característiques essencials ha de ser l'intent de reflectir, si més no esquemàticament, les forces que actuen en un terreny determinat, incloses les seves interaccions i la direcció dels seus efectes.

Un marc conceptual útil no és ni abstracte ni estàtic. El fet d'estar dinàmicament acoblat a un sistema de mesura li proporciona la seva utilitat empírica i la seva adaptabilitat. Per una banda, els marcs conceptuals haurien de guiar l'evolució d'aquells sistemes d'informació que quantifiquen les interaccions mostrades pel marc. D'una altra banda, els sistemes d'informació haurien de tenir una profunda influència sobre el marc conceptual en evolució: haurien de conduir a l'eliminació de les relacions insignificants o no pertinents (per exemple, del dogma ideològic!), i a l'elaboració continuada de les més importants. A més, l'anàlisi de dades podria descobrir relacions completament noves i amb el temps pot conduir a una revisió del marc.

(iv) La cooperació és una condició necessària per al progrés.

Aquesta mena d'evolució estadística representa una tasca molt ambiciosa. L'esforç queda justificat per la necessitat aclaparadora de millorar el nostre coneixement de les forces que hi ha darrere dels problemes econòmics i socials més peremptoris. Aquesta condició és clarament necessària per aconseguir polítiques i programes més efectius. El cost, tot i que no és negligible, és insignificant al costat del cost dels programes estatals destinats a tractar problemes que no es comprenen.

Aquest esforç no pot tenir èxit sense una col·laboració intensiva a tres bandes que inclogui el departament del sector interessat, l'institut d'estadística i la comunitat d'estudiosos de les ciències socials. El recolzament per part del departament ha de prendre almenys tres formes: recolzament moral, suport econòmic directe (o, alternativament, suport per als requisits econòmics de l'institut d'estadística), i col·laboració efectiva per explotar els models i els sistemes d'informació existents durant el decurs de l'anàlisi dels programes estatals existents i de les opcions polítiques.

La participació dels sociòlegs és fonamental per assegurar que les teories predominants es posin de relleu, la qual cosa ajudarà a dissenyar els instruments que il·luminaran els fenòmens relacionats, provant i modificant teories. El seu treball analític hauria de jugar un paper importantíssim a l'hora de destacar quina part de la informació necessària per provar les teories existents falta. Al mateix temps, les ciències socials es beneficien de la disponibilitat d'informació quantitativa que es pot fer servir per separar les especulacions sense fonament de les hipòtesis validades empíricament.

Naturalment, el paper dels instituts d'estadística en aquesta mena de col·laboració a tres bandes és molt significatiu. Han de dur la iniciativa a l'hora de convèncer els que han de prendre les decisions de la importància crítica que té per a ells mateixos posar en marxa el procés necessari a llarg termini per obtenir la informació precisa en el moment adequat. També han de prendre la iniciativa a l'hora d'identificar i, si es troba finançament, desenvolupar els sistemes d'informació necessaris, amb l'ajut d'analistes polítics i sociòlegs. L'experiència ens diu que, en terrenys com la salut, l'educació, el mercat laboral i els ingressos, aquests sistemes tendeixen a procedir, cada cop més, d'enquestes longitudinals o de dades administratives. Les dades longitudinals permeten associar resultats amb una gamma de variables causals possibles molt millor que les dades puntuals. Aquesta condició és indispensable per identificar relacions causals.

És fonamental que hi hagi la col·laboració necessària i que sigui productiva. La iniciativa per començar el procés pot arribar de qualsevol part, incloent-hi els estadístics. No cal que esperem que els altres ens vinguin a demanar ajuda. Al contrari, hauríem de tenir un programa analític ben desenvolupat, que ens permetés assenyalar els buits de dades, i també aquelles àrees de la política pública on no és possible «prendre decisions basant-se en l'evidència» per culpa de la manca de marcs conceptuals i de sistemes d'informació de suport.

A Statistics Canada hem engegat, en col·laboració amb els nostres companys acadèmics i de govern, una sèrie d'iniciatives per desenvolupar sistemes d'informació ideats per aportar una millora significativa en el coneixement i la comprensió de terrenys importants dins la política pública. Més endavant descriuré breument alguns exemples.

(v) El més pràctic i útil és abordar el problema per sectors

Malgrat que cada cop es reconeix més generalitzadament que els fenòmens socials i econòmics estan estretament interrelacionats, hi ha raons pràctiques per desenvolupar els nous sistemes d'informació seguint línies funcionals o sectorials. Hi ha ministeris diferents i altres institucions que tenen polítiques diferents en els camps dels recursos humans, mercats laborals, comerç, indústria, benestar, salut, educació, justícia, etc. Els sistemes d'informació només poden ser d'interès polític quan el destinatari és l'electorat, la funció del qual és jutjar les conseqüències que es deriven dels resultats. Però, més enllà d'aquesta consideració «mundana», el fet és que les mesures dels resultats quasi sempre es formulen d'acord amb objectius concrets per sectors[4]: reduir l'atur, millorar la salut de la població, millorar l'efectivitat del sistema educatiu, etc. Si pretenem que els sistemes d'informació ens ajudin a avaluar els nostres resultats, els hem de desenvolupar específicament per millorar el nostre coneixement del sector objecte [5]. Sabem que hi ha interaccions importants entre sectors: per exemple, és sabut que el nivell d'ingressos i l'educació afecten la salut, i també que la salut afecta l'educació. Però aquests efectes es poden acomodar dins els models sectorials com a variables exògenes.

5.2. Exemples de nous sistemes d'informació a Statistics Canada

Statistics Canada està involucrat en el desenvolupament de nous sistemes d'informació en diversos terrenys, força coincidents amb les àrees de repte polític identificades pel «Policy Group». Una característica comuna a aquestes iniciatives és l'objectiu d'il·luminar un terreny polític concret, per exemple identificar les principals forces que hi actuen i mesurar la seva importància relativa. Tot i que hi ha un cert consens sobre quines són aquestes forces, la tasca principal consisteix en quantificar la seva influència relativa sobre els resultats que ens interessen. Aquest és el cas, per exemple, en l'àrea de la dinàmica del mercat laboral i el nivell d'ingressos. En altres casos es necessita un esforç inicial important per esbossar un marc conceptual que posteriorment es desenvolupa mitjançant dades i es va modificant amb l'ús. És el cas de la salut i de la

[4]La paraula «sector» s'utilitza aquí per denominar un terreny subjecte de polítiques i programes estatals concrets: salut, educació, mercats laborals, polítiques macroeconòmiques, etc. Habitualment, s'assigna tota la responsabilitat, o almenys el paper principal, respecte de les polítiques i programes del sector a un departament o ministeri del govern.

ciència i la tecnologia. En el cas de l'educació i el desenvolupament dels infants, només vam poder identificar una llista, raonablement exhaustiva, del que nosaltres i els nostres col·legues vam considerar factors *possibles* (sense fer consideracions sobre les complexes interaccions que es produeixen entre ells). En aquest cas vam decidir començar fent una enquesta molt extensa, esperant poder elaborar un marc conceptual més complet mitjançant l'anàlisi de les dades obtingudes.

La majoria d'aquestes iniciatives prenen la forma d'enquestes longitudinals. Això no sorprèn, doncs les enquestes puntuals serveixen per observar i controlar els fenòmens, però només es pot arribar a correlacionar els diferents resultats (condició indispensable per poder analitzar la importància relativa de ressorts polítics alternatius i altres variables exògenes) mitjançant enquestes longitudinals.

(i) Dinàmica familiar, de treball i d'ingressos

En el terreny socioeconòmic, el principal problema de la majoria de països del G7 és l'alta taxa d'atur, amb la pobresa que se'n deriva, i la possibilitat que aparegui una classe de marginats permanents (els atrapats en el «pou de la pobresa»). En general, la clau del problema és la relació entre els ingressos, la participació en el mercat laboral i les circumstàncies personals i familiars. Hi ha una sèrie de qüestions importants: en quines circumstàncies pot una família pobra fugir de la pobresa? Quines característiques personals i quins programes estatals semblen ajudar més els pares i mares solteres a reeixir? En quines condicions poden els joves aturats, especialment si no han rebut educació postsecundària, trencar el cercle viciós «cal tenir experiència per trobar feina, però l'experiència només s'adquireix treballant»? Quins són els factors responsables de l'èxit o el fracàs en les transicions escola-treball, treball-atur, treball-jubilació?

Al Canadà hem posat en marxa un programa continu per mirar d'esclarir aquests temes tan complexes. Es basa en una mostra de famílies, cada membre de les quals serà seguit durant almenys sis anys en els quals viurà diferents experiències dins el mercat laboral, en alguns casos se separarà de la família original i potser formarà una família nova, etc. Les característiques clau del programa són la seva dimensió longitudinal i l'objectiu explícit d'associar causes i efectes, l'estreta col·laboració entre Statistics Canada, el departament implicat i la comunitat acadèmica a l'hora de concebre i desenvolupar l'enquesta i la seva anàlisi, i la intenció explícita de mantenir la capacitat d'evolució de l'enquesta. Es va considerar que, en aquest terreny, la bibliografia existent ja havia identificat els principals factors d'influència i que el que calia eren dades adientment organitzades sobre com aquests factors es manifesten al Canadà.

L'enquesta inclou preguntes bàsiques sobre ingressos, experiències laborals i característiques familiars, però també es deixa espai per qüestions suplementàries que permetin explorar noves hipòtesis sorgides de l'anàlisi de les dades.

(ii) La relació entre el rendiment dels treballadors i els resultats de l'empresa

Hi ha una sospita creixent de que la productivitat només es pot estudiar a nivell microeconòmic i de que certes pràctiques empresarials poden estar relacionades en últim terme amb la producció per treballador. Per exemple, en quina proporció i amb quins efectes s'utilitza tecnologia de la informació en el lloc de treball? Implica l'ús de tecnologia una major inversió en personal qualificat? Corren els treballadors menys qualificats el risc de deixar de ser útils a l'empresa? La flexibilitat del mercat laboral (ús creixent de treballadors subcontractats, eventuals o temporals) contribueix significativament a l'èxit de l'empresa?

Hem dut a terme una experiència pilot per veure si seria factible una enquesta longitudinal continuada d'empreses, amb una submostra dels seus treballadors que serien seguits durant almenys un any després d'acabar contracte amb l'empresa objecte de l'estudi. L'enquesta que volem fer ens donaria informació sobre l'ús de les noves tecnologies, sobre els programes de formació per treballadors, les estratègies empresarials (p.ex., el paper dels departaments de recerca i desenvolupament, l'èmfasi posat en els nous productes, l'expansió a nous mercats geogràfics, la col·laboració amb altres empreses en l'àmbit d'R+D, en producció o en marketing, etc.), les estratègies de mercat laboral (p.ex., regulació de plantilles, reenginyeria de processos, augment de la contractació de treballadors a temps parcial o temporals, increment d'hores extraordinàries en lloc de nous contractes), el posicionament al mercat o els canvis en la composició de la oferta de treball. La informació recollida es relacionarà amb indicadors «objectius» de rendiment empresarial tals com els valors de la producció, les vendes, les exportacions, els beneficis, etc. També relacionarà el comportament de l'empresa i els impactes sobre els seus treballadors: les relacions entre la formació que reben els treballadors i l'adquisició de noves capacitats, entre el grau d'ús de tecnologia i els salaris, entre formació i rendiment dels treballadors i la relació de tots els factors anteriors amb l'estabilitat laboral.

(iii) Enquesta sobre infants

La nostra comprensió de l'educació i el desenvolupament d'un infant és molt incompleta, fins i tot més que la de la dinàmica laboral i salarial. ¿Quines són les influències principals que permeten o impedeixen el desenvolupament de membres de la societat productius i feliços? El meu tercer exemple involucra una enquesta longitudinal molt ambiciosa sobre nens, inicialment de 0 a 11 anys d'edat, que intenta aportar una llum sobre aquesta qüestió, força bàsica. Donat que les investigacions existents indiquen clarament que els factors causals en aquest terreny operen durant llargs períodes de temps, l'objectiu és seguir la mateixa mostra d'infants fins que arribin a ser joves adults (de fins a 20 anys). No obstant això, l'enquesta està organitzada per produir resultats d'importància analítica i política contínuament. Estem recollint una àmplia gamma de variables possiblement relacionades amb el tema: salut de la mare durant l'embaràs, situació socioeconòmica de la família, estil dels pares (naturalesa dels estímuls i les

interaccions entre els pares i els fills), símptomes precoços de problemes emocionals o d'aprenentatge, relacions socials (amb companys i amics potencials), tests escolars, avaluació de l'infant per part del professor i del director de l'escola.

En aquest cas no disposàvem d'un marc conceptual que guiés el desenvolupament de l'enquesta. Tanmateix, treballant en la col·laboració més estreta possible amb importants investigadors i personal acadèmic vam poder identificar una llarga llista de variables que podrien tenir un impacte sobre el desenvolupament d'un infant. El resultat va ser la gamma excepcionalment alta de variables que s'inclouen en l'enquesta. En lloc de començar amb un marc conceptual totalment desenvolupat, pensem anar-lo refinant mitjançant l'anàlisi de les dades de l'enquesta. De fet, hem preparat una sèrie de contractes i altres formes de col·laboració per assegurar l'aprofitament complet de la base de dades.

Com en els exemples anteriors, l'enquesta es pot adaptar de manera que reflecteixi la nostra millora gradual en la comprensió, o simplement per recollir informació addicional de manera eficaç i barata.

(iv) Salut de la població

La majoria dels països del G7 es gasten entre el 7 i el 10 per cent del seu PNB en salut. A més, la salut és un tema permanentment situat entre les principals preocupacions dels canadencs. Tanmateix, en aquest cas els reptes polítics també superen de llarg la nostra capacitat de proporcionar informació que doni suport a la presa de decisions, basades en l'evidència, respecte dels factors determinants de la salut de la població i l'impacte a llarg termini de les intervencions en aquest camp. La nostra tercera enquesta longitudinal està dissenyada per aprofundir en els nostres coneixements en aquest àmbit.

L'enquesta segueix a una mostra d'individus durant un període d'anys encara per determinar. Inclou una sèrie de preguntes bàsiques, sobre l'estat de salut, la incapacitat, l'ús del sistema sanitari, els problemes de salut, la situació familiar, i l'activitat laboral o altres activitats importants. A més, cada enquesta inclou també una sèrie de preguntes més específiques de cada cicle (sent el primer cicle concentrat en la salut mental). També està preparat per vincular registres de l'enquesta amb els arxius provincials de l'administració sanitària, per poder incorporar a la base de dades de la mostra els contactes de les persones de la mostra amb el sistema sanitari.

Abans de desenvolupar l'enquesta vam fer un esforç considerable per desenvolupar un marc conceptual explícit que ens servís de guia. Com en els altres exemples, el disseny de l'enquesta i les preguntes suplementàries incorporades en cada fase s'han desenvolupat en estreta i contínua col·laboració amb els principals responsables federals i provincials en el camp de la salut, i també amb un grup àmpliament representatiu d'assessors.

(v) Ciència i tecnologia

La necessitat d'entendre l'impacte de la ciència i la tecnologia sobre la societat s'ha fet patent en la nostra agenda política. De tot el que s'ha escrit sobre la importància d'una inversió adequada en ciència i tecnologia, molt poc ha estat corroborat. ¿És veritat que hi ha un estret vincle entre la inversió d'un país en ciència i tecnologia i la seva taxa de creixement econòmic? Quins impactes té sobre l'ocupació? I sobre el medi físic? I sobre la cohesió social? Quina és la contribució relativa dels diferents sectors de la societat (estat, universitat, empresa, etc.) a la generació de coneixements i quins en són els resultats? Què podem dir de l'emmagatzematge dels coneixements (tant informalment, al cap de les persones, com formalment en objectes com llibres, disquets, etc.)? Com s'utilitza i amb quins efectes? Com es finança la generació de coneixements?

Per intentar respondre a aquestes preguntes estem seguint diverses pistes. Amb l'ajut d'un comitè consultiu extern, vam començar a desenvolupar un marc conceptual. Mentre es va definint aquest marc, anem revisant la informació existent per avaluar quines parts del marc estan ben recolzades i a on falten dades. Al mateix temps, estem intentant entendre les principals qüestions polítiques que s'haurien de contestar. Això ens permetrà perfilar un programa multianual de desenvolupament de la informació que ens ajudarà a ampliar el nostre coneixement sobre els factors clau que afecten la nostra economia i la nostra organització social. I, mentre es desenvolupa el treball conceptual, ja estem recollint informació que serà clarament necessària: sobre innovació, sobre l'adopció de tecnologies avançades, i sobre els fluxos de coneixements des de les universitats cap a les empreses.

(vi) Productivitat

El meu últim exemple tracta el tema de la productivitat. Col·lectivament, estem fent fortes inversions en les noves tecnologies, representades pels ordinadors i el programari i les telecomunicacions relacionades. En el seu número del 28 de setembre del 1996, The Economist estima que aquesta inversió combinada arriba al 12% del capital, el mateix nivell d'inversió que es va assolir en ferrocarrils durant l'apogeu de l'era del ferrocarril a segle XIX. Normalment, quan es fan inversions importants en una nova tecnologia, s'espera un augment significatiu de la productivitat. Però en la majoria dels països del G7 el creixement de la productivitat ha caigut durant els últims vint anys, comparats amb els vint anteriors. The Economist anomena aquesta troballa «forat negre estadístic.»

Molts creuen que hi ha un problema de mesura, i té a veure amb la mesura del rendiment de diverses empreses de serveis amb un creixement alt: banca, telecomunicacions, assesories, etc. En aquests sectors, el problema és que si definir una unitat de rendiment ja és prou difícil, encara ho és més si es tenen en compte les millores de qualitat a les quals estan constantment sotmesos.

Aquesta àrea també serveix com a exemple de col·laboració internacional. Els instituts d'estadística de diversos països van decidir treballar junts, cada un aprofitant-se de la força relativa dels altres. De fet, els seus representants es van posar d'acord per desenvolupar un marc conceptual i les corresponents «enquestes model» per a cada tipus d'indústria de serveis en particular, sovint en col·laboració amb les empreses líders dels seus respectius països. S'han fet experiències pilot amb aquestes enquestes model en altres països voluntaris i s'han comparat els resultats. D'aquesta manera, no només es milloren les tècniques de mesura, sinó que paral·lelament s'assoleix un subproducte de comparabilitat de dades.

6. REPERCUSSIONS PER ALS INSTITUTS D'ESTADÍSTICA: L'ADAPTABILITAT COM A ESTRATÈGIA CLAU

6.1. El programa bàsic

Encara que alguns terrenys polítics han guanyat importància, la majoria dels problemes dels anys de la postguerra continuen vigents. No hi ha dubte que continuarem recollint dades sobre creixement econòmic, inflació, atur i ocupació, evolució dels ingressos, nivell d'educació, salut, etc. Continua sent intrínsecament important *controlar* aquests fenòmens encara que no tinguem molt clar com fer per *millorar* la nostra actuació en cadascun d'aquests terrenys. No pararem de prendre-li la temperatura a un pacient només perquè no acabem d'entendre la raó de la febre, o perquè no sabem com curar-la. Per exemple, potser no hem descobert com curar l'atur, però la societat ha après a fer algunes adaptacions. Algunes tenen a veure amb el comportament individual. Altres són més programàtiques i intenten alleujar els pitjors efectes (com, per exemple, el subsidi d'atur), o produir millores (com els programes de formació ocupacional). Podem ajustar la dieta del pacient, mirar de fer baixar la temperatura per evitar complicacions secundàries, etc.

Encara que la nostra funció bàsica continuarà existint, hem d'esperar reptes importants, tant vells com nous. Per començar, el repte de trobar la mescla adequada entre la continuïtat i l'adaptació dels nostres conceptes a la realitat canviant continuarà present. Aquí, com en tants altres llocs, la col·laboració internacional és productiva i necessària. Davant de dificultats conceptuals, podem accelerar realment el progrés si unim les nostres respectives forces. A més, allà on els fonaments conceptuals són més febles, la pràctica i la convenció internacional els dóna un grau de legitimitat. Els criteris internacionals resulten encara més importants en el moment en què una proporció creixent dels nostres clients operen en un context global i volen informació que es pugui fer servir com a base per fer comparacions vàlides de l'actuació de diferents països en diferents terrenys.

En segon lloc, i això és el nostre programa bàsic, cal saber on van a parar la major part de les nostres despeses i en quins aspectes hem de millorar la nostra eficiència. I segur que necessitarem tenir estalvis perquè és inevitable que haguem de contribuir al finançament de les noves iniciatives estadístiques. L'experiència em diu que per obtenir finançament del govern és indispensable que se'ns vegi eficients a l'hora de dur a terme el nostre programa bàsic i mostrar una capacitat per identificar i eliminar programes de menys prioritat. Però per fer-ho hem de tenir un sistema de planificació efectiu [3].

El tercer repte respecte de les nostres activitats bàsiques és millorar el nostre programa de difusió de dades, especialment quan afecta les necessitats de la població en general, que en la majoria dels casos rep la informació estadística a través dels mitjans de comunicació. Hem de millorar molt quan es tracta d'*informar* la població sobre les nostres *troballes*, i no només de *publicar dades*. Posar èmfasi en les troballes en lloc de limitar-se a publicar les dades té un impacte enorme sobre la idea que la població es fa de la nostra importància [2].

6.2. Capacitat per realitzar enquestes «ad hoc»

Quan un client demana una informació que l'institut d'estadística no pot aconseguir mitjançant un programa finançat regularment, sorgeix la necessitat de dur a terme enquestes especials. La capacitat per respondre a encàrrecs «ad hoc», donant per descomptat que aquests encàrrecs vinguin acompanyats del finançament necessari, representa una forma important de flexibilitat/adaptabilitat del sistema estadístic. L'encàrrec pot tenir diverses formes, però aquí restringiré la meva atenció a enquestes o enquestes pilot. Tal com s'explica a [4], hi ha raons prou convincents per mantenir una bona capacitat per realitzar enquestes pagades pels clients.

El resultat d'aquestes enquestes especials és que es posa nova informació a disposició del públic, sovint en àrees noves. El fet de cobrar pel desenvolupament d'informació ad hoc ens proporciona una mena de test de mercat: si un departament està disposat a esmerçar una quantitat significativa de diners del seu propi pressupost, és probable que la informació estigui relacionada amb assumptes polítics seriosos. Tenint en compte que tota l'estadística oficial, independentment de la seva font de finançament, ha d'estar sempre a disposició del públic en general, el finançament extern d'informació relacionada amb assumptes polítics és sens dubte d'interès públic[5]

[5]Tot i que és important, i indubtablement beneficiós, tenir una capacitat sòlida per respondre els encàrrecs «ad hoc» dels departaments de clients, hi ha motius importants, tant de política estadística com d'eficàcia, per preferir que les necessitats estadístiques *regulars* dels clients se satisfacin a partir del pressupost regular de l'institut d'estadística.

Altres beneficis de tenir la capacitat per realitzar enquestes per encàrrec són:

- Les enquestes finançades pel client són una forma de respondre a la demanda que no es pot satisfer dins dels ajustats pressuposts de l'institut d'estadística. La flexibilitat en el moment de respondre, per tant, té un impacte important sobre el grau de satisfacció del client amb el sistema estadístic.

- Les enquestes especials fetes per encàrrec estan generalment relacionades amb àrees noves i, per tant, suposen innovació. D'aquesta manera contribueixen a mantenir un entorn obert a noves idees. De fet, algunes de les operacions en les quals hem recuperat la inversió han estat enquestes pilot dissenyades per provar noves maneres d'abordar un tema abans de cercar finançament regular per fer-les.

- Com que els preus inclouen totes les despeses, incloses les indirectes, aquesta mena d'encàrrecs contribueixen a mantenir una capacitat operativa.

- Statistics Canada ha establert dues divisions, totes dues amb capacitat per expandir-se ràpidament, el pressupost de les quals prové en la seva totalitat d'enquestes per encàrrec. El personal treballa en estreta col·laboració amb els principals departaments de clients i han après a «vendre» no només la seva capacitat operativa, sinó també les seves idees. Aquestes divisions desenvoluparan una cultura d'orientació envers el client els beneficis de la qual poden tenir un gran abast.

Com es pot crear i mantenir la capacitat per fer enquestes especials? De fet, qualsevol institut d'estadística pot realitzar aquesta mena de treball, fins a un cert límit, mobilitzant els recursos adients. Tanmateix, crec que hem d'anar més enllà. Hem de crear una cultura organitzativa que rebi aquesta mena de treball amb satisfacció i, fins i tot, que el busqui; sinó, corre el risc de convertir-se en una tasca addicional que s'accepta de mala gana. Els principals mitjans que hem fet servir a Statistics Canada per assolir aquest objectiu són:

- La creació de les dues divisions mencionades, que operen, respectivament, en el camp social/familiar i en l'estadística econòmica. Totes dues entitats s'han d'autofinançar mitjançant treballs per contracte.

- Tota l'organització ha adoptat una orientació de mercat, amb objectius explícits quant a beneficis.

- L'entorn de la direcció de personal i financera encoratja el desplegament de personal per dur a terme tasques específiques durant períodes específics.

- Cal tenir una sòlida capacitat operativa, capaç d'expandir-se i contraure's segons les necessitats. Més endavant tornaré a tractar el tema de la capacitat operativa.

Tot i que fomentem decididament el treball per encàrrec, mantenim unes normes bàsiques:

- No concedim cap privilegi al client quan ens encarrega una feina (per exemple, els resultats mai no poden ser privats).
- Statistics Canada manté un absolut control professional, només subjecte a les necessitats principals del client. El control professional inclou els continguts dels qüestionaris, el disseny de la mostra, l'operació de recollida de dades i el seu processament. També es reserva el dret de fer servir la informació resultant en les seves pròpies publicacions analítiques.
- Statistics Canada no fa enquestes que siguin incompatibles amb el seu mandat. Això inclou enquestes sobre intenció de vot i altres que el públic pugui considerar ofensives.
- Per últim, Statistics Canada no fa la competència al sector privat: es dedica a enquestes a gran escala i de gran qualitat, per les quals està més preparada i per les quals és important portar el segell del professionalisme i la legitimitat de l'institut.

6.3. Elements de capacitat operativa i professional

Si es vol tenir adaptabilitat, és imprescindible mantenir una sòlida capacitat professional i operativa. És especialment important protegir-la durant períodes de reducció de pressupost, perquè la infraestructura representa un blanc molt temptador, ja que el seu debilitament no té un impacte sobre el rendiment visible a curt termini. La capacitat d'investigació, d'anàlisi i de metodologia són especialment vulnerables.

Per recuperar-se de l'adversitat és especialment necessari tenir una sòlida **capacitat professional i analítica**. De fet, sense ella ens podem veure ficats en una espiral descendent de credibilitat i recursos. Necessitarem que el nostre personal professional desenvolupi programes com tan bon punt es presenti l'oportunitat. A més, potser podran crear una demanda informada a través del seu treball analític, que destacarà la importància de la informació estadística i, quan la situació ho demani, identificarà forats importants a la base empírica necessària per donar suport a conclusions significatives. La capacitat analítica de la plantilla professional també és necessària a l'hora de desenvolupar o perfeccionar els marcs conceptuals. Aquests marcs són, a la vegada, indispensables per desenvolupar nous sistemes d'informació, tal com ja s'ha explicat.

La **metodologia** forma part de la capacitat professional essencial de l'Institut. La nostra reputació depèn de la solidesa de la nostra metodologia estadística. Se'm podria contestar que en temps de reducció de pressupostos no cal tenir una gran capacitat per dissenyar enquestes, ja que no és gaire probable que n'encetem de noves. Però nosaltres ens hem adonat que millorant el mètode científic podem contribuir decisivament a millorar l'eficàcia en general. Això es pot assolir mitjançant millores en el disseny de les enquestes i el desenvolupament dels mètodes generals de mesura i de les eines

de processament. Una sòlida capacitat metodològica pot ser destruïda en mesos, però costa anys o fins i tot dècades assolir-la.

Es pot dir el mateix de la barreja d'eines i competència que donen la **capacitat operativa**. Això inclou un directori d'empreses cuidat i ben classificat; un nucli de personal supervisor d'equips de treball, al voltant del qual podem augmentar o reduir el personal operatiu d'acord amb les necessitats; tota la gamma d'habilitats necessàries per mantenir una capacitat informàtica flexible, moguda per la demanda; etc.

No n'hi ha prou amb simplement «preservar» aquestes àrees de capacitat. Cadascuna d'elles ha de preparar-se pel futur adaptant els seus processos perquè puguin utilitzar noves tecnologies i noves metodologies, i adaptar-se a possibles canvis d'actitud dels enquestats.

El que és realment necessari, més enllà d'elements particulars de flexibilitat operativa i professional, és mantenir en tot moment un **esperit d'investigació i d'innovació**. Sense això, cap institut d'estadística pot sobreviure gaire temps. Limitar-se a continuar amb un mateix programa, reduint-lo cada cop que hi ha un retall de pressupost, és la recepta ideal per acabar sent intranscendents. No obstant això, és molt difícil seguir sent innovador en períodes de retalls de pressupost, ja que la innovació ha de competir per un finançament que és necessari per mantenir els treballs importants ja existents. Per això demana una atenció especial. A Statistics Canada tenim un sistema de planificació [3] que ens facilita el procés. Cada any, sigui bo o dolent, destinem del 2 al 3 per cent del nostre pressupost a noves iniciatives. Aquesta quota ajuda a mantenir aquesta curiositat intel·lectual tan característica de les organitzacions sanes. Una part dels fons assignats s'utilitza per finançar enquestes pilot i proves a petita escala que es poden fer servir com a demostració de cara als departaments de clients de com la nova informació podria ajudar-los a l'hora d'anticipar, decidir i controlar polítiques i programes.

Una forma de flexibilitat organitzativa consisteix en facilitar al màxim **la reutilització de les dades**. Entrar a descriure amb detall els elements d'una estratègia ens apartaria massa del tema d'aquest article [2]. Aquí només m'agradaria insistir en la necessitat de tenir en compte tres maneres diferents d'encarar el tema. Per començar, hem de crear i mantenir una única finestra electrònica oberta a tota l'estadística nacional disponible per al públic (o sigui, no confidencial). Aquesta hauria de ser la infraestructura de suport de qualsevol mena de difusió, des de publicacions fins a l'accés a Internet. En segon lloc, darrere d'una base de dades estadístiques agregades a disposició del públic, hauríem de crear i mantenir una base de microdades que abastés totes les enquestes, estigués ben documentada i fos totalment accessible per al personal intern. Per últim, estic a favor de totes les mesures destinades a posar les microdades a disposició del públic (respectant, naturalment, la confidencialitat). Degut a que quasi totes les distribucions importants no segueixen una normal, no hem trobat la manera de publicar les microdades de la majoria de les enquestes econòmiques. Les que sí que publiquem, en la seva majoria, i després d'un tractament adequat, són les enquestes socials i familiars. Això facilita

que siguin utilitzades per investigadors externs, inclosos els departaments, que les fan servir com a entrades en models de polítiques concretes.

Com ja he dit, un **sistema de planificació** és un factor determinant en la flexibilitat organitzativa [3]. La planificació, però, ha de tenir el suport d'un **sistema de comptabilitat de costs** detallat basat en un projecte específic. Aquests dos sistemes són imprescindibles si volem tenir la capacitat de revisar regularment el cost dels nostres productes, d'avaluar la prioritat de cada producte en cada moment, i d'estimar el cost d'afegir o eliminar activitats concretes. Hauria estat extremadament difícil dirigir de manera efectiva la nostra actuació durant els últims quinze anys de retalls pressupostaris regulars, amb dues importants injeccions de capital intercalades, sense la possibilitat d'avaluar regularment les conseqüències generals i econòmiques de modificar la línia del nostre producte.

Possiblement, el factor més determinant de la flexibilitat organitzativa sigui la seva **estratègia de recursos humans** [8]. Les decisions de planificació sempre impliquen una reassignació de recursos, i el personal sempre representa el component més important de totes les despeses d'un projecte. Si no aconseguim redistribuir fàcilment el nostre personal d'acord amb les necessitats, el nostre sistema de planificació perdrà efectivitat. Statistics Canada, com la majoria d'instituts d'estadística, es caracteritzava per una estructura molt vertical quant a l'evolució del personal a dins de l'institut: si començaves treballant en salut, educació, ocupació o indústria, era molt probable que et retiressis treballant encara en el mateix camp.

Fa deu anys aproximadament ens vam adonar que, per diverses raons, no ens podíem permetre continuar pel mateix camí:

- No ens podíem permetre l'assignació de recursos tan rígida que això suposava. De fet, era imprescindible per a nosaltres poder ajustar els nostres programes segons les necessitats del client i del pressupost disponible, però sense la preocupació addicional d'haver de redistribuir gent amb capacitats determinades i intransferibles.
- La reducció continuada dels pressupostos reduïa les oportunitats de progressar i havíem de trobar una manera alternativa de mantenir l'interès i la motivació. Vam trobar que, per a la majoria de la gent, l'oportunitat i el repte d'una nova tasca funcionava prou bé.
- Malgrat la llarga sèrie de reduccions de pressupost, volíem que l'institut tingués un grau de solidesa que li permetés respondre a nous reptes. Això només es podia assolir adquirint una plantilla flexible i ben formada[6], per a la qual acceptar noves tasques fos quelcom normal.

[6]Durant els últims anys hem desenvolupat i posat en pràctica un programa de formació totalment integrat. També hem triplicat les nostres despeses en formació: d'un 1% a un 3% del pressupost total.

La nostra solidesa organitzativa està, en aquests moments, sotmesa a una prova important. Hem rebut una forta injecció de capital per realitzar una expansió molt important del nostre programa d'estadística econòmica. El nou capital representa el 10 per cent del nostre pressupost total, però implica doblar la plantilla en algunes divisions. A més, els resultats són necessaris per a un important programa estatal i, com és habitual, es necessiten amb urgència. Sense els deu anys precedents de rotació de plantilla i de programes de formació a gran escala, simplement no hauríem pogut mobilitzar el personal necessari en el poc temps de què disposàvem.

7. REPERCUSSIONS PER ALS SISTEMES ESTADÍSTICS: ELEMENTS D'UNA ESTRATÈGIA EXTERNA

De la mateixa manera que cal tenir una preparació interna per encarar els reptes futurs, els instituts d'estadística han de tenir també una «estratègia externa». Naturalment, les dues estratègies interactuen i han d'estar harmonitzades.

L'estratègia externa hauria de tenir almenys tres suports. El primer, que s'entén tan bé que no el comentaré aquí, està relacionat amb els nostres «valors bàsics»: mantenir la integritat científica del sistema estadístic; protegir la confidencialitat dels resultats estadístics identificables; respectar les normes de privacitat de la societat i minimitzar la molèstia (especialment a les empreses petites) fent servir els registres de l'administració, el mostreig i altres tècniques estadístiques.

Els altres dos suports de l'estratègia externa són la pertinència i la independència política. Existeix un conflicte potencial entre aquests dos objectius bàsics: com més prop del procés polític se situa el sistema estadístic, més alt és el seu potencial de pertinència (o almenys això és el que es diu); però aquesta proximitat pot fer disminuir l'objectivitat política, o almenys la seva percepció. La solució d'aquest conflicte potencial depèn de les circumstàncies nacionals.

7.1. Assolir un alt nivell de pertinència

Per assolir objectius abstractes com la pertinència cal confiar en mecanismes particulars. Els més importants per Statistics Canada són els següents:

(i) Interacció estreta i productiva amb els nivells més alts de la burocràcia

Jo no estic d'acord amb la teoria de que l'estadística oficial només ha de satisfer les necessitats del govern nacional, ni tan sols de tots els nivells de govern. Però sí que crec que la nostra prioritat hauria de ser proporcionar informació per a l'acció política. El subministrament obert d'informació objectiva sobre temes de política pública beneficia

no només el govern, sinó també l'oposició, els grups d'interès i, de fet, tota la població. Per tant, és molt important estar ben connectats amb les més altes esferes polítiques per tal d'obtenir indicis de les actuals preocupacions i les futures prioritats del govern amb la màxima antelació.

Aquesta interacció estreta i productiva, tan necessària, no es dóna espontàniament. Evoluciona amb el temps d'acord amb disposicions organitzatives i iniciatives personals. Per exemple, al Canadà, el Cap del servei estadístic («Chief Statistician») és un membre de ple dret del consell de viceministres («Permanent Secretary») i participa en les seves reunions setmanals, en concentracions periòdiques per discutir a fons totes les possibles conseqüències de les prioritats del govern, i en freqüents grups de treball per analitzar a fons temes específics. Pertànyer a aquest «club» de viceministres obre oportunitats per presentar determinats temes d'acord amb la informació estadística o per cridar l'atenció sobre nous punts de vista en el moment de la seva publicació. L'objectiu primordial és aconseguir que el procés polític, durant el seu desenvolupament, aprofiti al màxim tots els punts de vista que es poden derivar de la informació estadística. Un objectiu secundari, però gens negligible, és generar, al nivell més alt de la burocràcia, una consciència de la importància de la informació estadística per al procés polític, i de que l'objectivitat apolítica, lluny de disminuir la seva utilitat, la realça.

Malgrat ser imprescindible, la interacció al més alt nivell no és suficient. És essencial que hi hagi també una xarxa d'interaccions bilaterals [4] amb els principals departaments i amb tots aquells que tinguin sistemes de registres amb un interès estadístic potencial, com les autoritats duaneres i d'hisenda.

(ii) Activitats analítiques

Un bon programa analític intern ens ajudarà a entendre millor les necessitats del analistes externs, dins o fora del govern. És important que entenguem aquestes necessitats per poder identificar els buits de dades més importants, aquells que, si s'omplissin, contribuirien decisivament a la bona comprensió de temes polítics clau. Aquestes orientacions són imprescindibles per donar suport a noves iniciatives.

Els bons analistes tenen una gran motivació personal per explorar nous terrenys, i sovint les seves exploracions els porten a desenvolupar dades o a identificar-ne buits importants. En qualsevol cas, ells lideraran la causa de la millora de la informació o d'un marc conceptual més fructífer.

(iii) Una xarxa extensa d'«assessories»

En l'anàlisi final, l'assignació de prioritats és subjectiva. Per això és molt important que la nostra avaluació es basi en una informació àmplia i equilibrada, amb el suport de diversos mecanismes de consulta, formals i informals. Degut a l'estructura federal del Canadà, tenim mecanismes consultius amb les províncies en tots els camps en els

quals estan interessades. Rebem opinió externa experta des d'una dotzena de comitès assessors, cadascun d'ells dedicat a un tema específic. Els nostres executius comptables, que treballen amb els clients més importants, recullen punts de vista addicionals del món empresarial. Mantenim una connexió activa amb les principals organitzacions empresarials i amb representants del sector de la petita empresa. A l'àpex dels nostres mecanismes consultius està el «National Statistics Council», un selecte comitè d'assessors.

(iv) Cooperació amb la comunitat acadèmica

A través de les seves activitats analítiques, la comunitat acadèmica pot derivar algunes conclusions significatives de les bases de dades estadístiques. També pot col·laborar en la construcció de marcs conceptuals; cridar l'atenció sobre la necessitat de nous productes d'informació; revisar plans per a noves enquestes; participar en els comitès assessors, revisar productes analítics, etc. Com passa amb les cooperacions de tota mena, fer que resulti productiva implica un cert esforç. Nosaltres treballem en estreta col·laboració amb ells per conèixer les seves necessitats d'accés a bases de dades estadístiques, donem oportunitats perquè alguns d'ells passin temps sabàtic amb nosaltres, escrivim articles en equip, participem en l'organització i el finançament de conferències científiques, etc.

Al llarg del temps, aquesta col·laboració pot proporcionar a la comunitat acadèmica molts coneixements sobre el sistema estadístic. Això els permet ser més efectius a l'hora de detectar i cridar l'atenció sobre problemes i tendències emergents. Alguns acadèmics tenen contactes freqüents amb els mitjans de comunicació respecte dels temes relatius a les seves especialitats, i per a Statistics Canada és beneficiós que comentin les noves troballes, analítiques o de dades.

(v) Relacions amb els mitjans de comunicació

La pertinença queda determinada no només per la utilitat potencial de la informació que produeix l'institut d'estadística, sinó també pel grau en el qual aquesta informació ajuda a comprendre millor determinats temes. El paper dels mitjans de comunicació és molt important perquè, per a la major part de la població, inclosos molts dels nostres representants electes, aquest és l'únic canal pel qual reben informació estadística. Per aquest motiu és d'interès públic que els mitjans de comunicació informin sovint sobre estadística. Això també beneficia l'institut d'estadística, doncs les referències freqüents i objectives dels mitjans de comunicació als seus productes tenen un impacte positiu acumulatiu sobre l'opinió pública envers l'institut.

L'aspecte més important de les relacions amb els mitjans de comunicació és la necessitat de que cada publicació estadística vagi acompanyada d'un resum analític fàcilment comprensible de les troballes més significatives. Altres possibles mesures de cara als

mitjans de comunicació serien: accés lliure a totes les publicacions de l'institut, que haurien d'anar acompanyades d'un portaveu competent; ser proactius a l'hora de cridar l'atenció sobre errors o problemes amb les dades; respondre per escrit a articles equivocats o que indueixin a conclusions errònies; disponibilitat del personal de certa jerarquia per a ser entrevistat pels mitjans de comunicació; subministrament de detalls d'interès local, quan això pugui augmentar la cobertura per part dels mitjans de comunicació regionals.

7.2. El tema de l'objectivitat política i la seva percepció

El valor de l'estadística per a la societat depèn directament de la confiança que hi hagi envers els que la produeixen. Com que hi ha molt pocs usuaris que realment puguin discutir les estadístiques oficials, la seva disposició per fer-les servir és, en últim terme, una reflexió de la seva confiança en la integritat professional dels estadístics i en la seva capacitat per realitzar la seva feina lliures de qualsevol influència política perniciosa [1]. La importància fonamental de l'objectivitat s'ha fet encara més patent arran de l'escepticisme de la població respecte del procés polític i de la creixent substitució de l'estadística «objectiva» per l'opinió a l'hora de distribuir el minvant finançament estatal.

Vull cridar l'atenció sobre tres temes bàsics, deixant de banda tècniques tan especifiques i útils com avisar amb antelació de la data de publicació en totes les sèries importants, comitès externs de revisió, etc.:

(i) Relacions institucionals

En altres ocasions he comentat [4] els arguments generals a favor i en contra d'un sistema estadístic centralitzat. Tanmateix, és indubtable que la centralització facilita la independència política. Per començar, la protecció d'aquesta independència és una de les principals responsabilitats del cap de l'institut. Com més alta sigui la seva posició dins la burocràcia, millor podrà desenvolupar aquesta funció. Això no ve donat pel poder en sí, sinó més aviat pel fet d'estar més exposat a l'opinió pública, cosa que el col·loca sota «l'amenaça» d'haver de dimitir, amb possibles conseqüències de gran abast. Per tant, des de la perspectiva de l'objectivitat política, és preferible la centralització, ja que quasi sempre assegura una posició jeràrquicament més elevada pel cap de l'institut d'estadística.

Un altre problema estructural bàsic està relacionat amb el caràcter formal de la relació entre el procés polític i el sistema estadístic. Aquí crec que hi ha una relació d'equilibri entre pertinència i objectivitat política. Per una banda, en aquest article s'ha defensat que l'estatus de viceministre (o sigui, cap d'un departament de govern) del cap del servei estadístic té una importància extraordinària de cara a mantenir bones relacions amb

els altres departaments (que al seu temps són factors clau de pertinència a llarg termini). No obstant això, un viceministre respon davant del ministre. En el model de govern de Westminster el ministre és responsable davant del Parlament, però el funcionari públic no.

Certament, una relació de subordinació a un ministre pot produir interferència política. Una alternativa que preservaria els avantatges de la centralització seria que l'institut d'estadística passés a dependre explícitament del Parlament, igual que el banc nacional i els organismes de control del govern en molts països. Això representa la manera més segura d'aïllar l'institut d'estadística de les influències polítiques, però també existeix el risc de que aquest quedi aïllat de la «maquinaria» del govern, perdent pertinència.

Statistics Canada sempre ha depès de ministres que tenien com a principal responsabilitat política l'institut d'estadística. Aquests mantenen una relació distant amb Statistics Canada en tots els temes de política i programes estadístics: totes les qüestions tècniques i de prioritats programàtiques o bé són consultades amb el Cap del servei estadístic o bé se n'encarrega ell directament. Aquesta tradició es veu mantinguda i reforçada pel servei públic i el «Privy Council Office» (el departament personal del Primer Ministre), que entenen perfectament la importància política de tenir un institut d'estadística amb una bona credibilitat.

Tothom és perfectament conscient, a hores d'ara, que la tradició d'independència política és tan fortament arrelada a Statistics Canada que els mitjans de comunicació de seguida se n'adonarien de qualsevol intent deshonest d'interferir.

Amb aquestes circumstàncies tan favorables, la jerarquia «departamental» de l'institut d'estadística només ofereix avantatges. Però en altres circumstàncies, podríem arribar a conclusions totalment diferents, especialment si no es pogués comptar amb els nivells superiors del servei públic per donar suport a la independència política del sistema estadístic.

(ii) Control del pressupost

El control del pressupost és un aspecte bàsic per mantenir la independència política. Si el govern pogués controlar programes estadístics específics a través dels pressupostos, tindria l'oportunitat de sabotejar aquells programes que poguessin donar una mala imatge política de la seva gestió. La mera existència d'aquesta possibilitat ja podria influenciar el comportament d'ambdues parts.

Durant els últims 12-15 anys, Statistics Canada ha experimentat continus retalls pressupostaris. No obstant això, l'institut ha gaudit de completa llibertat per administrar aquestes reduccions. Naturalment, això ha significat que havíem d'estar preparats per defensar les nostres decisions. De fet, la nostra forma de dirigir el procés ens ha fet mereixedors d'una considerable dosi de credibilitat professional i administrativa, que

ha contribuït al nostre èxit posterior en el moment d'obtenir finançament addicional per algunes noves iniciatives força importants.

¿És una contradicció mantenir el control dels pressuposts i cercar finançament per a iniciatives específiques? No necessàriament. El finançament ha estat dedicat a noves activitats a les quals hem assignat, en col·laboració amb altres institucions, màxima prioritat. A més, un cop rebuts, els diners passen a formar part del nostre pressupost general. Tot i que òbviament estem obligats per la nostra paraula a gastar-los en els nous programes anunciats, això no és un requisit formal, ni ho és a perpetuïtat.

(iii) El paper de l'anàlisi substantiva

Presentar un flux de producció analítica objectiva i equilibrada dóna una imatge de professionalitat i d'independència política, i això és essencial pels instituts d'estadística, perquè els ajuda, possiblement més que qualsevol altra cosa, a diferenciar la seva imatge pública de «la del govern.»

La producció analítica d'Statistics Canada pren diverses formes, de les quals la més visible és la que nosaltres anomenem «publicacions insígnia.» Mensual o trimestralment, publiquem anàlisis d'alta qualitat sobre l'estat de l'economia, el mercat laboral i el nivell d'ingressos, sobre les tendències en salut i educació, etc. A part de publicar una àmplia gamma d'informes analítics, la nostra política especifica és que totes les nostres publicacions estadístiques han d'anar acompanyades d'un resum de les troballes més significatives en els terrenys econòmic i social.

Tant l'objectivitat com la pertinència són importants. Objectivitat significa explorar totes les vessants d'un tema, evitant fer propaganda política o caure en prejudicis, destacant les troballes més importants, independentment de la imatge que donin del govern. La pertinència té a veure amb l'elecció dels temes a tractar: aquests han de tenir una importància clara, tot i que alguns puguin ser discutibles. Com totes les nostres publicacions, els estudis analítics inclouen resums de fàcil lectura i són citats amb freqüència pels mitjans de comunicació.

Tot i que un flux regular i visible de producció analítica pot contribuir a la nostra imatge positiva, un programa d'aquesta mena ha d'estar molt ben dirigit. La producció ha d'estar sotmesa a revisió per part dels companys per verificar la validesa científica del mètode analític. Però també ha d'estar sotmesa al que nosaltres anomenem una «revisió institucional,» que comprovi que l'anàlisi és neutral, que explori el tema d'una manera equilibrada i que no ultrapassi la fina línia que separa l'anàlisi de la propaganda.

7.3. Col·laboració internacional

L'últim element de l'estratègia externa que vull tractar és la necessitat de participar en treballs internacionals. Crec que l'escenari internacional és molt més que un luxe opcional. Pren almenys les tres formes següents:

(i) Les funcions internacionals tradicionals

En aquesta categoria hi cau tot el treball prou reconegut per la nostra professió des de fa més de 150 anys: els esforços per harmonitzar conceptes i classificacions, l'estimulació professional mútua, i l'aprenentatge dels uns dels altres. La necessitat de l'harmonització, que sempre havia estat important, ha augmentat espectacularment i ho continuarà fent. Aquest fet es produeix per diverses causes: corporacions transnacionals; negociadors internacionals; organitzacions internacionals que posen els preus i distribueixen els beneficis; investigadors que fan servir les comparacions internacionals com a cotes o punts de referència naturals, i el públic en general, que ja gaudeix d'una facilitat d'accés a les dades nacionals sense precedents, gràcies a Internet.

(ii) La unió dels esforços intel·lectuals

Tot i que la categoria anterior ja incloïa col·laboració, estava relacionada bé amb les interaccions professionals tradicionals, bé amb el treball realitzat per i sota els auspicis d'organitzacions internacionals (com el desenvolupament de classificacions internacionals i de normes). En els últims anys, estimulats per la complexitat de determinats problemes (com la mesura de la producció de les indústries del sector de serveis), hem format diversos grups de treball, informals però estructurats, que es reuneixen regularment. La «taxa d'admissió» és la presentació del treball conceptual dut a terme entre reunions. Molts d'aquests fòrums han resultat força productius.

(iii) Dimensions transnacionals

Hi ha una tercera categoria de treball internacional, les dimensions de la qual encara són borroses, però la necessitat de la qual es veu clarament. Està relacionada amb el problema que representa seguir la pista de les transaccions i contribucions econòmiques de les empreses transnacionals. Cap institut d'estadística nacional no pot dur un control del seu funcionament, ja que operen realment per sobre de les fronteres. Considerem una fàbrica al Canadà, part d'una empresa transnacional, que produeix frens i els exporta a tot el món per ús d'aquesta mateixa empresa. Aquesta fàbrica canadenca informarà dels guanys derivats de l'exportació, amb valor afegit, beneficis, inventari, etc., d'acord amb les convencions de comptabilitat de l'empresa. Però aquests informes poden perfectament canviar al llarg del temps en resposta a la seva avaluació dels beneficis que poden derivar de les diferències en els imposts de cada país.

Per molt complicat que sembli el problema respecte de les mercaderies, ho és encara molt més respecte dels serveis, molts dels quals poden fins i tot travessar fronteres nacionals electrònicament i són, per tant, indetectables. És evident que, en aquest tema de creixent importància, només es poden assolir objectius mitjançant la col·laboració entre instituts d'estadística nacionals en fòrums i fórmules que encara s'han d'articular.

CONCLUSIÓ

A mida que ens acostem al final del mil·lenni, es va posant de moda parlar del futur: identificar les tendències emergents i analitzar amb erudició les seves conseqüències. La meva experiència amb exercicis similars no ha estat favorable: retrospectivament, es pot dir que les tendències que finalment resulten importants no acostumen a ser les que s'havien anticipat. Abunden els exemples de diagnòstics desencertats. Un dels meus favorits és la famosa frase del periodista americà Lincoln Steffens, en tornar d'una visita a la Unió Soviètica al 1919: «He vist el futur: i funciona ...».

Fins i tot quan encertem el pronòstic, les dificultats que ens trobem per respondre als canvis acostumen a ser força diferents del que havíem esperat. Com ja us n'haureu adonat, el resultat és que he preferit amagar-me darrere les observacions d'un grup d'analistes polítics canadencs més o menys anònims, evitant donar la meva opinió personal sobre quins seran els reptes polítics clau del futur més visible.

Podria haver intentat impressionar-vos amb escenaris futuristes de comunicacions per ordinador i amb impactes sobre la societat i l'institut d'estadística; em podia haver desfet en un discurs eloqüent sobre el final dels estats-nació, o sobre tot el contrari (totes dues posicions van ser defensades en el recent número de commemoració del 75è aniversari de Foreign Affairs, ni més ni menys que per autoritats del calibre d'Arthur Schlesinger, Peter Drucker i Anne-Marie Slaughter [6]); em podia haver posat el meu barret d'analista polític i mirar d'esbrinar si la recent reducció del paper de l'estat és un fet puntual o part d'un moviment comparable al d'un pèndol gegant. Prefereixo estalviar-vos-ho. Bàsicament, crec que no només és possible trobar una estratègia prou sòlida per no dependre de la nostra habilitat com a futuròlegs, sinó que és absolutament necessari fer-ho, i jo he intentat comentar els elements interns i externs que ha de tenir aquesta estratègia.

La principal característica de la capacitat interna és la construcció i el manteniment, fins i tot davant de retalls pressupostaris, d'una capacitat professional i operativa sòlida, capaç d'adaptar-se a l'entorn. He comentat les adaptacions que considero més importants en els terrenys amb més importància política. En la majoria dels casos, això implica treballar tant l'aspecte conceptual com les dades, i jo defenso que tots dos s'han de treballar en paral·lel, de manera iterativa, començant amb alguns marcs concep-

tuals i després posant-los en pràctica i millorant-los a través de nous sistemes d'informació.

Els aspectes clau de l'estratègia externa asseguren l'excel·lència dels receptors, que filtren els senyals del nostre entorn. Hem de donar absoluta prioritat a les orientacions necessàries per seguir sent pertinents (la qual cosa implica, entre d'altres coses, mirar d'anar més enllà del nostre paper tradicional de *controlar* per intentar *il·luminar* els temes que tractem, incloent-hi el paper dels «ressorts polítics»). Hem de fer un bon servei a la societat, a tots els grups principals, d'acord amb *les seves necessitats*, no *la nostra conveniència.* Per últim, però no menys important, hem de fer tot el que calgui per conservar i reforçar la nostra independència professional i apolítica.

REFERÈNCIES

[1] **Fellegi, I.P.** (1991). «Maintaining Public Confidence in Official Statistics». *Journal of the Royal Statistical Society*, Series A, 1-22.

[2] **Fellegi, I.P.** (1991). «Marketing at Statistics Canada». *Statistical Journal of the United Nations ECE*, 295-306.

[3] **Fellegi, I.P.** (1992). «Planning and Priority Setting. The Canadian Experience», in *Statistics in the Democratic Process at the End of the 20th Century. Anniversary Publication for the 40th Plenary Session of the Conference of European Statisticians 1992*, edited by Hölder, Malaguerra and Vukovich.

[4] **Fellegi, I.P.** (1996). «Characteristics of an Effective Statistical System». *International Statistical Review*, Vol. 64, Number 2, August 1996, 165-198.

[5] **Fellegi, I.P.** and **M. Wolfson** (1997). «Towards Systems of Social Statistics». *Bulletin of the International Statistical Institute*, 51st Session, Istanbul, 1997.

[6] **Foreign Affairs** (1992). *75th Anniversary Issue*, September/October 1997.

[7] **Picot, G.** (1997). «Statistics and Public Policy: The Case of Labour Markets and Firms». Paper presented at the meetings of the Statistical Society of Canada. Statistics Canada, June 1997.

[8] **Statistics Canada** (1997). *Human Resource Development at Statistics Canada.* Statistics Canada, July 21, 1997.

ENGLISH SUMMARY

STATISTICAL SERVICES. PREPARING FOR THE FUTURE*

I. P. FELLEGI
Statistics Canada**

The objective of the paper is to look forward: to the challenges facing statistical agencies and how to prepare for them. It reviews the context within which statistical offices must evolve: the main forces at work that are modifying the economy and society; the key policy issues that appear to be in need of new or additional statistical information; and changes in the nature of government that have significant implications for statistical offices.

The bulk of the paper proposes an internal strategy for statistical offices. This includes the development of new types of data systems that go beyond the traditional function of «monitoring» and attempt to shed light on the main forces at work. Another key component of the must be to ensure a high level of adaptability. Such adaptability requires specificic initiatives.

Finally, the paper explores the elements of an «external strategy»: achieving and maintaining a high level of relevance; political objectivity and its perception; and international collaboration.

Keywords: Official statistics, policy challenges, data systems, flexibility, relevance, international collaboration.

AMS Classification: 62P25, 92G99.

---

* This paper is an authorised version of the conference by Ivan. P. Fellegi at «1997 UK Statistics Users Conference» (London, 11th November 1997). The original version in English will be published shortly in the «Journal of Official Statistics». Consequently, the full English summary is not appropriated in this ocasion, except for this brief abstract.

** Dr. Ivan P. Fellegi, Chief Statistician of Canada, Statistics Canada, 26-A, R.H. Coats Building, Ottawa, Canada K1A 0T6. E-mail: fellegi@statcan.ca.

QÜESTIIÓ, vol. 23, 2, p. 297-320, 1999

ENLACE DE ENCUESTAS: UNA PROPUESTA METODOLÓGICA Y APLICACIÓN A LA ENCUESTA DE PRESUPUESTOS DE TIEMPO

M. J. BÁRCENA
F. TUSELL
Universidad del País Vasco*

Tratamos el problema de completar dos ficheros con registros conteniendo un subconjunto común de variables. La técnica investigada utiliza árboles de regresión y/o clasificación. Se propone y estudia una extensión para variables respuesta multivariantes, ilustrando su empleo sobre la Encuesta de Presupuestos de Tiempo (EPT-93).

Linking surveys: a method and application to the Survey of Time Uses.

Palabras clave: Enlace de ficheros, inserción de encuestas, imputación, árboles de regresión y clasificación.

Clasificación AMS: 62G05, 62D05.

Este trabajo ha sido realizado como parte integrante de la tesis doctoral del primer autor. Agradecemos la financiación recibida del MEC y de la UPV/EHU (proyecto PB95-0346), los comentarios de asistentes a un seminario del Departamento, particularmente de Vicente Núñez, Eva Ferreira y Karmele Fernández, y de un evaluador. Los errores, carencias o inexactitudes que subsistan son nuestros.

* Departamento de Estadística y Econometría. Facultad de CC.EE. y Empresariales. Avda. del Lehendakari Aguirre, 83. 48015 Bilbao. E-mail: etptupaf@bs.ehu.es.

- Recibido en junio de 1998.
- Aceptado en marzo de 1999.

1. INTRODUCCIÓN

Nuestro punto de partida son dos ficheros, A y B con un total de $N = N_A + N_B$ observaciones. Cada fichero tiene la siguiente estructura:

Fichero A

$X_1, \ldots, X_p$	$Y_1, \ldots, Y_q$	desconocido

Fichero B

$X_1, \ldots, X_p$	desconocido	$Z_1, \ldots, Z_r$

Estos ficheros contienen, en nuestra aplicación, datos de dos encuestas diferentes con un grupo de variables en común, $X_1, \ldots, X_p$, cuyos valores son conocidos para los N individuos. Cada encuesta tiene, además, otras variables específicas que sólo son medidas para los individuos de esa encuesta: $Y_1, \ldots, Y_q$ y $Z_1, \ldots, Z_r$ respectivamente, en nuestra notación.

En este trabajo estudiamos cómo completar los ficheros A y B, utilizando los valores de las variables comunes X: tratamos de imputar las áreas sin sombreado en la Figura 1.

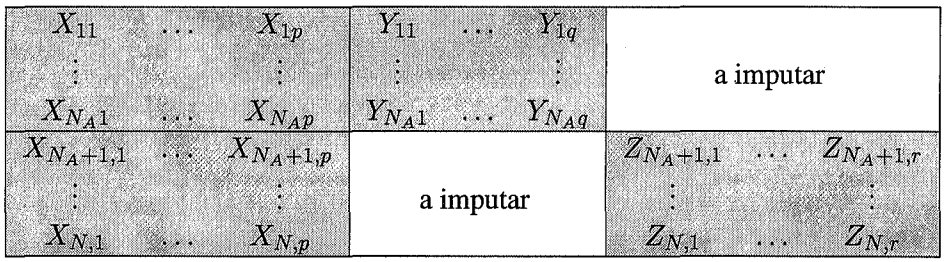


Figura 1. Matriz de observaciones disponibles en un problema de enlace de encuestas.

El objetivo es permitir un análisis más rico de la información al poner en relación las variables específicas de ambas encuestas, es decir, aproximarnos al caso de tener datos completos de (X, Y, Z) para los N individuos. Sin embargo, hay una limitación importante: sólo podemos aspirar a reconstruir la parte de la relación entre Y y Z que puede transmitirse a través de X ya que, con los datos disponibles, es imposible conocer la relación entre Y y Z cuando X es constante. Por tanto, los resultados obtenidos con datos completados se aproximarán más a los que se obtendrían con datos completos, es decir, el enlace será mejor:

1. Cuanto menor sea la relación entre Y y Z cuando X es constante.
2. Si tenemos información adicional sobre la relación entre Y y Z cuando X es constante, y realizamos el enlace teniendo en cuenta esa información.

En lo que sigue veremos una breve descripción de algunas técnicas de enlace de encuestas, desarrollaremos una nueva que emplea árboles de regresión y/o clasificación y aplicaremos esta última a los datos de la Encuesta de Presupuestos de Tiempo, en lo sucesivo designada EPT-93.

La EPT-93 fue realizada por el EUSTAT (Instituto Vasco de Estadística) en otoño de 1992 y primavera de 1993, entrevistando a 5040 individuos. Para cada individuo se recogen sus características personales y el tiempo en minutos que dedica a diferentes actividades (dormir, higiene, trabajo, etc.). La información es recogida mediante un diario que cada encuestado rellena en un sólo día previamente establecido. Como se indica en EUSTAT (1997), pág XII, el diario semanal es una alternativa mucho mejor, pero

> «... su calidad disminuye conforme avanza la semana, los diarios incompletos se multiplican y, además, es imposible de realizar sin un incentivo económico alto.»

Por ello se optó por utilizar cuestionarios de un sólo día, pero distribuyéndolos aleatoriamente en cuatro bloques semanales (de lunes a jueves, viernes, sábados y domingos). Así, para cada individuo se conocen sus características personales, el tiempo dedicado a las actividades que realiza en un día dado y en cuál de esos bloques semanales ha respondido a la encuesta.

Es necesario analizar por separado las respuestas en días laborables de las obtenidas en días festivos, ya que la utilización de tiempo es diferente en días de distinto tipo. Por ello, hemos separado los datos de la EPT-93 en dos ficheros: `trabajo.dat` (con las observaciones para los 2521 individuos que responden un día de lunes a viernes) y `fiesta.dat` (con los datos para los 2519 que responden un día sábado o domingo).

El objetivo de este trabajo es presentar la metodología empleada para enlazar la información de `trabajo.dat` y `fiesta.dat`, e ilustrar su comportamiento. De esta forma, tratamos de aproximarnos a los resultados que se habrían obtenido con un cuestionario semanal. Esto puede verse como un caso particular del problema general descrito más arriba, considerando `trabajo.dat` y `fiesta.dat` como ficheros A y B.

El resto de este artículo se organiza de la siguiente manera: la Sección 2 contiene una breve descripción de algunos de los métodos empleados para enlazar encuestas;

en la Sección 3 proponemos un método de enlace de encuestas que emplea árboles de regresión y/o clasificación; examinaremos sus ventajas e inconvenientes frente a los métodos señalados en la Sección 2 en el caso más simple y los problemas que plantea su extensión multivariante; en la Sección 4 mostraremos un procedimiento para solventar dichos problemas, y en la Sección 5 aplicaremos el algoritmo resultante a datos reales (de la EPT-93).

2. TÉCNICAS DE ENLACE DE ENCUESTAS

Una idea que surge de modo natural es utilizar los valores disponibles de (X, Y) y (X, Z) que aparecen en la Figura 1 para regresar las variables Y y Z sobre las X. Podríamos luego imputar los valores desconocidos por los valores ajustados $\hat{Y}$ o $\hat{Z}$, mediante las regresiones obtenidas. Esta idea se remonta al menos a 1960 (véase Buck (1960)), y posteriormente ha habido numerosas contribuciones en esta línea. También, si puede especificarse la función de verosimilitud, puede emplearse el algoritmo EM en la estimación de los valores perdidos en la Figura 1 (véase Dempster et al. (1976)).

En lugar de emplear un modelo paramétrico para el ajuste pueden emplearse métodos de reemplazamiento, sustituyendo los valores de las variables desconocidas para un individuo mediante los valores que esas variables toman en otro individuo «próximo» a él, según una métrica predefinida en el espacio de las variables comunes X. Nos encontramos entonces con diferentes versiones de la idea de vecinos más próximos, entendiendo la proximidad en el espacio $\mathcal{X}$ de las variables X y basándonos en alguna noción útil de distancia. A este método se le denomina «hot-deck imputation» (ver, por ejemplo, Little y Rubin (1987), p. 60).

Se han propuesto también métodos para desvelar la relación entre Y y Z en casos como el que nos ocupa. En Aluja y Rius (1994) y Aluja et al. (1995), se estudia cómo proyectar la información de una encuesta en los planos factoriales obtenidos del análisis de la otra (la encuesta de referencia), mediante técnicas de análisis factorial (análisis de correspondencias múltiples, análisis de componentes principales, etc.) (ver también Bárcena et al. (1997) sobre el mismo tema). Métodos similares y estrechamente relacionados son el método de componentes principales de Dear y el método de descomposición en valores singulares de Krzanowski: en Bello (1993) se da una breve descripción de cada uno y un estudio de simulación para comparar su aplicación.

Las redes neuronales tienen la ventaja de su extrema flexibilidad para modelizar las dependencias más diversas y no requerir la especificación de un modelo *a priori*. Ejemplos recientes del uso de redes neuronales en problemas de imputación son Villagarcía y Muñoz (1997) y Nordbotten (1996).

La técnica conocida como imputación múltiple es un complemento interesante a lo presentado hasta ahora. En Little y Rubin (1987) se presenta convincentemente su fundamento y sus beneficios (ver también Rubin (1986)).

La imputación múltiple permite tener una idea de la variabilidad de las estimaciones entre imputaciones. Tenemos así no sólo un resultado, sino también una idea de la imprecisión introducida en el mismo por el proceso de imputación.

3. ENLACE DE ENCUESTAS MEDIANTE ÁRBOLES DE REGRESIÓN Y/O CLASIFICACIÓN

Proponemos el uso de árboles de regresión y/o clasificación como método de imputación. Desde nuestro punto de vista, utilizar árboles soluciona varios problemas: proporciona un tratamiento unificado tanto para variables cualitativas como cuantitativas; de su empleo se derivan resultados que pueden utilizarse para valorar el enlace, y permite realizar fácilmente imputación múltiple. Además los árboles tienen en ésta como en otras aplicaciones ventajas bien conocidas: flexibilidad, escasez de supuestos, resistencia a *outliers*, etc. El trabajo seminal Breiman et al. (1984) describe bien estas ventajas.

3.1. Grupos de variables específicas, Y y Z, univariantes

Estudiamos aquí el caso más sencillo en que tenemos p variables comunes X y $q = r = 1$, es decir, una única variable específica a imputar en cada encuesta (nos referimos a la Figura 1). Relegamos el caso $q > 1, r > 1$ a la Sección 3.2, que muestra los problemas que plantea la imputación multivariante. Un método para llevarla a cabo se presenta en la Sección 4.

Denotaremos $\mathcal{X}$ el espacio formado por todos los posibles valores de X. La idea es construir dos particiones de $\mathcal{X}$, tales que en cada clase de la primera (segunda) partición tengamos valores lo más parecidos posible de Y (respectivamente, Z). Dado que no imponemos restricciones sobre el tipo de variables, la metodología CART de Breiman et al. (1984) es adecuada. Remitimos al lector a dicha fuente para una presentación de dicha metodología, y especialmente de las técnicas para construir y podar árboles.

Nuestra primera aproximación viene recogida en el Algoritmo 1. Nótese que el método descrito en el mismo se presta bien a la imputación múltiple, ya que cada individuo cae en un nodo terminal que normalmente contiene casos con diferentes valores de Z (o de Y). Ello permite tomar uno o varios al azar. Podemos también imputar mediante la media, mediana, etc. de los valores que encontremos, o siguiendo la regla de la mayoría en el caso de que la variable a imputar sea cualitativa.

El método propuesto, en esencia, puede verse como un método de regresión en el que se emplea un árbol en lugar de un modelo paramétrico, o como un método «hot deck» en que la selección del individuo donante (o una pluralidad de ellos, si se recurre a imputación múltiple) se hace con ayuda de un árbol.

Algoritmo 1. Imputación univariante por árbol.

---

1. Construir un árbol $\mathcal{Y}_X$ «regresando» Y sobre las X, usando validación cruzada y las observaciones $i = 1, \ldots, N_A$. Denotamos las hojas por $\mathcal{Y}_1, \ldots, \mathcal{Y}_a$. Forman la partición $\mathcal{Y}$.
2. Construir un árbol $\mathcal{Z}_X$ «regresando» Z sobre las X, usando validación cruzada y las observaciones $i = N_A + 1, \ldots, N$. Denotamos las hojas por $\mathcal{Z}_1, \ldots, \mathcal{Z}_b$. Forman la partición $\mathcal{Z}$.
3. Para imputar un valor de Z correspondiente a un caso $i = 1, \ldots, N_A$, lo dejamos caer por el árbol $\mathcal{Z}_X$. Si cae en la hoja $\mathcal{Z}_{\delta(i)}$, imputamos en función de los casos de la muestra de entrenamiento que han caído en dicha hoja.
   Procedemos del mismo modo para imputar valores de Y para los casos en que falta dicha variable ($i = N_A + 1, \ldots, N$).

A la vista del Algoritmo 1 cabe preguntarse: ¿Podemos mejorarlo? ¿Podemos aprovechar para predecir, por ejemplo, Z, que conocemos no sólo los valores de X sino también los de Y? La respuesta es: no. Con la muestra disponible (ver Figura 1), no hay información en la Y que pueda ayudarnos a predecir Z salvo la ya contenida en las X (ver Bárcena y Tusell (1998b) a este respecto). Sólo si tuviésemos información adicional sobre la relación entre Y y Z dada X podríamos aprovechar que conocemos los valores de Y además de los de X; esta información adicional podría obtenerse mediante una muestra de observaciones completas de los tres grupos de variables (X, Y, Z) o mediante información extra-muestral (por ejemplo, conocimiento experto). El uso de árboles puede orientar sobre cómo tomar una muestra adicional de observaciones completas de (X, Y, Z), si ello fuera posible: Clases o nodos terminales con una variabilidad intra-clase alta indican estratos de población donde conviene tomar una muestra completa (X, Y, Z).

Como conclusión, creemos que el uso de árboles a la hora de enlazar encuestas tiene las siguientes ventajas:

1. Su uso es sencillo y los resultados fáciles de interpretar.
2. No es necesario que las variables sigan una determinada distribución, ni que la relación entre la variable a explicar y los «regresores» tenga una forma funcional determinada.

3. Cada caso a imputar cae en un nodo terminal en el que normalmente hay más de un caso de la encuesta donante. Esto permite realizar de forma sencilla imputación múltiple.
4. Las hojas o clases con elevada dispersión sugieren estratos de población donde sería más conveniente tomar una muestra adicional de observaciones completas de (X, Y, Z), si existiera la posibilidad de hacerlo.
5. El empleo de variables suplentes *(surrogate splitting)* permite emplear el árbol incluso con valores perdidos entre las X.

3.2. Generalización: Y y Z multivariantes

Al tratar de generalizar el método a situaciones en que Y y Z son multivariantes ($q > 1$ y $r > 1$) nos encontramos con un problema: la metodología existente sobre árboles de regresión y/o clasificación sólo contempla variables respuesta univariantes. La idea más inmediata sería construir árboles para variables respuesta multivariantes que proporcionasen una partición de $\mathcal{X}$ en cada una de cuyas regiones estuviesen los individuos con valores «semejantes» del vector respuesta. No hay una manera única de definir semejanza en este contexto. Como alternativa, pensamos en varias opciones que se describen a continuación. Por brevedad, nos centramos en el caso de imputar valores de las variables Y; para imputar Z se procede de modo análogo.

Opción 1: Imputar cada variable $Y_1, \ldots, Y_q$ por separado, construyendo un árbol para cada una de ellas sobre las variables comunes X. Se emplean para ello los datos completos en (X, Y) en la encuesta A.

Esto supone convertir un problema multivariante en varios univariantes, prescindiendo de las relaciones entre las variables Y. Por tanto, esta opción parece apropiada si las variables a imputar están próximas a la independencia. Si éste no es el caso y entre las variables a imputar existe una estructura de correlación que interesa preservar, nos planteamos las alternativas siguientes.

Opción 2: Efectuar un cambio de variables, reemplazando $Y_1, \ldots, Y_q$ por sus componentes principales $U_1, \ldots, U_q$. Se imputan entonces las componentes principales y, hecho esto, se deshace el cambio, convirtiendo los valores imputados de las componentes a valores imputados de las variables originales.

Esta segunda opción tiene la ventaja (respecto a la Opción 1) de que parte de la relación entre las variables a imputar (aquélla de que dan cuenta las componentes principales) queda recogida. Tiene el inconveniente de que los árboles construidos son más difíciles de interpretar.

Al imputar cada variable o componente por separado, los valores imputados pueden pertenecer a individuos diferentes y, por tanto, no hay garantía de que esos valores imputados sean coherentes entre sí, en el sentido de verificar una relación lógica o matemática que necesariamente debería darse entre los valores imputados. Por ejemplo, en el caso de la encuesta EPT-93, si imputásemos el tiempo dedicado a cada actividad por separado nada impediría que la suma de tiempos fuese diferente de 24 horas (algo que necesariamente debe ocurrir).

Opción 3. Imputar a cada caso i de una encuesta todo el vector de variables desconocidas de una vez, tomando los valores del vector homólogo en un individuo «semejante» de la otra encuesta. Este modo de operar parece ser comúnmente aceptado (véase Lejeune (1995), pág. 140 y Lebart y Lejeune (1995) a este respecto).

Así aseguramos que los valores imputados son coherentes; los toma un individuo de esa población (estamos suponiendo que las encuestas se han realizado sobre la misma población). Por ello, consideramos esta opción la más adecuada. La Sección 4 describe un método para implementarla.

4. DE LA IMPUTACIÓN SIMPLE A LA CONJUNTA: UN NUEVO MÉTODO

Para imputar simultáneamente $\mathbf{Y_i} = (Y_{i1}, \ldots, Y_{iq})$ utilizamos los resultados de los árboles $\mathcal{Y}_X^{(j)}$ construidos para cada una de las variables Y_j, $j = 1, \ldots, q$, del modo que se explica a continuación y se formaliza en el Algoritmo 2. Como anteriormente, discutimos sólo el caso de imputar las variables Y; se procede de modo análogo para las Z.

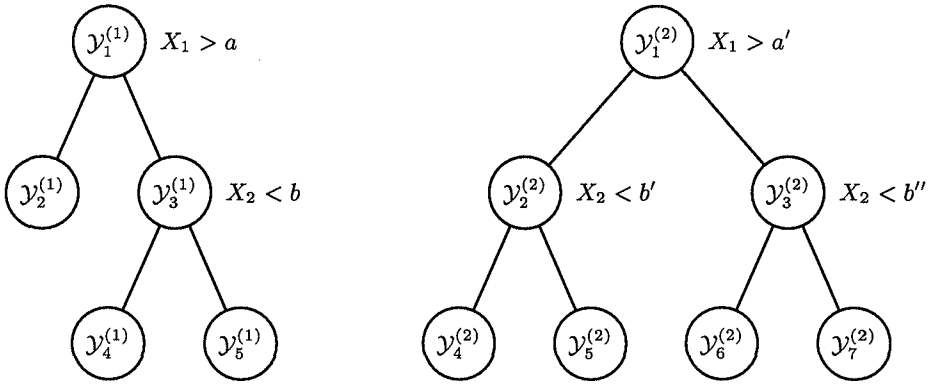


Figura 2. Árboles $\mathcal{Y}_X^{(1)}$ e $\mathcal{Y}_X^{(2)}$. Junto a cada nodo no terminal aparece la condición que, de verificarse, da lugar a clasificar en el hijo derecho.

Supondremos los nodos de cada árbol, terminales o no, numerados. Sea $\mathcal{Y}_k^{(j)}$ el k-ésimo nodo (terminal o no) del árbol $\mathcal{Y}_X^{(j)}$. Designaremos con el mismo símbolo $\mathcal{Y}_k^{(j)}$ el nodo del árbol, la región de $\mathcal{X}$ con valores de X que pueden dar lugar a que un caso transite por, o finalice en, dicho nodo, y el subconjunto de casos que lo hacen en la encuesta donante.

Por ejemplo, supongamos $q = 2$ (hay por tanto dos variables Y_1 e Y_2 a imputar en la encuesta A, para las que hemos construido dos árboles $\mathcal{Y}_X^{(1)}$ e $\mathcal{Y}_X^{(2)}$). Imaginemos que los árboles construidos tienen una forma tan simple como la recogida en la Figura 2.

Las particiones del espacio $\mathcal{X}$ realizadas por ambos árboles pueden entonces verse una junto a otra en la Figura 3. En cada caso se han rotulado las regiones de $\mathcal{X}$ con los nombres de la hojas correspondientes.

Para cualquier q-tupla $(\alpha_1, \ldots, \alpha_q)$ tal que α_j ($j \in \{1, \ldots, q\}$) sea etiqueta de un nodo en el árbol j, definimos

$$\mathcal{C}_{\alpha_1,\ldots,\alpha_q} = \mathcal{Y}_{\alpha_1}^{(1)} \cap \mathcal{Y}_{\alpha_2}^{(2)} \cap \ldots \cap \mathcal{Y}_{\alpha_q}^{(q)}. \tag{1}$$

Consecuentes con la definición anterior, $\mathcal{C}_{\alpha_1,\ldots,\alpha_q}$ designará a un tiempo un subconjunto de casos de la muestra de entrenamiento —el de aquéllos que al ser procesados por los árboles $\mathcal{Y}^{(1)}, \ldots, \mathcal{Y}^{(q)}$ respectivamente transitan por, o finalizan en, los nodos $\mathcal{Y}_{\alpha_1}, \ldots, \mathcal{Y}_{\alpha_q}$ —y una región de $\mathcal{X}$— la formada por los valores de las variables comunes X que dan lugar a que un caso esté en la clase $\mathcal{C}_{\alpha_1,\ldots,\alpha_q}$. Designaremos, finalmente, por $\mathcal{Y}_{(\uparrow\alpha_k)}^{(j)}$ al nodo «padre» del $\mathcal{Y}_{\alpha_k}^{(j)}$ en el árbol $\mathcal{Y}_X^{(j)}$. Cuando no exista duda del árbol al que nos referimos, designaremos el nodo simplemente por α_k y a su padre por $(\uparrow \alpha_k)$.

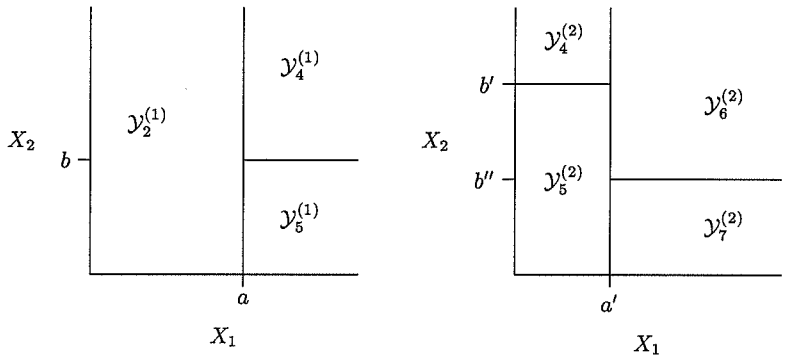


Figura 3. Particiones del espacio $\mathcal{X}$ realizadas respectivamente por los árboles $\mathcal{Y}_X^{(1)}$ e $\mathcal{Y}_X^{(2)}$.

Sea ahora el caso i con $i \in \{N_A + 1, \ldots, N\}$, para el que deseamos imputar $\mathbf{Y_i}$. Imaginemos que al clasificar dicho caso con ayuda de los árboles construidos para las

variables Y, finaliza en las hojas $\mathcal{Y}_{i_1}^{(1)}, \ldots, \mathcal{Y}_{i_q}^{(q)}$. La idea es entonces imputar $\mathbf{Y_i}$ como función de los vectores $\mathbf{Y}$ correspondientes a casos en la encuesta donante pertenecientes a $\mathcal{C}_{i_1,\ldots,i_q}$, es decir, que al ser procesados por cada uno de los árboles finalizan precisamente en las mismas hojas que el caso a imputar i. De nuevo, al igual que en el Algoritmo 1, las opciones son ahora muchas: imputar mediante un vector tomado al azar de $\mathcal{C}_{i_1,\ldots,i_q}$, mediante la media de todos o algunos, etc.

En el ejemplo anterior, supongamos un caso con $a' < X_1 < a$ y $X_2 < b''$; terminará en las hojas $\mathcal{Y}_2^{(1)}$ e $\mathcal{Y}_7^{(2)}$ al dejarlo caer respectivamente por los árboles $\mathcal{Y}_X^{(1)}$ e $\mathcal{Y}_X^{(2)}$. La intersección de dichas dos hojas,

$$\mathcal{C}_{2,7} = \mathcal{Y}_2^{(1)} \cap \mathcal{Y}_7^{(2)}, \tag{2}$$

es el conjunto de casos en la muestra de entrenamiento que caen en los mismos nodos que el caso analizado o, alternativamente, que tienen valores de (X_1, X_2) en la región señalada como $\mathcal{C}_{2,7}$ de la Figura 4.

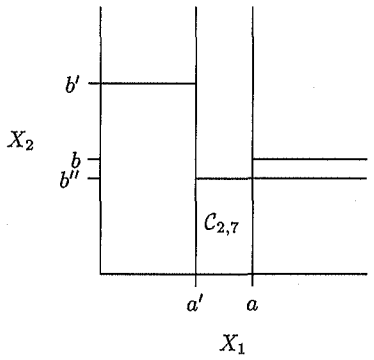


Figura 4. Particiones superpuestas del espacio $\mathcal{X}$ realizadas respectivamente por los árboles $\mathcal{Y}_X^{(1)}$ e $\mathcal{Y}_X^{(2)}$, mostrando $\mathcal{C}_{2,7}$.

Si entre las variables Y_1 e Y_2 existe relación, tiene sentido imputarlas conjuntamente en función de los valores que esas variables toman en los casos de la muestra de entrenamiento que caen en la intersección. Buscar intersecciones supone afinar las particiones de $\mathcal{X}$ que determinan cada uno de los árboles $\mathcal{Y}_X^{(1)}$ e $\mathcal{Y}_X^{(2)}$ por separado, considerando «semejantes» al caso a imputar aquéllos que han caído en las mismas hojas tanto de $\mathcal{Y}_X^{(1)}$ como de $\mathcal{Y}_X^{(2)}$.

Un problema que puede presentarse en el método propuesto es que no exista ningún individuo de la muestra de entrenamiento que caiga en los mismos nodos que el caso a

imputar i; es decir, que la intersección $\mathcal{C}_{i_1,\ldots,i_q}$ de las hojas en las que cae dicho caso i sea vacía. Una solución a este problema es, partiendo de las hojas $\mathcal{Y}_{i_1}^{(1)},\ldots,\mathcal{Y}_{i_q}^{(q)}$ donde ha caído el individuo i, ir «trepando», esto es, sustituyendo sucesivamente los nodos por sus «padres». Treparíamos por los árboles $\mathcal{Y}_X^{(1)},\ldots,\mathcal{Y}_X^{(q)}$ hasta encontrar una intersección no vacía. La idea es eliminar paulatinamente divisiones en algunos árboles para conseguir una intersección no vacía, procurando, al hacerlo, que se pierda lo mínimo posible en calidad de imputación.

Veamos cómo hacerlo. En lo que sigue, como hasta ahora, nos referimos exclusivamente a la imputación de las variables Y; con las Z se procedería de modo análogo. Por concreción supondremos también que las Y son continuas, y estamos ante árboles de regresión, pero la idea es generalizable a variables Y cualitativas y árboles de clasificación.

Como sabemos (ver, por ejemplo, Breiman et al. (1984), Cap. 3), en el proceso de construcción de un árbol de regresión se subdividen los nodos en tanto ello reporte alguna mejora en términos de desviación *(deviance)* —por lo general, en árboles de regresión, suma de cuadrados residual—. En el contexto de árboles de regresión, llamamos $R(t)$ a la desviación en el nodo t, y $R(T)$ a la desviación total del árbol T, definidas respectivamente así:

$$R(t) = \sum_{i\in t}(y_i - \overline{y}_t)^2 \tag{3}$$

$$R(T) = \sum_{t\in\tilde{T}} R(t), \tag{4}$$

en que $\tilde{T}$ denota el conjunto de hojas o nodos terminales del árbol T e $\overline{y}_t$ es la media aritmética de los valores que la variable respuesta toma en los casos caídos en el nodo t.

El coste de trepar en un árbol cualquiera $\mathcal{Y}_X^{(j)}$, $j = 1,\ldots,q$, del nodo hijo t_h al nodo padre t_p puede evaluarse de la siguiente forma:

$$c^{(j)}(t_h) = \frac{\sum_{i=1}^{N_p}(y_{ij} - \overline{y}_{j,t_p})^2}{N_p} - \frac{\sum_{i=1}^{N_h}(y_{ij} - \overline{y}_{j,t_h})^2}{N_h} \tag{5}$$

$$= \hat{R}(t_p)/N_p - \hat{R}(t_h)/N_h \qquad \forall j = 1,\ldots,q, \tag{6}$$

en que $\hat{R}(t_h)$ y $\hat{R}(t_p)$ son estimaciones por resustitución de la desviación en el nodo «hijo» t_h (es decir, el nodo desde el que queremos trepar) y en su nodo «padre» t_p respectivamente, N_p y N_h son el número de casos en los nodos t_p y t_h, e $\overline{y}_{j,t_p}, \overline{y}_{j,t_h}$ las medias respectivas de la variable Y_j en dichos nodos.

Con la notación anterior, estamos ya en situación de especificar un algoritmo de imputación; es el recogido como Algoritmo 2. Nótese que el modo en que se hace la

imputación a partir de una intersección $\mathcal{C}_{\alpha_1,\ldots,\alpha_q}$ queda sin especificar: cualquiera de las opciones mencionadas anteriormente (media, mediana, imputación múltiple, etc.) es utilizable. En la medida en que el procedimiento selecciona grupos de casos de la encuesta donante similares (en lo que se refiere a los valores de las variables X) al que tratamos de imputar, la imputación múltiple es fácil de hacer. Nótese también que es posible hacer la imputación sobre las variables originales o transformaciones de las mismas, como componentes principales. Esto último se ha hecho en la aplicación que mostramos en la Sección 5.

Algoritmo 2. Imputación multivariante mediante árboles.

---

1. *(opcionalmente)* Calcular las componentes principales del grupo de variables Y a imputar utilizando la muestra de entrenamiento (la encuesta donante).
2. Construir los árboles $\mathcal{Y}_X^{(1)},\ldots,\mathcal{Y}_X^{(q)}$.
3. **for** $i \in \{$Casos a imputar$\}$ **do**
4. Dejarlo caer por cada árbol, y a partir de las hojas $\mathcal{Y}_{\alpha_1}^{(1)},\ldots,\mathcal{Y}_{\alpha_q}^{(q)}$ en que cae determinar $\mathcal{C}_{\alpha_1,\ldots,\alpha_q}$.
5. **if** $\mathcal{C}_{\alpha_1,\ldots,\alpha_q} \neq \emptyset$ **then**
6. **break**
7. **else**
8. **while** $\mathcal{C}_{\alpha_1,\ldots,\alpha_q} = \emptyset$ **do**
9. Computar los costes $c^{(1)}(\alpha_1),\ldots,c^{(q)}(\alpha_q)$ de trepar desde los nodos actuales.
10. Seleccionar k tal que la trepa desde el nodo α_k sea de mínimo coste.
11. $\alpha_k \leftarrow \quad (\uparrow \alpha_k)$; sustituir el nodo α_k por su padre.
12. **end while**
13. **end if**
14. Imputar el caso i a partir de $\mathcal{C}_{\alpha_1,\ldots,\alpha_q}$.
15. **end for**
16. *(si procede)* Reconstruir las variables originales a partir de los valores imputados de las componentes principales.

---

5. APLICACIÓN A LA ENCUESTA DE PRESUPUESTOS DE TIEMPO (EPT-93)

A continuación ilustraremos el funcionamiento del método enlazando los ficheros `trabajo.dat` y `fiesta.dat`, mencionados en la Introducción. En cada uno de los ficheros hay dos tipos de variables:

- Variables comunes o de caracterización X cuya descripción y modalidades aparecen en la Tabla 1.

Tabla 1. Variables comunes (o de caracterización) junto a sus modalidades.

Variable	**Descripción**	**Código**	**Modalidad**
X_1	Edad	EDA1	Hasta 34 años
		EDA2	Entre 35 y 59 años
		EDA3	60 años o más
X_2	Sexo	VARO	Varón
		MUJE	Mujer
X_3	Estado civil	SOLT	Soltero
		CASD	Casado
		REST	Resto
X_4	Nivel de instrucción	PRIM	Primarios
		MEDI	Medios
		SUPE	Superiores
X_5	Relación con la actividad	SRMI	Servicio militar
		OCUP	Ocupados
		PARA	Parados
		JUBI	Jubilados
		ESTD	Estudiantes
		LAHO	Labores del hogar
		OTRS	Otros

- Variables específicas: tiempo en minutos al día dedicados a las actividades que aparecen en la Tabla 2.

Tabla 2. Variables específicas de uso del tiempo. Las variables $Y_1, \ldots, Y_{24}$ corresponden a usos del tiempo en días laborables, y las $Z_1, \ldots, Z_{24}$ son las variables homólogas para los días no laborables.

Variables	**Descripción**
Y_1, Z_1	Dormir
Y_2, Z_2	Higiene y cuidado personal
Y_3, Z_3	Comer
Y_4, Z_4	Actividades privadas y actividades no descritas
Y_5, Z_5	Trabajo
Y_6, Z_6	Formación
Y_7, Z_7	Tareas domésticas (cocinar, fregar, limpiar la casa, arreglo y cuidado ropa y trabajos diversos)
Y_8, Z_8	Compras
Y_9, Z_9	Gestiones
Y_{10}, Z_{10}	Actividades de semi ocio (punto, costura, pintura, escultura, reparaciones, bricolaje, jardinería, cuidado de animales, etc.)
Y_{11}, Z_{11}	Cuidado de niños y adultos
Y_{12}, Z_{12}	Reuniones de tipo familiar (comidas, defunciones, bodas, visitas hospitalarias, etc.)
Y_{13}, Z_{13}	Reuniones con amigos, fiestas, ir de potes o copas, etc.
Y_{14}, Z_{14}	Participación religiosa o política
Y_{15}, Z_{15}	Gimnasia y deporte
Y_{16}, Z_{16}	Excursiones y paseos
Y_{17}, Z_{17}	Ocio en el hogar (TV, vídeo, música, radio, etc.)
Y_{18}, Z_{18}	Ocio fuera del hogar (cine, teatro, conciertos, museos y exposiciones, espectáculos deportivos, etc.)
Y_{19}, Z_{19}	Otras actividades de ocio (micro informática, fotografía, cartas, juegos, crucigramas, etc.)
Y_{20}, Z_{20}	Trayecto al trabajo o formación
Y_{21}, Z_{21}	Acompañar a otros
Y_{22}, Z_{22}	Esperas en el trabajo o formación
Y_{23}, Z_{23}	Esperas en cuidados médicos y gestiones administrativas
Y_{24}, Z_{24}	Otras esperas

Ambos ficheros proceden de una misma operación de muestreo, y pueden verse como muestras independientes procedentes de la misma población; pueden verse detalles sobre el mismo problema en Bárcena y Fdez. de Agirre (1998). Se trata de un caso particular del problema de enlace de encuestas descrito en la Sección 1, en que `trabajo.dat` y `fiesta.dat` son respectivamente los ficheros A y B. Si colocamos los datos de ambos ficheros formando una matriz incompleta (como la que aparece en la Figura 1), tenemos en este caso: p=5, $q = r$ =24, $N_A = 2521$ y $N_B = 2519$.

Hemos enlazado los ficheros `trabajo.dat` y `fiesta.dat` siguiendo el método explicado en la sección anterior (Algoritmo 2). Los cálculos necesarios se han realizado mediante varias funciones programadas en el lenguaje específico del paquete estadístico S-PLUS. Una descripción de dicho paquete puede verse en Becker et al. (1988) y Chambers y Hastie (1992).

Primero se realizó un Análisis en Componentes Principales (ACP) tipificado de los valores disponibles de las variables Y y Z; el análisis se realiza tipificado ya que las variables Y y Z tienen muy diferente dispersión.

A continuación se construyó el árbol de regresión sobre las variables comunes X de cada componente principal. Para cada componente se obtiene una sucesión de árboles $T_1 \succ T_2 \succ ..... \succ \{t_1\}$ ordenados de más a menos frondoso: $\{t_1\}$ es el árbol degenerado que consiste sólo en el nodo raíz. Se estima mediante validación cruzada la desviación $R(T)$ para cada árbol de la sucesión y se selecciona como árbol T^* de tamaño óptimo el que tenga menor $\hat{R}^{vc}(T)$ (desviación estimada). La validación cruzada se realizó dividiendo aleatoriamente la muestra en 10 bloques, y utilizando por turno las observaciones de cada bloque para estimar la tasa de error de los árboles construidos con los nueve bloques restantes.

Una vez construidos los árboles para cada una de las componentes principales, se ejecuta el bucle principal del Algoritmo 2. En esta particular aplicación optamos por conservar todas las componentes principales dando lugar a árboles con más de un nodo (todas menos dos), dada la casi total falta de correlación entre las variables originales, que no aconsejaba reducir su dimensionalidad. Optamos también por imputar al azar dentro de cada clase de equivalencia: cuando se encuentra una intersección $C_{\alpha_1,...\alpha_q}$ no vacía se escoge aleatoriamente un sólo caso de la misma y se asignan los valores de sus componentes principales al caso a imputar.

El número de individuos que han requerido trepar para encontrar una intersección es de 33 para las variables Y (tan sólo un 1.31% de los casos) y de 48 para las Z (1.91%). Hemos comprobado que para los individuos para los que es necesario trepar, se trepa principalmente por los árboles de las últimas componentes principales (como era de esperar, ya que son los árboles de las componentes con menos dispersión los que cabe imaginar menos costosos de trepar). En varios casos es necesario trepar hasta el nodo raíz de algunos árboles.

El resultado final es una matriz de datos como la representada en la Figura 1 pero completada.

Con el fin de estudiar la calidad del enlace comparamos, desde un punto de vista descriptivo, las distribuciones de las variables Y y Z para valores imputados con las obtenidas para valores disponibles. Un buen enlace de encuestas no debe alterar la distribución de Y ni de Z, y es práctica habitual (véase Lebart y Lejeune (1995)) comparar las distribuciones marginales de las variables en la encuesta donante con las de las mismas variables imputadas en la encuesta receptora.

Tabla 3. Imputación de variables de uso del tiempo en días laborables. Estadísticos de posición y dispersión para datos observados e imputados de variables Y. Se han señalado en negrita las variables cuya diferencia de medias entre datos observados y completados excede de dos desviaciones típicas.

	Valores observados						**Valores imputados**					
	Min.	q_1	Me	$\overline{x}$	q_3	Max.	Min.	q_1	Me	$\overline{x}$	q_3	Max.
Y_1	179	450	499	511.30	559	1214	179	450	495	508.50	555	975
Y_2	0	20	35	42.23	55	295	0	20	35	41.90	55	295
Y_3	0	75	105	109.90	135	374	0	75	100	109.90	135	374
Y_4	0	0	0	0.33	0	105	0	0	0	0.25	0	75
Y_5	0	0	0	196.50	455	990	0	0	0	194.70	460	929
$\mathbf{Y_6}$	0	0	0	23.84	0	760	0	0	0	56.25	0	760
$\mathbf{Y_7}$	0	0	75	133.10	235	770	0	0	45	107.60	185	630
$\mathbf{Y_8}$	0	0	0	30.22	55	355	0	0	0	25.34	45	330
Y_9	0	0	0	2.73	0	415	0	0	0	3.05	0	240
$\mathbf{Y_{10}}$	0	0	0	16.90	0	634	0	0	0	13.81	0	634
$\mathbf{Y_{11}}$	0	0	0	20.53	0	735	0	0	0	15.71	0	625
$\mathbf{Y_{12}}$	0	0	0	7.32	0	630	0	0	0	8.90	0	630
$\mathbf{Y_{13}}$	0	0	0	40.43	60	510	0	0	5	47.35	75	510
Y_{14}	0	0	0	4.64	0	365	0	0	0	5.12	0	350
$\mathbf{Y_{15}}$	0	0	0	6.91	0	315	0	0	0	8.39	0	230
Y_{16}	0	0	0	54.83	90	600	0	0	0	53.92	90	600
$\mathbf{Y_{17}}$	0	74	140	163.00	225	964	0	65	135	157.70	224	730
$\mathbf{Y_{18}}$	0	0	0	1.48	0	239	0	0	0	2.09	0	239
Y_{19}	0	0	0	10.76	0	405	0	0	0	9.59	0	350
Y_{20}	0	0	0	20.48	30	340	0	0	0	21.14	30	340
$\mathbf{Y_{21}}$	0	0	10	33.32	50	355	0	0	15	39.25	60	355
Y_{22}	0	0	0	1.49	0	165	0	0	0	1.81	0	165
Y_{23}	0	0	0	0.50	0	165	0	0	0	0.56	0	165
Y_{24}	0	0	0	0.79	0	75	0	0	0	0.94	0	75

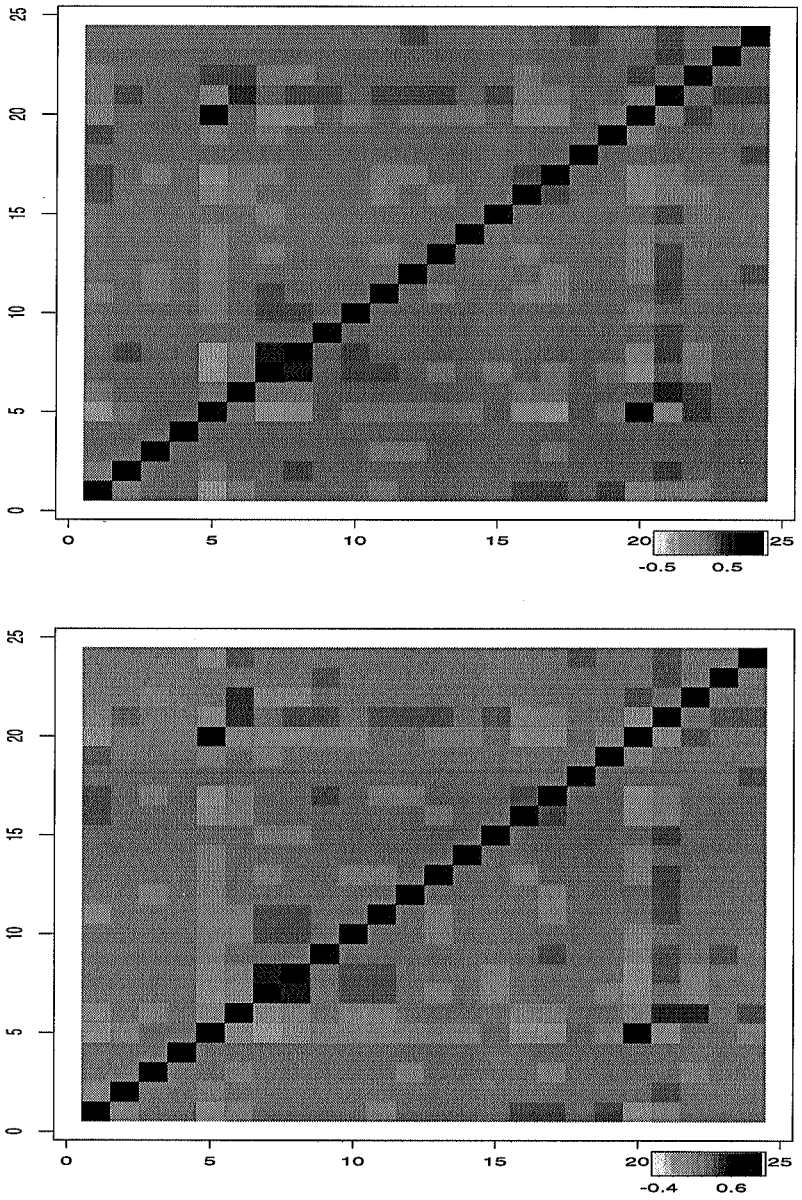


Figura 5. Representación gráfica de las matrices de correlación de las Y observadas (arriba) e imputadas (abajo). Nótese la disposición de las variables en los ejes y la gran similitud entre ambas.

Así lo hemos hecho nosotros comparando media, mediana, cuartiles y valores extremos de los valores disponibles de las variables específicas (Y y Z) y los obtenidos por imputación. Los resultados para Y se muestran en la Tabla 3 (fueron análogos para Z, y no se reproducen en interés de la brevedad.) Las principales diferencias se dan en la variable Y_7 («Tareas domésticas») e Y_{13} («Reuniones con amigos»).

También hemos comparado las matrices de correlación para valores disponibles e imputados. De nuevo se representan solamente (en la Figura 5) las correspondientes a las variables Y; análogos resultados se obtuvieron para las Z. Se observa su gran similitud aparente: la máxima diferencia en valor absoluto es de 0.1124 (que corresponde a la correlación entre Y_5 e Y_6).

Como conclusión de esta aplicación observamos que el método de enlace descrito es factible, reproduce aceptablemente las distribuciones marginales de las variables Y y Z y sus respectivas estructuras de correlación. La información recuperada sobre la relación entre las Y y las Z (a la que hemos omitido en lo que antecede) es escasa, confirmando la necesidad de un conjunto de variables comunes X suficientemente rico y descriptivo (cf. Lejeune (1995)). Detalles adicionales sobre la aplicación pueden obtenerse de Bárcena y Tusell (1998a).

REFERENCIAS

Aluja, T., R. Nonell, R. Rius y M. Martínez (1995). «File grafting». En Mola y Morineau (1997), 23–32.

Aluja, T. y R. Rius (1994). «Inserción de datos de encuesta mediante análisis de componentes principales». *Presented at the XXI Congreso nacional de Estadística e Investigación Operativa (SEIO). Calella.*

Bárcena, M. y K. F. de Agirre (1998). «Visualization Techniques Related to Survey Linking: Two Alternative Ways and Bootstrap Validation». *BILTOKI DT98.14*, Universidad del País Vasco.

Bárcena, M., K. Fdez. Aguirre y F. Tusell (1997). «Survey grafting with common structural base: Time use Survey and Health Survey». En Mola y Morineau (1997).

Bárcena, M. y F. Tusell (1998a). «Enlace de encuestas: una propuesta metodológica y aplicación a la Encuesta de Presupuestos de Tiempo». *Inf. Téc. 98.07*, Departamento de Econometría y Estadística.

Bárcena, M. y F. Tusell (1998b). «Linking surveys using reciprocal classification trees». En Fernández y Morineau (1998), 67–76. CISIA-CERESTA.

Becker, R., J. Chambers y A. Wilks (1988). *The New S Language. A Programming Environment for Data Analysis and Graphics.* Wadsworth & Brooks/Cole, Pacific Grove, California.

Bello, A. (1993). «Choosing among imputation techniques for incomplete multivariate data: a simulation study». *Communications in Statistics - Theory and Methods*, 22(3), 853–877.

Breiman, L., J. Friedman, R. Olshen y C. Stone (1984). *Classification and Regression Trees.* Wadsworth, Belmont, California.

Buck, S. (1960). «A method of estimation of missing values in multivariate data suitable for use with an electronic computer». *Journal of the Royal Statistical Society, Ser. B*, 22, 302–306.

Chambers, J. y T. Hastie (1992). *Statistical Models in S.* Wadsworth & Brooks/Cole, Pacific Grove, Ca.

Dempster, A., N. Laird y D. Rubin (1976). «Maximum likelihood from incomplete data via the EM algorithm». *Journal of the Royal Statistical Society, Ser. B*, 39, 1–38.

EUSTAT (1997). *Análisis de Tipologías de Jornadas Laborales.* Instituto Vasco de Estadística, (EUSTAT), Vitoria/Gazteiz.

Fernández, K. y Morineau, A. (eds.) (1998). *Analyse Multidimensionnelle des Données - NGUS'97.* Centre Int. de Statistique et d'Informatique Appliquées, CISIA-CERESTA.

Lebart, L. y M. Lejeune (1995). «Assessment of Data Fusions and Injections». En *Encuentro Internacional AIMC sobre Investigación de Medios*, 1–18. Madrid.

Lejeune, M. (1995). «De l'usage des fusions de données dans les études de marché». En *Proceedings of the IASS Meeting, Beijing.*

Little, R. y D. Rubin (1987). *Statistical Analysis with Missing Data.* John Wiley and Sons, New York.

Mola, F. y A. Morineau (eds.) (1997). *Actes du IIIème Congrès International d'Analyses Multidimensionnelles des Données - NGUS'95.* Centre Int. de Statistique et d'Informatique Appliquées, CISIA-CERESTA.

Nordbotten, S. (1996). «Neural Network Imputation Applied to the Norwegian 1990 Population Census Data». *Journal of Official Statistics*, 12(4), 385–401.

Rubin, D. (1986). «Statistical Matching Using File Concatenation With Adjusted Weights and Multiple Imputations». *Journal of Business and Economic Statistics*, 4(1), 87–94.

Villagarcía, T. y A. Muñoz (1997). «Imputación de datos censurados mediante redes neuronales: una aplicación a la EPA». *Cuadernos Económicos de I.C.E.*, (63), 193–204.

ENGLISH SUMMARY

LINKING SURVEYS: A METHOD AND APPLICATION TO THE SURVEY OF TIME USES

M. J. BÁRCENA
F. TUSELL
Universidad del País Vasco*

We address the problem of completing two files with records containing a common subset of variables. The technique investigated involves the use of regression and/or classification trees. An extension is proposed to deal with multivariate response variables. The method is demonstrated on a real problem, merging two files of the Survey of Time Uses (EPT-93).

Keywords: File matching, survey linking, imputation, regression trees, classification trees.

AMS Classification: 62G05, 62D05.

We thank for support the Spanish MEC (grant PB95-0346) and UPV/EHU (grant PB95-0346). We gratefully acknowledge comments from Vicente Núñez, Eva Ferreira and Karmele Fernández. Any errors and obscurities that remain are our own.

* Departamento de Estadística y Econometría. Facultad de CC.EE. y Empresariales. Avda. del Lehendakari Aguirre, 83. 48015 Bilbao. E-mail: etptupaf@bs.ehu.es.

– Received June 1998.
– Accepted March 1999.

1. INTRODUCTION

Our starting point are two files, A and B with a total of $N = N_A + N_B$ observations and the following structure:

File A

$X_1, \ldots, X_p$	$Y_1, \ldots, Y_q$	unknown

File B

$X_1, \ldots, X_p$	unknown	$Z_1, \ldots, Z_r$

In the problem that motivates this research, these two files contain data from two sample surveys which share a common set of questions. The population sampled is assumed to be the same for both surveys. Variables $X_1, \ldots, X_p$ are known for all N cases. Beyond this common set of variables, each file contains also data on a specific set of variables: $Y_1, \ldots, Y_q$ and $Z_1, \ldots, Z_r$ in our notation.

In this paper we deal with the problem of imputing the missing values in the files A y B, using the values of the common variables $X_1, \ldots, X_p$. The goal is to create a data set as close as possible to what would have been obtained, had we had the chance to gather complete information on variables (X, Y, Z) for all N cases.

The EPT-93 (Survey of Time Uses), compiled by the EUSTAT (Basque Institute of Statistics), is a survey comprising responses of 5040 individuals which were asked about the time devoted daily to different purposes, grouped by us in 24 categories. The field work was done in the Fall of 1992 and Spring of 1993; refer to EUSTAT (1997) for a description of the survey.

Ideally, each individual should answer the survey both on working and non working days; but when individuals were asked to compile a daily breakdown of their uses of time for a period one week long, it was found that the quality of the diaries degraded. In the end, single day questionnaires were administered, to be completed for a specific day of the week. We therefore have 2521 individuals who answered on working days (in file `trabajo.dat`) and 2519 who answered on weekend days (in file `fiesta.dat`). The information on these two sets of answers compose files A and B. In either file we have answers to a common set of characterization variables $X_1, \ldots, X_p$, and then information regarding the use of time in a working day (variables $Y_1, \ldots, Y_q$, in file A) or a weekend day (variables $Z_1, \ldots, Z_r$, in file B).

2. IMPUTATION USING REGRESSION OR CLASSIFICATION TREES

We propose the use of regression and/or classification trees to impute missing values. To simultaneously impute $\mathbf{Y_i} = (Y_{i1}, \ldots, Y_{iq})$ we use the univariate trees $\mathcal{Y}_X^{(j)}$ cons-

tructed for each of the variables Y_j, $j = 1, \ldots, q$, as described next and formalized in Algorithm 1. In the interest of brevity we address the problem of imputing the variables in Y; to impute Z we proceed likewise.

Let the nodes of each tree be numbered, and let $\mathcal{Y}_k^{(j)}$ be the k-th node of tree $\mathcal{Y}_X^{(j)}$. We use $\mathcal{Y}_k^{(j)}$ to denote the node, the subset of cases ending in, or going through, that node, and the corresponding region of $\mathcal{X}$.

For each q-tuple $(\alpha_1, \ldots, \alpha_q)$ such that α_j ($j \in \{1, \ldots, q\}$) is the label of a node in tree $\mathcal{Y}_X^{(j)}$, we define

$$\mathcal{C}_{\alpha_1,\ldots,\alpha_q} = \mathcal{Y}_{\alpha_1}^{(1)} \cap \mathcal{Y}_{\alpha_2}^{(2)} \cap \ldots \cap \mathcal{Y}_{\alpha_q}^{(q)}. \tag{1}$$

Finally, let node $\mathcal{Y}_{(\uparrow\alpha_k)}^{(j)}$ be the "father" of node $\mathcal{Y}_{\alpha_k}^{(j)}$ in tree $\mathcal{Y}_X^{(j)}$. When there is no ambiguity about the tree referred to, we will simply refer to nodes $(\uparrow \alpha_k)$ and α_k.

Algorithm 1. Multivariate imputation using trees.

---

1. *(optionally)* Compute the principal components of the variables Y to impute using the training sample.
2. Construct trees $\mathcal{Y}_X^{(1)}, \ldots, \mathcal{Y}_X^{(q)}$.
3. **for** $i \in$ {Cases to impute} **do**
4. Drop case i down the trees, and determine the intersection $\mathcal{C}_{\alpha_1,\ldots,\alpha_q}$ of the leaves $\mathcal{Y}_{\alpha_1}^{(1)}, \ldots, \mathcal{Y}_{\alpha_q}^{(q)}$ where it falls.
5. **if** $\mathcal{C}_{\alpha_1,\ldots,\alpha_q} \neq \emptyset$ **then**
6. **break**
7. **else**
8. **while** $\mathcal{C}_{\alpha_1,\ldots,\alpha_q} = \emptyset$ **do**
9. Compute the costs $c^{(1)}(\alpha_1), \ldots, c^{(q)}(\alpha_q)$ of climbing from the current nodes.
10. Select k such that climbing from node α_k is of minimal cost.
11. $\alpha_k \leftarrow (\uparrow \alpha_k)$; replace node α_k by its father.
12. **end while**
13. **end if**
14. Impute i from $\mathcal{C}_{\alpha_1,\ldots,\alpha_q}$.
15. **end for**
16. *(if required)* Reconstruct the original variables from the imputed principal components.

Consider now case i, ($i \in \{N_A + 1, \ldots, N\}$), for which an imputation of $\mathbf{Y_i}$ is sought. Assume that when dropping that case through the trees built for each of the variables in Y, it ends in the leaves $\mathcal{Y}_{i_1}^{(1)}, \ldots, \mathcal{Y}_{i_q}^{(q)}$ and hence belongs to $\mathcal{C}_{i_1,\ldots,i_q}$. The simple idea in our method is to impute $\mathbf{Y_i}$ as a function of the values $\mathbf{Y}$ from cases in the training sample (file A) which also belong to $\mathcal{C}_{i_1,\ldots,i_q}$. Those cases have values for each variable $Y_1, \ldots, Y_q$ which, as far as the relevant trees can ascertain, are indistinguishable from the ones of the case to impute.

With the previous notation, we can specify Algorithm 1.

3. IMPUTATION IN THE SURVEY OF USES OF TIME (EPT-93)

The concepts above about linkage of files have been tested on `trabajo.dat` and `fiesta.dat` referred to in Section 1. The two files have been linked using Algorithm 1. The computation has been performed with functions programmed using the S-PLUS language and primitives: see Becker et al. (1988) and Chambers and Hastie (1992) for a description of that package.

In this application we opted for single imputation using one case taken at random from $\mathcal{C}_{\alpha_1,\ldots\alpha_q}$, the first non empty intersection. We kept track of how many times climbing was necessary, because the intersection of leaves reached in the first instance was empty. Only 33 cases (1.31% of the total) required climbing when imputing the variables Y, and 48 cases (1.91% of the total) when imputing the Z's. Climbing was mainly up the trees built for the last principal components —the less costly. In a few cases, climbing all the way up to the root of one tree was required.

The imputed values for Y reproduced reasonably well the distribution of the observed values.

4. SUMMARY AND CONCLUSIONS

A new method for imputation has been presented and its feasibility demonstrated on a real data set. It can cope with a large variety of problems, because of the generality of the tool used for approximation —classification or regression trees. It makes few assumptions, is computationally feasible, and appears to give good results. Simulations seem to confirm this; the method works well whenever the common variables X are good predictors for the Y's and Z's and the functional relationship among predictors and responses can be reasonably well approximated by a tree.

We see room for improvement, specially in the climbing strategy, at the expense of increased complexity and computational burden. Our work proceeds along this line. Further work is also required in comparing our method to other flexible, all-purpose methods of imputation, like those using neural networks.

QÜESTIIÓ, vol. 23, 2, p. 321-342, 1999

QUALITÉ DE L'INFORMATION DANS LES ENQUÊTES

L. LEBART
Ecole Nationale Supérieure des Télécommunications*

Cet article tente de montrer les contributions des analyses statistiques multidimensionnelles et des analyses textuelles à l'amélioration de la qualité des résultats d'une enquête. L'idée de base est la suivante: chaque phase de l'enquête et de sa réalisation sur le terrain peut donner à la mesure ou au relevé de certaines variables de contrôles. Le fichier des données individuelles se voit alors doublé d'un fichier de variables de contrôles, pouvant être numériques (chronometrage de l'entrevue) nominales (circonstances de l'entrevue, appréciations) ou textuelles (commentaires libres de l'enquêteur et de l'enquêté). Les méthodes précitées permettent alors de visualiser et de confronter ces deux fichiers, et donc de juger la cohérence globale de l'information, en positionnant les variables de contrôles parmi les variables de l'enquête.

Quality of data in sample surveys

Mots clé: Qualité de l'information. Enquêtes socio-économiques. Analyse des données. Analyses statistiques textuelles.

AMS Classification: 62H25, 62H30, 62D05, 62P25

* Ludovic Lebart. Centre National de la Recherche Scientifique. Ecole Nationale Supérieure des Télécommunications. 46 Rue Barrault. 75013 Paris.

– Reçu en décembre de 1998.
– Accepté en mai de 1999.

1. INTRODUCTION

Pour les statisticiens, le mot enquête désigne le plus souvent une enquête par sondage (enquêtes démographiques, socio-économiques, socio-politiques, épidémiologiques, enquêtes audimétriques ou de marketing). L'enquête est un recueil d'une certaine ampleur, destiné à mesurer et/ou explorer un phénomène. Il est considéré aujourd'hui comme indispensable d'étudier la cohérence et la validité de l'information produite par cet instrument d'observation complexe et délicat à mettre en oeuvre. Cette étude s'appuie notamment sur les techniques actuelle de traitement des données d'enquêtes. Ces techniques ont été profondément modifiées par l'analyse des données qui intervient, dans une phase préliminaire, pour apprécier la qualité de l'information, synthétiser cette information, et orienter la suite des traitements.

La démarche *Data Mining* reprend cette approche globale des grands fichiers, dans le contexte de la diffusion et de la banalisation de la puissance de calcul (cf. Hand, 1998). Après Fayyad *et al* (1996) on peut définir le Data Mining comme la recherche de *patterns* (de traits structuraux) dans de vastes ensembles de données, ces patterns devant être «valides, nouveaux, potentiellement utiles, et, si possible, compréhensibles ou explicables». Les fichiers peuvent être très grands (des millions d'enregistrements), non structurés et non représentatifs. L'objectif ultime est alors d'extraire des données des informations nouvelles de la façon la plus automatique possible.

L'analyse des données textuelles permet d'étendre ce programme aux informations non numériques (réponses libres, textes). Sous le nom de *Text Mining* elle permet de traiter dans la même optique que le Data Mining des corpus de lettres de réclamations, de questions ouvertes dans les enquêtes de satisfaction ou de marketing, des documents Web et internet.

2. LES CONTRÔLES DE BASE DE L'ENQUÊTE

Chaque phase de la réalisation de l'enquête donne lieu au relevé de *variables de contrôles*, qui vont participer activement au traitement de l'information et à l'évaluation de la qualité de l'information.

2.1. La conception du questionnaire

Le questionnaire est indissociable du contenu de l'enquête et des disciplines concernées. Il serait vain de chercher des règles communes à des questionnaires concernant une enquête nutritionnelle dans un pays en développement, une enquête de satisfaction vis-à-vis de produits financiers, une enquête d'audience de Télévision. Au plus peut-on faire

des recommandations générales concernant: la clarté d'expression et de présentation, la lisibilité, la facilité de manipulation et d'encodage. Pour les enquêtes d'opinions, une série de travaux a porté sur les influences des libellés de questions, de l'ordre des questions, de la nature des questions (fermées ou ouvertes) de la longueur des questionnaires sur les résultats (cf. Schuman et Presser, 1981).

Les premiers contrôles font l'objet d'une étude qualitative sur l'accueil et la compréhension du questionnnaire par les personnes interrogées. D'autre contrôles portent sur la formation des enquêteurs.

Les variables de contrôles attachées à chaque entrevue peuvent concerner l'ordre des questions (cas de permutations aléatoires de titres dans les enquêtes d'audience), les questions ouvertes pourquoi, à la suite des questions fermées dont l'interpretation est susceptible d'être variable (cf. section 5).

2.2. Les modes de questionnement

Actuellement, il existe cinq modes principaux de «passation» d'un questionnaire ou de relevé d'informations de base:

- Le mode externe ou observateur.
- Le mode face à face (direct, ou systèmes dits «CAPI» –computer assisted personal interview–).
- Le mode téléphonique (systèmes dits «CATI» –computer assisted telephone interview–).
- Le mode postal (auto-administré).
- Le mode télématique (panel audimétriques, par exemple).

Les modes informatisés (CATI et CAPI) permettent de relever automatiquement un grand nombre de variables de contrôles, comme, par exemple, le chronométrage détaillé de l'entrevue, la durée totale de l'entrevue, l'heure de l'entrevue).

2.3. Le plan de sondage et son exécution

Cette phase donne lieu à des contrôles par inspecteurs, par contre-visites ou contre-appels. Puis, s'il y a lieu, par une analyse de la distribution des poids de redressement.

2.4. Terrain et recueil

La phase 2.4 donne lieu à des contrôles par inspecteurs, par contre-visites ou contre-appels, et par des questions de contrôles portant sur les caractéristiques et appréciations des enquêteurs (niveau de coopération, atmosphère de l'entrevue). Ces questions peuvent comporter les caractéristiques de l'enquêteur (pour tester le niveau éventuel d'*interaction sociale*), le lieu de l'interview, la présence d'autres personnes. Des questions ouvertes sur l'appréciation de l'interveiw peuvent être posées à l'enquêté et à l'enquêteur.

3. LA VISUALISATION DE DONNÉES D'ENQUÊTES

Il est toujours possible de calculer des distances entre les lignes (individus) et entre les colonnes (variables) d'un tableau rectangulaire de valeurs numériques, mais il n'est pas possible de visualiser ces distances de façon immédiate: il est nécessaire de procéder à des transformations et des approximations pour en obtenir une ou plusieurs représentations planes.

3.1. Méthodes factorielles

C'est une des tâches dévolues à l'analyse factorielle au sens large d'opérer une réduction de certaines représentations «multidimensionnelles». On recherche donc des sous-espaces de faibles dimensions (une, deux ou trois par exemple) qui ajustent au mieux le nuage de points-individus et celui des points-variables, de façon à ce que les proximités mesurées dans ces sous-espaces reflètent autant que possible les proximités réelles. On obtient ainsi un espace de représentation, l'espace factoriel.

Mais la géométrie des nuages de points et les calculs de proximités ou de distances qui en découlent diffèrent selon la nature des lignes et des colonnes du tableau analysé.

Les colonnes peuvent être des variables continues ou des variables nominales ou des catégories dans le cas des tables de contingences. Les lignes peuvent être des individus ou des catégories.

La nature des informations, leur codage, les spécificités du domaine d'application vont introduire des variantes au sein des méthodes factorielles.

On rappelle brièvement ici trois techniques fondamentales:

- *L'analyse en composantes principales (ACP)* (Hotelling, 1933) s'applique aux tableaux de type «variables-individus», dont les colonnes sont des variables à valeurs

numériques continues (exemples: enquêtes de consommation, enquêtes de budget-temps) et dont les lignes sont des individus, des observations, des objets, etc. Les proximités entre variables s'interprètent en termes de corrélation; les proximités entre individus s'interprètent en termes de similitudes des valeurs observées. Elle peut donner lieu à de nombreuses variantes en s'appliquant par exemple à un tableau de rangs (diagonalisation de la matrice de corrélation des rangs de Spearman), ou encore après l'élimination de l'effet de certaines variables (analyses locales ou partielles).

- *L'analyse des correspondances (AC)* (Benzécri, 1973) s'applique aux tableaux de contingences (croisement de deux variables nominales). L'analyse fournit des représentations des associations entre lignes et colonnes de ces tableaux, fondées sur une distance entre profils (qui sont des vecteurs de fréquences conditionnelles) désignée sous le nom de distance du *Chi-deux*.
- *L'analyse des correspondances multiples (ACM)* est une extension du domaine d'application de l'analyse des correspondances, avec cependant des procédures de calcul et des règles d'interprétation spécifiques. Son champ d'application est considérable. Elle est particulièrement adaptée à la description de grands tableaux de variables nominales dont les fichiers d'enquêtes socio-économiques ou médicales constituent des exemples privilégiés. Les lignes de ces tableaux sont en général des individus ou observations (il peut en exister plusieurs milliers); les colonnes sont des modalités de variables nominales, le plus souvent des modalités de réponses à des questions (cf. Lebart, 1975; Lebart *et al.*, 1995; Saporta, 1990).

3.2. Méthodes de classification

Il existe plusieurs familles d'algorithmes de classification: les algorithmes conduisant directement à des *partitions* comme les méthodes d'agrégation autour de centres mobiles; les *algorithmes ascendants* (ou encore agglomératifs) qui procèdent à la construction des classes par agglomérations successives des objets deux à deux, et qui fournissent une hiérarchie de partitions des objets. On se limitera ici à ces deux techniques de classification:

- Les groupements peuvent se faire par recherche directe d'une partition, en affectant les éléments à des centres provisoires de classes, puis en recentrant ces classes, et en affectant de façon itérative ces éléments. Il s'agit des techniques *d'agrégation autour de centres mobiles*, apparentées à la méthode des «nuées dynamiques», ou méthode «k-means», qui sont particulièrement intéressantes dans le cas des grands tableaux (Ball et Hall, 1967; Diday, 1971).
- Les groupements peuvent se faire par agglomération progressive des éléments deux à deux. C'est le cas de la classification ascendante hiérarchique qui peut fonctionner avec plusieurs critères d'agrégation. La technique d'agrégation «selon la variance»

ou de Ward est intéressante par la compatibilité de ses résultats avec certaines analyses factorielles.

Il est aussi possible d'envisager une stratégie de classification basée sur un *algorithme mixte*, particulièrement adapté au partitionnement d'ensembles de données comprenant des milliers d'individus à classer. Un des avantages des méthodes de classification est de donner lieu à des éléments (les classes) souvent plus faciles à décrire automatiquement que les axes factoriels.

Enfin, la pratique montre que l'utilisateur a intérêt à utiliser de façon conjointe les méthodes factorielles et les méthodes de classification (cf. section 4.3 ci-dessous).

4. LE TRAITEMENT GLOBAL DES DONNÉES D'ENQUÊTES

Rappelons la démarche du statisticien lors du dépouillement traditionnel d'une enquête sur ordinateur avec les outils logiciels disponibles (cf. Grangé et Lebart, 1993). Ce dépouillement met en ouvre des techniques simples, éprouvées, faciles à interpréter: les tris, les tableaux croisés, c'est-à-dire des calculs de pourcentages d'individus pour chaque modalité d'une variable nominale (avec ou sans filtre préalable) et des calculs de moyennes de variables numériques ou quantitatives (qui peuvent être ventilées selon les catégories d'une ou de plusieurs variables nominales).

Des méthodes statistiques plus élaborées viennent parfois compléter ces premiers résultats: régressions, analyses de la variance ou de la covariance, modèles log-linéaires.

Les techniques d'analyse des données (analyses descriptives multidimensionnelles) présentées en section 2 modifient profondément les premières phases du traitement des données d'enquête. Elles vont en fait bouleverser l'enchaînement des tâches, et définir une méthodologie nouvelle.

4.1. Les étapes du traitement

Dans le cadre de cette méthodologie, les étapes du traitement des données d'enquêtes sont, brièvement, les suivantes:

1) Descriptions élémentaires (tri-à-plat, histogrammes, calculs de statistiques élémentaires, moyennes, écarts-types, valeurs extrêmes, quantiles). Retour éventuel aux données de base pour une nouvelle saisie partielle ou pour des corrections.
2) *Épreuves de cohérence globale*; Épreuves d'hypothèses larges (par hypothèses larges, on entend: hypothèses générales permises par les nouveaux outils de description). Structuration des données, typologies, sélection de tableaux croisés.

3) Épreuves d'hypothèses classiques (tests statistiques usuels, régression, discrimination, analyses de la variance, modèles log-linéaires...).

4) Conclusions: Critique de l'information de base: lacunes dans le choix des variables, déséquilibre de l'échantillon ou du champ d'observation, biais ou erreurs. Choix de modèles, énoncés des résultats, rejets d'hypothèses, suggestions de nouvelles hypothèses.

La phase 2 est encore souvent absente des logiciels classiques. Lors de cette phase, la cohérence globale du recueil de données peut en effet être éprouvée de façon systématique, des panoramas globaux peuvent être dressés, permettant de critiquer l'information, mais aussi d'orienter la suite des traitements, de choisir les tableaux croisés les plus pertinents. Les typologies (classification des individus en prenant en compte simultanément plusieurs réponses ou plusieurs caractéristiques de base), les outils de visualisation (plans factoriels) fournissent de nouveaux matériaux d'analyse.

Ces opérations, intervenant au début de la chaîne de traitement, permettent de piloter la suite du dépouillement de l'enquête. Le choix des modèles n'est plus fait de façon aveugle en fonction des hypothèses de base: ces hypothèses pourront souvent être critiquées, d'autres hypothèses pourront être suggérées.

Notons que les règles d'interprétation des représentations obtenues à l'issue des techniques de réduction présentées en section 1 n'ont pas la simplicité de celles de la statistique descriptive élémentaire. Une formation et une expérience pratique s'avéreront nécessaires.

4.2. Le modèle de base: éléments actifs et illustratifs (ou supplémentaires)

Il est très intéressant de positionner dans les sous-espaces de représentation des lignes ou des colonnes supplémentaires du tableau de données (Cazes, 1981). On peut ainsi illustrer les plans factoriels par des informations n'ayant pas participé à la construction de ces plans, ce qui va avoir des conséquences importantes au niveau de l'interprétation des résultats.

Les éléments ou variables servant à calculer les plans factoriels sont appelés éiéments actifs ou variables actives: ils doivent former un ensemble homogène pour que les distances entre individus ou observations s'interprètent facilement. Ils sont en général relatifs à un même thème de l'enquête. Les éléments illustratifs peuvent être très hétérogènes.

Cette dichotomie entre *variables actives* et *variables illustratives* est du même ordre que la distinction que l'on établit entre variables exogènes (explicatives) et endogènes (à expliquer) dans les modèles de régression multiple. D'un point de vue géométrique,

les deux situations sont d'ailleurs très similaires. Les variables exogènes engendrent un sous-espace sur lequel seront projetées les variables endogènes. Les variables actives engendrent aussi un sous-espace, que l'on va réduire pour le visualiser, et c'est sur cet espace réduit que l'on projette les variables illustratives.

4.3. Complémentarité de la classification

Dans le cas du traitement statistique des fichiers d'enquêtes en vraie grandeur, la démarche précédente fondée sur des représentations graphiques a deux graves inconvénients:

1) Les visualisations sont limitées à deux ou en général à très peu de dimensions, alors que le nombre d'axes significatifs peut être plus élevé.
2) Ces visualisations peuvent inclure des centaines de points, et donner lieu à des graphiques chargés ou illisibles. Il faut donc à ce stade faire appel de nouveau aux capacités de gestion et de calcul de l'ordinateur pour compléter, alléger et clarifier la présentation des résultats.

L'utilisation conjointe de la classification automatique et des analyses factorielles permet de remédier à ces lacunes. Lorsqu'il y a trop de points sur un graphique, il paraît utile de procéder à des regroupements en familles homogènes. Mais les algorithmes utilisés pour ces regroupements fonctionnent de la même façon, que les points soient situés dans un espace à deux ou à 30 dimensions. Autrement dit, l'opération de regroupement va présenter un double intérêt: allégement des sorties graphiques d'une part, prise en compte de la dimension réelle du nuage de points d'autre part.

Une fois les individus regroupés en classes, il est facile d'obtenir une description automatique de ces classes: on peut en effet, pour les variables numériques comme pour les variables nominales, calculer des statistiques d'écarts entre les valeurs internes à la classe et les valeurs globales; on peut également convertir ces statistiques en valeurs-test et opérer un tri sur ces *valeurs-test.* On obtient finalement, pour chaque classe, les modalités et les variables les plus caractéristiques.

4.4. Sélection raisonnée des tableaux croisés et noyaux factuels

Prenons l'exemple d'une enquête nationale représentative. Étant donnée la structure de la population, les caractéristiques de base (sexe, niveau de vie, statut matrimonial, niveau d'instruction, profession...) ne sont pas indépendantes. Il est utile de décrire le réseau d'interrelations entre toutes ces caractéristiques de base, puis de positionner les autres thèmes de l'enquête en tant qu'éléments illustratifs.

Les caractéristiques des personnes qui répondent sont alors visibles immédiatement dans un cadre qui tient compte des interrelations existant entre ces caractéristiques. Les consultations classiques (sans visualisation factorielle préalable) de tableaux croisés sont en effet redondantes lorsque les caractéristiques qui servent à établir ces tableaux sont liées entre elles.

Le système de projection de variables supplémentaires permet donc d'économiser du temps et d'éviter des erreurs d'interprétation. Chaque variable illustrative fournit une information qui ne pourrait être acquise que par la lecture de nombreux tableaux croisés.

Les noyaux factuels

On désigne par noyaux factuels des groupes d'individus les plus homogènes possibles vis-à-vis de leurs caractéristiques de base.

On aimerait en effet croiser des caractéristiques telles que l'âge, le sexe, la profession, le niveau d'instruction, de façon à étudier des groupes d'individus tout à fait comparables entre eux du point de vue de leur situation objective (réaliser, dans la mesure du possible, le *toutes choses égales par ailleurs*). Mais de tels croisements conduisent vite à des milliers de modalités, dont on ne sait que faire lorsqu'on étudie un échantillon lui-même de l'ordre de quelques milliers d'individus. De plus, les croisements ne tiennent pas compte du réseau d'interrelations existant entre ces caractéristiques: certaines sont évidentes (il n'y a pas de «moins de 30 ans» retraités), d'autres ont un caractère plus statistique (il y a plus de femmes dans la catégorie «plus de 60 ans»).

Une classification des individus décrits par la batterie active des caractéristiques de base va permettre de regrouper les individus ayant, dans l'échantillon, le maximum de caractéristiques en commun. En pratique, elle fournira des regroupements opératoires en une vingtaine de classes pour un échantillon de l'ordre de 2 000 individus.

Le tableau croisant une des variables nominales de l'enquête avec la partition en noyaux factuels résume pratiquement tous les tableaux obtenus en croisant cette même variable avec chacune des caractéristiques de base. De plus, certaines interactions indécelables à partir de ces tableaux binaires peuvent être détectées.

4.5. Itération de traitements. Articulation description-inférence

La plupart des techniques évoquées plus haut, dans le cadre d'une première approche, peuvent être mises en oeuvre directement à partir de logiciels standards. Mais l'exigence de l'utilisateur croît avec la connaissance progressive qu'il acquiert de son sujet. Il lui faut croiser des variables de base, regrouper des modalités d'autres

variables, diviser en classes certaines variables continues... en somme préparer les données en vue d'analyses plus fines.

Les *opérations de recodage* font partie d'un processus itératif qui converge vers une connaissance et une assimilation optimale de l'information de base.

4.6. Estimations de données manquantes

La panoplie du statisticien contient des modèles, permettant, à partir de variables quelconques, de prévoir une variable numérique (régression, analyse de la variance et de la covariance), une variable nominale (analyse discriminante, régression logistique: cf.: Bardos, 1989; Celeux et Nakache, 1994; Hand, 1997), d'étudier les associations dans les tables de contingence (modèles d'association, modèles log-linéaires).

La régression et l'analyse discriminante par arbre (Breiman *et al.*, 1984), qui améliore les méthodes classiques de segmentation utilisées en marketing.

Enfin les réseaux de neurones (Hérault et Jutten, 1994; Thiria *et al.*, 1997) constituent des modèles souples et non-linéaires qui généralisent la plupart des méthodes précitées. Leur fonctionnement en tant que *boite noire*, la difficulté d'interprétation des paramètres, les problèmes de convergence numérique font que ces méthodes ne se substituent pas totalement aux méthodes statistiques plus classiques.

Une des difficultés majeures de l' articulation description-modèles tient au fait qu'on ne peut de façon valide tester sur des données un modèle statistique découvert sur ces mêmes données (Cox, 1977). Il va de soi que le traitement des données d'enquêtes n'est pas le seul domaine où ces problèmes se rencontrent. Des techniques du type «échantillon test» ou «validation croisée» pourront aider à contourner ces obstacles (cf. McLachlan, 1992).

Tous ces modèles permettent dans de nombreux cas d'estimer des données manquantes. Les méthodes de fusions de fichiers (cf. par exemple Aluja *et al.* 1997), permettent également de procéder à des imputations de blocs de variables.

5. VISUALISATION DE DONNÉES TEXTUELLES

Les analyses statistiques de textes peuvent intervenir à deux niveaux dans un contexte industriel et commercial: au niveau du traitement des lettres de réclamations, des cahiers de doléances ou de suggestion, au niveau de questions ouvertes dans des enquêtes postales ou téléphoniques. Les questions ouvertes les plus simples et les plus

fréquentes sont d'une part la question «pourquoi» posée après une question fermée, et d'autre part les questions du type «autre, préciser», comme item de réponse complémentaire à une question fermée. Le traitement proposé va produire de façon automatique des *mots caractéristiques* et des *réponses caractéristiques* pour diverses catégories de répondants.

5.1. Questions ouvertes dans les enquêtes

Il peut donc être intéressant, dans un certain nombre de situations d'enquête, de laisser ouvertes des questions, dont les réponses se présenteront sous forme de textes de longueurs variables. Les outils de calcul et les méthodes statistiques descriptives multidimensionnelles apportent une aide au traitement de ce type d'information, évidemment complexe. Plus généralement, le *Text Mining* désigne l'analyse exploratoire de très grands recueils de textes. Le cas des questions ouvertes est un cas favorables de text mining, puisque les textes ont une homogénéité exceptionnelle (réponses à une même question) et sont accompagnés d'informations complémentaires très riches (questions fermées).

Bien que les réponses libres et les réponses aux questions fermées fournissent des informations de natures différentes, les premières sont plus économiques que les secondes en temps d'interview et génèrent moins de fatigue. Une simple question ouverte (par exemple: «Avez-vous des réclamations à formuler concernant ce produit?») peut remplacer de très longues listes d'items. Notons que les questions ouvertes sont considérées comme peu adaptées aux problèmes de mémorisation de comportement. «Quels sont les noms des magazines que vous avez lus la semaine dernière?». Pour ces questions qui font l'objet d'enquêtes périodiques, il a été prouvé maintes fois que les questions fermées donnent des taux d'oubli plus faibles (Belson et Duncan, 1962).

5.2. Les traitements statistiques de textes

Il existe deux grandes séries d'applications des analyses statistiques de textes, selon que l'on s'intéresse à la forme ou au contenu:

- Les applications à des textes littéraires (attributions d'auteurs, datation, par exemple) qui cherchent à saisir des caractéristiques de forme et de style à partir des distributions statistiques de vocabulaire, d'indices ou de ratios, ou encore à partir de corpus partiels de mots-outil (articles, conjonction, etc.). (cf. par exemple Holmes, 1985, pour une revue de ces travaux).
- D'autre part les applications réalisées en recherche documentaire (Salton, 1988), en codification automatique, dans le traitement des réponses à des questions ouvertes, qui s'intéressent principalement au contenu, au sens des textes.

Cependant, lors du traitement statistique de réponses à des questions ouvertes, ou lors des analyses d'entretiens, le socio-linguiste peut être aussi intéressé par la forme, par les connotations véhiculées par exemple par certains synonymes, certaines tournures (cf. par exemple: Achard, 1993). Les méthodes d'analyses de réponses libres dans les enquêtes relèvent de cette seconde famille d'application (Lebart et al, 1999).

5.3. Les unités statistiques découpées dans les textes

Les formes graphiques

L'unité statistique de base est la forme graphique, suite de caractères non-délimiteurs (en général des lettres) entourée par des caractères délimiteurs (blanc, points, virgules...). Un même mot pourra souvent donner lieu à plusieurs formes graphiques, selon son cas ou son genre dans le texte. Une même forme graphique peut renvoyer à plusieurs mots (en français, avions renvoie à un nom, mais aussi au verbe avoir). Cela n'est pas toujours un inconvénient grave, car les formes graphiques ne seront pas traitées isolément. Les traitements statistiques concerneront en effet les profils de fréquences de formes graphiques, c'est-à-dire les vecteurs dont les composantes sont les fréquences de chacune des formes utilisées par un individu ou un groupe d'individus. Ces profils contiennent une information extrêmement riche. Plus précisément, les techniques mettront en évidence les différences entre profils de formes graphiques.

«Mots-outil», parties du discours

Des progrès importants ont été réalisés dans le domaine de l'analyse syntaxique automatisée des textes, comme en témoigne, par exemple l'amélioration constante des correcteurs orthographiques. Des analyseurs syntaxiques permettent de calculer la proportion de noms, de verbes, d'adjectifs, etc.

Notons que si l'isolement de mots-outil (encore appellés mots vides ou mots grammaticaux) demande une désambiguïsation du texte —cas de la forme pas en français, par exemple— il existe aussi des locutions contenant des mots pleins qui sont des substituts de mots-outil (*de façon que, en même temps que, sans oublier*, etc.).

Les unités lemmatisées

Un autre type de traitement préliminaire du texte consiste à procéder à une lemmatisation. Cette opération, difficile à réaliser de façon entièrement automatique, consiste à remplacer les formes par l'entrée du dictionnaire correspondant (infinitif pour les verbes, masculin singulier pour les adjectifs, formes non élidées à la place des formes

élidées, etc.). Elle est parfois complétée par la suppression de certains mots-outils (articles, conjonctions, etc., cf. par exemple Reinert, 1986). En documentation automatique, cela permet de travailler avec un nombre restreint de mots-clé dont les occurrences sont fréquentes. Une lemmatisation complète demande une analyse morpho-syntaxique approfondie, et ne peut être entièrement automatique (cf. Charniak, 1993). En traitement de questions ouvertes, cette opération n'est pas toujours souhaitable a priori car elle détruit certaines locutions. En revanche, elle peut intervenir comme complément, car elle fournit un point de vue différent de celui fourni par une analyse entièrement automatique sur les formes graphiques du texte. Dans le cas d'entretiens non directifs peu nombreux, la lemmatisation permet de travailler avec des seuils de fréquences de mots plus élevés que ceux nécessités par l'analyse des formes graphiques.

5.4. Les analyses statistiques; les trois outils de base

Une numérisation préliminaire (qui est aussi une compression) consiste à affecter à chaque nouvelle forme graphique un numéro d'ordre qui sera associé à toutes les occurrences de cette même forme. Ces numéros seront stockés dans un dictionnaire de formes, ou vocabulaire, propre à chaque exploitation. Les trois outils de base sont l'analyse des correspondances des tableaux lexicaux, les sélections de formes caractéristiques, les sélection de réponses modales.

a) *Analyse des correspondances des tableaux lexicaux*

Les analyses des correspondances (cf. section 3) peuvent décrire les tables de contingence croisant les réponses et les formes graphiques, ou des groupes de réponses (par exemple regroupement selon le niveau d'instruction des répondants) et les formes graphiques. Elles permettent de visualiser les associations entre mots (formes) et groupes ou modalités. Ainsi, une visualisation des proximités entre mots et catégories socioprofessionnelles pourra aider la lecture des réponses de chacune de ces catégories.

Avec ce type de représentation, la présence de mots-outils est parfaitement justifiée: si ces mots caractérisent électivement certaines catégories, ils se positionnent dans leur voisinage, et peuvent être intéressants à interpréter; si au contraire leur répartition est aléatoire, ils s'abîmeront dans la partie centrale du graphique, sans en encombrer la lecture.

b) *Formes ou segments caractéristiques (ou spécificités)*

Il est tentant de compléter les représentations spatiales fournies par l'analyse des correspondances par quelques paramètres d'inspiration plus probabiliste: les spécificités ou formes caractéristiques. Ce seront les formes «anormalement» fréquentes dans les réponses d'un groupe d'individus (cf. Lafon, 1980). Un test simple fondé sur la loi hypergéométrique permet de sélectionner les mots dont la fréquence dans

un groupe est notablement supérieure (ou inférieure pour les mots anti-caractéristiques) à la fréquence moyenne dans le corpus.

c) *Les sélections des réponses modales*

Pour un groupe d'individus donné, et donc pour le regroupement de réponses correspondant, les réponses modales (ou encore phrases caractéristiques, ou documents-type, selon les domaines d'application) sont des réponses originales du corpus de base, ayant la propriété de caractériser au mieux la classe. On peut, pour chaque regroupement, calculer la distances du profil lexical d'un individu au profil lexical moyen du groupement. On peut ensuite classer les distances par ordre croissant, et donc sélectionner les réponses les plus représentatives au sens du profil lexical, qui correspondront aux plus petites distances. On obtient ainsi une sorte de résumé des réponses de chaque regroupement, formé de réponses originales.

5.5. Stratégie de traitement

On a vu qu'il était souvent nécessaire de regrouper les réponses pour pouvoir procéder à des analyses de type statistique. Les profils lexicaux d'agrégats de réponses ont plus de régularité et de signification que ceux des réponses isolées. Ce regroupement a priori peut être réalisé à partir des variables disponibles, retenues en fonction de certaines hypothèses. Mais ceci suppose une bonne connaissance préalable du phénomène étudié, situation qui n'est en général pas réalisée dans les études dites exploratoires.

Regroupement par noyaux factuels

La technique dite des «noyaux factuels» déjà évoquée en section 4 va permettre de donner des éléments de réponse à ce problème. Étant donnée une liste de descripteurs ou de variables caractérisant les individus, le problème est de regrouper les individus en groupes les plus homogènes possibles vis-à-vis de ces caractéristiques, sans en privilégier certaines a priori. La partition obtenue est une sorte de «partition moyenne» qui résume les principales combinaisons de situations observables dans l'échantillon, et qui permet donc de procéder à des regroupements de réponses textuelles les moins arbitraires possibles.

Analyses directes sans regroupement

Une telle analyse produit une typologie des réponses, en général assez grossière, et produit de façon duale une typologie de mots ou de formes graphiques.

Il est donc possible d'illustrer ces typologies par les caractéristiques des individus interrogés qui auront le statut de variables supplémentaires ou illustratives. Ce traitement

direct des réponses pourra conduire à la réalisation d'un post-codage partiellement automatisé.

5.6. Conclusions sur les analyses de textes

Il s'agit avant tout d'une confrontation de questions ouvertes et de questions fermées. L'analyse est essentiellement différentielle, comparative, et en cela se distingue de l'analyse de contenu classique. Elle ne vise en effet qu'à décrire les contrastes entre plusieurs textes, ces textes étant les réponses originales, ou des regroupements de réponses réalisés à partir des questions fermées de l'enquête.

Pour une question ouverte et pour une partition de la population, on obtient donc, de façon intégralement automatisable:

- Une visualisation des proximités entre formes et catégories, par analyse des correspondances du tableau lexical agrégé, éventuellement complétée par une visualisation similaire des proximités entre segments et catégories.
- Les formes (et/ou segments) caractéristiques de chaque catégorie.
- Les réponses modales de chaque catégorie.

Ces résultats, obtenus sans codification ni intervention manuelle, fournissent des compléments et donnent des éléments critiques nouveaux pour juger à la fois la cohérence et la pertinence du questionnement, la compréhension des réponses, ainsi que le niveau d'implication ou de participation des répondants. Ils participent donc à l'amélioration de la qualité de l'information, et fournissent des éléments originaux au dossier des analyses de satisfaction.

6. PROBLÈMES DE QUALITÉ D'INFORMATION

Les visualisations de variables numériques ou textuelles qui viennent d'être évoquées permettent de prendre en compte certaines déficiences de l'information de base (non-réponses par exemple), ainsi que des variables de contrôles liées à la qualité du recueil de l'information de base (cf. ASU, 1992).

Conjectures sur les non-réponses

Les non-réponses sont des modalités comme les autres, qui peuvent être positionnées dans les espaces factoriels des thèmes, comme dans les espaces de la structure de base. Le traitement des non-réponses qui se prête mal aux tests statistiques usuels reçoit ici

une importante contribution, dans la mesure où l'on peut étudier le contexte de ces refus ou lacunes, soit en termes de caractéristiques des répondants, soit à partir des réponses effectives à d'autres thèmes.

Le positionnement des variables techniques

Dans l'espace des caractéristiques de base ou dans celui des principaux thèmes, que ceux-ci soient résumés par un plan factoriel ou par une partition, il est possible de placer les modalités de variables nominales dites «techniques» telles que: numéro ou nom de l'enquêteur, caractéristiques diverses de l'enquêteur, heure de l'interview, lieu et durée de l'interview, appréciation de l'enquêteur ou de l'enquêté sur l'interview, etc.

On obtient ainsi un panorama de la fabrication de l'information, permettant de rapprocher globalement les circonstances des interviews et les caractéristiques des personnes interrogées. Cette confrontation permet souvent d'apprécier la validité des données de base et de nuancer l'interprétation des résultats.

Les questions ouvertes

Alors que la question ouverte *pourquoi?* après une question fermée permet de vérifier la compréhension de la question, des questions de commentaires libres à l'issue de l'interview (questions qui peuvent être posés à l'enquêté, mais aussi à l'enquêteur) permettent de critiquer aussi bien le questionnaire que les conditions de sa passation.

Finalement, dans le traitement des données d'enquêtes l'approche globale et exploratoire est un élément important d'une *démarche qualité.* En effet, détecter des patterns, c'est évidemment dans une première phase détecter des anomalies, des incohérences, des points aberrants (*outliers*), des hétérogénéités inattendues. Dans le cas des données d'enquêtes, particulièrement en présence de questions ouvertes, cela peut amener non seulement une critique de la réalisation pratique de l'enquête et du recueil de l'information sur le terrain, mais cela peut dans certains cas aller jusqu'à une remise en cause de la conception de l'enquête et de son questionnaire.

RÉFÉRENCES

Achard, P. (1993). *La sociologie du langage.* PUF. Paris.

Aluja Banet, T., Rius, R., Nonell, R. and Martínez-Abarca, M.J. (1997). «Data Fusion and File Grafting», *Proc of the IV ème Congrès International d'Analyse Multidimensionnelle des Données. NGUS-97,* Bilbao, 22-28.

ASU, (Lebart, L., ed.) (1992). *La qualité de l'information dans les enquêtes.* Dunod, Paris.

Ball, G.H. and Hall, D.J. (1967). «A clustering technique for summarizing multivariate data», *Behavioral Sciences*, 12, 153-155.

Bardos, M. (1989). «Trois méthodes d'analyse discriminante», *Cahiers Economiques et Monétaires*, 33, 151-190.

Belson, W.A. and Duncan, J.A. (1962). «A Comparison of the check-list and the open response questioning system», *Applied Statistics*, 2, 120-132.

Benzécri, J.P. (1973). *L'Analyse des Données. Tome 1: La Taxinomie. Tome 2: L'Analyse des Correspondances*, (2de. éd. 1976). Dunod, Paris.

Breiman, L., Friedman, J.H., Ohlsen, R.A. and Stone, C.J. (1984). *Classification and Regression Trees*, Wadsworth, Belmont.

Cazes, P. (1981). «- Note sur les éléments supplémentaires en analyse des correspondances», *Les Cahiers de l'Analyse des Données*, 1, 9-23; 2, 133-154.

Celeux, G. and Nakache, J.P. (eds). (1994). *Analyse discriminante sur variables qualitatives*, Polytechnica, Paris.

Charniak, E. (1993). *Statistical Language Learning*, The MIT Press, Cambridge.

Cox, D.R. (1993). «The role of significance tests», *Scandinavian Journal of Statistics*, 4, 49-70.

Diday, E. (1971). «La méthode des nuées dynamiques», *Revue Statist. Appl.*, 19, 2, 19-34.

Fayyad, U., Piateski-Schapiro, G., Smyth, P. and Uthurasamy, R. (Eds) (1996). *Advances in Knowledge Discovery and Data Mining*. AAAI Press/The MIT Press.

Grangé, D. et Lebart, L. (1993). *Traitement statistique des enquêtes.* Dunod, Paris.

Hand, D.J. (1997). *Construction and Assessment of Classification Rules.* J. Wiley, Chichester.

Hand, D.J. (1998). «Data mining: Statistics and more?», *The American Statistician* (May issue).

Hérault, J. et Jutten, C. (1994). *Réseaux neuronaux et traitement du signal.* Hermès. Paris.

Holmes, D.I. (1985). «The analysis of literary style - A Review», *J.R. Statist. Soc.*, 148, 4, 328-341.

Hotelling, H. (1933). «Analysis of a complex of statistical variables into principal components», *J. Educ. Psy.*, 24, 417-441, 498-520.

Lafon, P. (1980). «Sur la variabilité de la fréquence des formes dans un corpus», *Mots*, 1, 127-65.

Lebart, L. (1975). «L'orientation de dépouillement de certaines enquêtes par l'analyse des correspondances multiples», *Consommation*, 2, 73-96.

Lebart, L., Morineau, A. y Fenelon, J.P. (1985). *Tratamiento Estadístico de Datos*. Marcombo, Barcelona.

Lebart, L., Salem, A. y Becue, M. (1999). *Análisis Estadístico de Textos*. Ediciones Milenio, Lleida.

McLachlan, G.J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. J. Wiley, New York.

Reinert, M. (1986). «Un Logiciel d'analyse lexicale», *Les cahiers de l'analyse des données*, 4, Dunod, 471-484.

Salton, G. (1988). *Automatic Text Processing: the Transformation, Analysis and Retrieval of Information by Computer*, Addison-Wesley, New York.

Saporta, G. (1990). *Probabilités, analyse des données et statistiques*. Technip, Paris.

Schuman, H., Presser (1981). *Questions and Answers in Attitude Surveys*. Academic Press, New York.

Thiria, S., Gascuel, O., Lechevallier, Y. et Canu, S. (1997). *Statistique et méthodes neuronales*. Dunod, Paris.

ENGLISH SUMMARY

QUALITY OF DATA IN SAMPLE SURVEYS

L. LEBART
Ecole Nationale Supérieure des Télécommunications*

This paper aims at highlighting the contribution of both exploratory multivariate methods and textual analysis to the assessment of data quality in surveys.
The basic idea is that each step in carrying out a survey, from the conception of the questionnaire to the very end of the interview, can provide us with a series of control variables. These variables could be numerical (e.g. duration of each interview), nominal (e.g. circumstances, location of the interview) or textual (e.g. open ended questions about the content of the survey).
The methods mentionned above allow for confrontations of initial data files and the control file, leading to an assessment of the obtained information.

Keywords: Data quality, Socioeconomic Surveys, Exploratory Multivariate Data Analysis, Textual Data Analysis.

AMS Classification: 62H25, 62H30, 62D05, 62P25

* Ludovic Lebart. Centre National de la Recherche Scientifique. Ecole Nationale Supérieure des Télécommunications. 46 Rue Barrault. 75013 Paris.

- Received December 1998.
- Accepted May 1999.

1. INTRODUCTION

Survey data processing will involve exploratory multivariate data analysis and textual data analysis, that are closely related to Data Mining and Text Mining.

2. BASIC CONTROLS OF A SURVEY

When carrying out a sample surveys, each step leads to specific controls. Besides, it provides the researcher with a set of control variables that can be added to the variables relating to the content of the survey.

3. VISUALIZATION OF SURVEY DATA

The techniques of visualization dealt with in this section are those adapted to the processing of large data tables. Principal axes methods (Principal component analysis, simple and multiple correspondence analysis) are the most common methods.

Among clustering techniques, the k-means method is certainly the most largely used, thanks to its ability to tackle huge data sets.

4. THE GLOBAL PROCESSING OF SURVEY DATA

The most common phase of the processing requires using the demographic variables to design the frequency distributions and the cross-tabulations. These same demographic variables can also be used to build a typology of the respondents. In fact, several typologies of the respondents can be derived from several sets of the so-called active variables (homogeneous set of variables relating to a particular theme of the questionnaire).. Each of these step defines a specific point of view. The basic procedure consists in positionning supplementary variables in the maps corresponding to each typology.

The complementarity of clustering is stressed, due to the limitation of the visualization to a small number of dimensions. Moreover, these visualizations can include several hundred points, and give rise to crowded or illegible graphs, and to lengthy lists of coordinates.

Thus it is important to make use of the data management and computational capabilities of the computer to complete and clarify the presentation of the results. The combined use of clustering methods and correspondence analysis can fill in the gaps. When there

are too many points on a graph it may be useful to group the data into homogeneous families. The algorithms used for developing these groups work the same way whether the points are located in a two-dimensional or a ten-dimensional space.

In other words, the process has two objectives: to minimize graphical printouts, on the one hand, and to work with the real dimensionality of the configuration of points, on the other hand. Once the individuals have been grouped into clusters, it is straightforward to obtain a description of these clusters: indeed, statistics related to differences between internal values for each cluster and overall values for the sample can be calculated for numerical variables and categorical variables. These statistics can also be converted into *test-values* and sorted on these test-values.

Finally, the most characteristic response categories and variables can be displayed for each cluster.

All the technical variables collected can be projected as illustrative (or: supplementary) variables. This helps us to detect possible interactions between some aspects of the content of the surveys and some apparently lateral features such as the gender of the surveyor, the day or the time of the interview, the presence of other persons during the interview, the duration of the interview, the opinion of the surveyor about the interview, etc.

5. VISUALIZING TEXTUAL DATA

Open ended questions in surveys can contribute to the assessment of data quality. Open-ended questions can be found in the questionnaires of many surveys performed in socioeconomy, epidemiology, advertising, business or political marketing. They become an essential part of these questionnaires when the scope of research goes beyond a simple tally, and when a complex and unknown topic is being explored.

Open ended questions are less costly in terms of interview time, and generate less fatigue and tension. But a fundamental advantage lies in their use *to probe the response to a closed-end question.*: This is the follow up additional question «*Why?*». Explanations concerning a response already given have to be provided in a spontaneous fashion. A battery of items might suggest new ideas that would only mar the authenticity of the explanation given.

The main tools are the *correspondence analysis of lexical tables*, showing on a series of mappings the associations between the words used and the characteristics of the respondents, the *characteristic words*, (list of words or phrases most associated with a specific category of respondent) and the characteristic responses, pinpointing the responses most characteristic of a category.

6. DATA QUALITY ASSESSMENT

The powerful tools briefly described above allow for a new level of assessment of consistency in survey data processing. All sorts of control variables, no-response items, open responses (through characteristics words or responses) can now be positionned at a low cost among the variables relating to the content of the survey.

QÜESTIIÓ, vol. 23, 2, p. 343-362, 1999

DISEÑO DE ENCUESTAS DE OPINIÓN: BARÓMETRO CIS

V. MARTÍNEZ
Centro de Investigaciones Sociológicas*

El diseño de las encuestas de opinión puede incluirse dentro de los diseños generales de las encuestas sobre población, cuyo objetivo es la obtención de información mediante la observación de los individuos que la componen. Las características diferenciales con respecto a los diseños generales son la utilización de tamaños de muestra reducidos y una selección no probabilística de los individuos. En este sentido, la calidad de los resultados obtenidos suele ser cuestionada y, por tanto, las extrapolaciones sobre la población deben ser tomadas con cautela. En este artículo se revisan y valoran estos preceptos en el marco de la encuesta de opinión que elabora mensualmente el Centro de Investigaciones Sociológicas-CIS (conocida como Barómetro CIS), cuya versión inicial fue presentada en las Primeras Jornadas Internacionales «Generación de Información Estadística: Calidad y Limitaciones» (Barcelona, 30 noviembre-1 diciembre 1998) que organizó la Xarxa Temàtica «Enquestes i Qualitat de la Informació Estadística».

Surveys Design for Opinions: Barometer CIS

Palabras clave: Marcos, cuestionario, diseño, tamaño muestral, unidades de muestreo, conglomerados, estratificación, afijación, tamaño del conglomerado, cuotas, estimadores, costes, calidad.

Clasificación AMS: 62D05, 62P25.

* Departamento de Banco de Datos. Centro de Investigaciones Sociológicas (CIS). Montalbán, 8. 28014 Madrid. E-mail: vmartinez@cis.es.

- Recibido en diciembre de 1998.
- Aceptado en abril de 1999.

1. INTRODUCCIÓN

El diseño de las encuestas de opinión se sitúa dentro de los que pueden denominarse diseños generales sobre la población, dado que su objetivo es medir el estado de la opinión pública en relación con la situación social, política, económica, etc. del país. De este modo, no existe ningún interés explícito en un subconjunto de la población o en un conjunto de variables, excepto en ciertas situaciones como elecciones, u otros fenómenos de mayor incidencia social donde, para su mejor estudio, se acota la población y se precisan las variables objetivo.

Por tanto, el contenido básico de las encuestas de opinión responde a fenómenos coyunturales de los cuales se facilita una «foto fija» de la opinión de la población en un momento del tiempo. Este marcado carácter coyuntural no impide establecer un bloque fijo de preguntas en el cuestionario con el fin de poder elaborar series e indicadores de algunas de las variables que faciliten información sobre la evolución de la opinión, así como algunos cruces con variables estructurales como pueden ser el sexo, la edad o la situación laboral.

Uno de los inconvenientes de las encuestas de opinión es que las extrapolaciones sobre la población, al emplear esquemas no probabilísticos en alguna de sus fases así como tamaños de muestra reducidos, deben de ser tomadas con la máxima cautela. De hecho, los resultados que se obtienen sobre ciertas variables estructurales estudiadas por otras encuestas muestran diferencias muy significativas; éste sería el caso de los resultados sobre situación laboral al ser comparados con la Encuesta de Población Activa –EPA– elaborada por el Instituto Nacional de Estadística.

Por último, los estudios sobre la opinión pública tienen un alto grado de oportunidad, en el sentido de que la encuesta debe realizarse cuando el fenómeno está vigente o, de lo contrario, no podrá ser recogido adecuadamente.

En este artículo se recogen los diferentes elementos que caracterizan a las encuestas de opinión y en particular a la encuesta mensual del CIS más conocida por Barómetro CIS.

2. UNIVERSO DE LA ENCUESTA

El universo de la encuesta se corresponde con la población de la cual quiere obtenerse información. De este modo puede hablarse de universo en general cuando la investigación viene referida a la población en su conjunto o universo particular cuando dicha investigación se pretende realizar sobre una población específica.

Así, cuando la investigación viene referida a un universo general, como es el caso de la encuesta mensual del CIS, su objetivo estará en investigar aquel grupo de la población que, en un sentido amplio, se considera más adecuado para responder a las necesidades de información y, por tanto, de poder ser preguntado sobre el conjunto de cuestiones que se planteen. En términos generales para este tipo de encuestas el grupo que se considera más adecuado es aquel que tiene derecho de voto, es decir, la población con 18 años y más.

Por otro lado, cuando la investigación se concentra en un subconjunto particular de la población deberá definirse, de manera muy precisa, cuál es el grupo de interés, ya sean poblaciones específicas como profesionales de un sector, diputados electos, etc. o bien subpoblaciones caracterizadas a través de ciertas variables: grupos de edad (jóvenes, mayores de 65 años, etc.), una situación socioeconómica, etc.

3. MARCOS DE MUESTREO

Definidos los objetivos de la encuesta, el investigador se enfrenta al primer problema que va a condicionar la totalidad de su trabajo y, en particular, de los diseños muestrales alternativos que puede considerar para alcanzar tales objetivos. Este problema es la elección del marco de muestreo.

Dado que la finalidad del diseño es la selección de un conjunto finito de unidades, el marco de muestreo puede definirse como la población finita de unidades donde se aplican las reglas de selección determinadas por el diseño muestral. Si el diseño es probabilístico, permitirá conocer la probabilidad de obtener tal subconjunto de unidades, en otro caso tal probabilidad no podrá ser conocida.

En general, el investigador no dispone de un listado exhaustivo de las unidades o personas que constituyen la población objetivo de su investigación, debiendo recurrir a las listas o directorios existentes que mejor se corresponden con la población a investigar. Por tanto, resulta habitual el empleo de una población instrumental que posibilite la realización de su trabajo, a esta población instrumental se la denomina población investigada o población de inferencia y constituye el marco de la encuesta.

El marco empleado puede caracterizarse por su estructura, es decir, por la información que contiene de la población a investigar. Esta información determinará el tipo de diseño muestral, así como los procedimientos de estimación que pueden ser utilizados. Marcos simples, sin información auxiliar, conducirán a diseños simples, y marcos complejos pueden facilitar la realización de diseños complejos.

Una vez que el investigador dispone del marco más adecuado, y previa la realización de cualquier otro tipo de actividad, es necesaria su depuración y actualización, en el

sentido de aplicar en la investigación la información más actual que permita la mejor aproximación a la población objetivo. Esta labor de depuración consiste en revisar su contenido, eliminar los elementos duplicados, etc.

Ejemplos de marcos utilizados en las encuestas de opinión son los censos de población y las listas telefónicas

* *Censos de población*

Este tipo de marco ofrece información de carácter general, dado que la legislación vigente no permite la difusión de datos individuales. En el caso español, la información a nivel más desagregado se obtiene de la sección electoral y, a partir de ésta, puede generarse a nivel de distrito, municipio, comarca, etc., mediante agregación simple. Esta información implica que el investigador no tiene otra opción que trabajar bajo un esquema de muestreo por conglomerados aunque, como se verá más adelante, el utilizar información adicional puede permitir la aplicación de esquemas mixtos de muestreo.

El empleo de la información censal implica una relación no directa entre las unidades seleccionadas e investigadas. Por ejemplo, si el marco es un listado de municipios, la selección de cualquiera de ellos implica una relación entre las unidades del marco y las unidades de análisis (individuos) que no es unívoca sino de uno a varios.

* *Listados telefónicos*

En la actualidad este tipo de listado es muy empleado para la realización de encuestas telefónicas, en particular aplicando los sistemas CATI[1]. Estas listas permiten la identificación directa de individuos, de tal modo que existe una relación unívoca entre registro del listado y elemento de la población que puede ser muy válida para encuestas de carácter general sobre la población, siempre que el grado de cobertura del servicio telefónico sea adecuado. En general, el empleo de estos listados necesita un gran esfuerzo para eliminar de los mismos elementos que no forman parte de la población, duplicados, etc.

En resumen,

- la mayor o menor aproximación del marco disponible a la población objetivo facilitará o complicará el alcanzar los objetivos marcados en el proyecto de investigación;

[1] Corresponde a las siglas inglesas Computer Assisted Telephone Interviewing.

- si la estructura del marco contiene información suplementaria sobre ciertas características de sus unidades, esta información permitirá establecer ciertas clasificaciones y, por tanto, realizar un diseño muestral más eficiente.[2]

4. CUESTIONARIOS

La elaboración de los cuestionarios debe dirigirse al cumplimiento de los objetivos de la investigación, mediante la elaboración de un conjunto de preguntas que permitan la medición acurada de la información necesaria.

Ciertas reglas deben ser seguidas para abordar con cierto grado de éxito la encuesta. Entre ellas cabe destacar las dos siguientes:

- Las preguntas deben formularse teniendo en mente al entrevistado.
- Las preguntas deben situarse de tal forma que permita un posterior tratamiento del flujo de las respuestas.

En este sentido hay que favorecer aquellas preguntas que no requieren un esfuerzo más que razonable por parte del entrevistado, tanto en el tipo de preguntas a realizar (ej. el recuerdo de hechos o acciones pasadas, cuestiones íntimas, etc.) como de duración global del cuestionario[3].

El investigador debe tener en consideración todo aquello que facilite la recogida de la información del entrevistado, en este sentido, un cuestionario de autocumplimentación en una encuesta por correo deberá incluir aquellos apoyos que sirvan para la correcta interpretación de la pregunta planteada.

5. RECOGIDA DE LA INFORMACIÓN

Como ya se avanzó en el punto anterior la recogida de la información requiere del cuidado del investigador, dado que los diferentes métodos a aplicar implican la elaboración del cuestionario adecuado, así como del desarrollo de otros elementos que permitan recoger la información con el mayor grado de calidad.[4]

---

[2]El término eficiente debe entenderse en un doble sentido: por un lado, el diseño debe conducir a una muestra con el menor error posible y, por otro lado, debe ser aquella que tenga el menor coste por unidad.

[3]El no tomar en consideración estos aspectos puede conducir a problemas graves de falta de respuesta.

[4]El objetivo global en el diseño del cuestionario, como en el método de recogida, es limitar el esfuerzo que debe realizar el entrevistado en el momento de facilitar la información.

En general, se distinguen cuatro métodos básicos de recogida: observación directa, entrevista personal, entrevista telefónica y correo. Algunas de las características de estos métodos son las siguientes:

a) Observación directa

Se aplica cuando el entrevistado no facilita la información que de él se requiere. Ejemplos de este tipo son los dispositivos de medición de las audiencias televisivas, el estudio de conductas donde interesa conocer lo que el individuo hace y no lo que dice que hace, etc.

b) Entrevistas personales

Es el método más utilizado en las encuestas dirigidas a los hogares a pesar de su elevado coste y de incorporar un elemento intermedio que es el entrevistador. El entrevistador es el elemento básico en este tipo de encuesta, debiendo recibir la formación necesaria para que conduzca la entrevista de manera que el entrevistado se sienta lo más cómodo posible, pueda resolver sus dudas de forma clara y rápida y evite ciertas formas de plantear las preguntas que puedan influenciar la respuesta y, por tanto, la introducción de sesgos.

Por otro lado, el investigador debe elaborar los manuales de apoyo para los entrevistadores, así como realizar una formación específica para la encuesta de que se trate. El objetivo en ambos casos es establecer claramente las reglas a aplicar y evitar que se den soluciones diferentes a problemas similares por parte del equipo de entrevistadores.

c) Entrevistas telefónicas

Análogamente al punto anterior, el elemento básico es el entrevistador, aunque empleando un medio diferente como es el teléfono. Aquí el entrevistador se encuentra limitado por el medio empleado.

d) Cuestionarios por correo

En este tipo de encuestas la interrelación entre personas no se produce, siendo necesario que el cuestionario enviado sea lo más claro posible e incluya todos los apoyos que faciliten su correcta cumplimentación. Por otro lado, debe incorporar algún incentivo que favorezca un mayor nivel de respuesta.

6. DISEÑO DE LA ENCUESTA

El diseño de una encuesta de opinión puede resumirse en la aplicación de dos reglas simples que favorezcan

- el control de la dispersión de la muestra en el espacio;
- la realización en un período de tiempo reducido.

Ambas reglas pretenden la coordinación y el mejor reparto de la carga de trabajo de la encuesta, de manera que los equipos de entrevistadores no se vean obligados a realizar un número elevado de desplazamientos o entrevistas que afectaría tanto a la calidad de la información recogida como a la oportunidad de la investigación por muestreo.

Estas reglas se aplican sobre el marco disponible, el cual, dependiendo de la información que contenga, permitirá al investigador la realización de un diseño simple o complejo. Un **diseño simple** puede establecerse si el marco consiste en un listado de las unidades de la población objetivo. A partir de la lista puede procederse a realizar una selección de las unidades, ya sea mediante una tabla de números aleatorios o una selección sistemática con arranque aleatorio.

En general, como ya se ha comentado, el investigador no suele disponer de la lista de unidades de la población objetivo. En esta situación, la opción más razonable es trabajar sobre la base de la información agregada existente sobre las unidades y, fijando una serie de pasos intermedios, llegar a las unidades finales de interés para la investigación. El resultado de emplear este esquema de muestreo conduce a los denominados **diseños complejos** sobre los cuales se consideran los siguientes aspectos: determinación del tamaño muestral, unidades de muestreo, estratificación de las unidades de primera etapa, distribución del número de entrevistas, determinación del número de entrevistas por unidad primaria y unidad secundaria, selección de las unidades de primera, segunda y tercera etapa, selección de las unidades finales y estimadores.

6.1. Determinación del tamaño muestral

Para establecer el tamaño de la muestra el investigador debe considerar dos aspectos: la precisión que se pretende obtener y el coste que supone su realización.

Con respecto de la precisión se deberá fijar un límite de error admisible como resultado de la elección consciente de estudiar un subconjunto de la población en lugar de la totalidad y, por otro lado, dado que los presupuestos no son ilimitados, se tenderá a llegar a un compromiso entre precisión y coste de forma que la encuesta sea útil para los objetivos planteados. En el caso que las restricciones presupuestarias no posibilitasen la realización de la encuesta bajo ciertas garantías de error, entonces, la investigación por muestreo debe ser abandonada.

Dado que el investigador puede plantear diseños alternativos que conducen a diferentes niveles de error y costes, tenderá a las siguientes opciones:

- Ante diseños de precisión similar se elegirá aquel que tenga un menor coste asociado.
- Entre diseños de coste similar se elegirá aquel que tenga asociado el mayor nivel de precisión.

En las encuestas de opinión, dado que el tamaño de la población a investigar es normalmente grande y la fracción de muestreo muy baja, se emplean las fórmulas correspondientes a un muestreo aleatorio con reemplazamiento y probabilidades iguales.

Así, en el caso de la estimación de proporciones, el tamaño muestral viene dado por

$$n = \frac{k^2 \cdot PQ}{\epsilon_a^2} \tag{1}$$

donde,

n, tamaño de la muestra
k, valor correspondiente al nivel de confianza establecido
P, valor de la proporción en la población
$Q = (1 - P)$
ϵ_a, nivel de error absoluto; definido como la diferencia entre el estimador y el valor desconocido
$\epsilon_a = \hat{P} - P$

A partir de la fórmula anterior se suelen establecer hipótesis sobre los valores de k y P, de manera que se obtenga una relación entre el tamaño de la muestra y el nivel de error. Las hipótesis más habituales son:

- Intervalo de confianza para la estimación 95%.
- Considerar los valores de P y Q que hacen máxima la varianza.

De este modo, $k = 2$ y $P = Q = 1/2$ y, sustituyendo estos valores en (1)

$$n = \frac{1}{\epsilon_a^2} \tag{2}$$

Por lo tanto, fijando el valor de n o de error ϵ_a, se obtendrá el nivel máximo de error ϵ_a[5] o el tamaño de muestra necesario, respectivamente.

[5]En el anexo que figura al final del artículo se incluyen los tamaños de muestra para diferentes valores de P y ϵ_a. Si se mantiene el nivel de error el tamaño de muestra aumentará conforme P se acerca al 50%; por otro lado, si se mantiene el valor de P el tamaño muestral decrecerá conforme aumenta el nivel de error.

N	$\epsilon_a(\%)$
40.000	0,5
10.000	1
2.500	2
1.111	3
625	4
400	5

En la encuesta mensual del CIS el tamaño de muestra utilizado es 2.500 entrevistas, que se corresponde con un error del 2%, bajo las hipótesis comentadas.

6.2. Unidades de muestreo

A partir de la información contenida en el marco disponible, el investigador establece una jerarquía de unidades según su mayor/menor nivel de agregación que facilite una relación entre las unidades más agregadas y las unidades finales que son los individuos. A partir de la jerarquía se seleccionarán aquellas unidades más adecuadas para el desarrollo del diseño muestral, las cuales definirán las diferentes etapas del muestreo.

Si el investigador dispone de la información censal puede definir diferentes unidades estableciendo una jerarquía que comprendería distintos niveles de agregación como son la provincia, el municipio, el distrito y la sección. A partir de estas unidades el investigador decide qué unidades resultan más adecuadas para el mejor desarrollo de la investigación.

En la encuesta mensual del CIS, las unidades utilizadas y su jerarquía son los siguientes:

Jerarquía	**Unidades**
Primera etapa	Municipios
Segunda etapa	Secciones
Tercera etapa	Viviendas
Última etapa	Individuos

6.3. Estratificación de las unidades de primera etapa

Si el investigador realiza una selección aleatoria de las unidades contenidas en el marco, se encuentra que tales unidades son muy heterogéneas con respecto a las características

estructurales así como a las variables que pretende medir. En este sentido, si se consideran como unidades de primera etapa los municipios, las características de una gran urbe serán muy diferentes de las de un municipio de pequeño tamaño.

Con el objetivo de evitar, en el mayor grado posible, la heterogeneidad comentada, el investigador puede establecer una estratificación de las unidades de manera que éstas formen parte de estratos con características más homogéneas. Este proceso de estratificación puede realizarse en base a diferentes criterios. Por ejemplo, si las unidades consideradas son los municipios puede utilizarse la población de derecho, algún tipo de ámbito geográfico, etc. o una combinación de criterios.

Los criterios utilizados en la encuesta mensual del CIS son los siguientes:

- Población de derecho de los municipios

Estrato	**Población de derecho (P)**
1	$P \leqslant 2.000$
2	$2.000 < P \leqslant 10.000$
3	$10.000 < P \leqslant 50.000$
4	$50.000 < P \leqslant 100.000$
5	$100.000 < P \leqslant 400.000$
6	$400.000 ,< P \leqslant 1.000.000$
7	$P > 1.000.000$

- Comunidad autónoma: 17 estratos
- Área metropolitana, que se define por la pertenencia o no del municipio al área de influencia de un municipio importante.

Por lo tanto, se establecen 238 estratos teóricos (7x17x2), algunos de los cuales están vacíos.

6.4. Distribución del número de entrevistas

Una vez establecida la estratificación de las unidades se puede proceder a la distribución de la muestra en los estratos o afijación de la muestra. En esta etapa el investigador debe decidir qué tipo de distribución considera más adecuada para los objetivos de la investigación. En general se distinguen tres afijaciones básicas.

a) Proporcional: el número de entrevistas en cada estrato se asigna en función de la proporción de la población objetivo en dicho estrato:

$$n_h = n \cdot \frac{N_h}{N} = n \cdot W_h$$

donde,

n_h, número de entrevistas a realizar en el estrato h
n, tamaño de muestra (ver apartado 6.1)
N, número de individuos en la población objetivo
N_h, número de individuos de la población objetivo en el estrato h
W_h, proporción de población objetivo correspondiente al estrato h

b) Uniforme: se asigna el mismo número de entrevistas en cada uno de los estratos.

c) Mixta: se realiza una combinación entre los esquemas a) y b) anteriores. Por ejemplo, puede considerarse una distribución uniforme a nivel de las comunidades autónomas y realizar una distribución proporcional según el tamaño de los municipios.

Con el esquema proporcional, que es el utilizado en la encuesta mensual del CIS, la distribución de las entrevistas refleja la proporción de población que tiene cada comunidad autónoma; por consiguiente, aquéllas con mayor población tendrán asignado un mayor número de entrevistas y aquéllas con menor población tendrán un número menor. De este modo, considerando la relación entre tamaño de muestra y error dada por (2), los niveles de error serán muy diferentes entre las comunidades autónomas.

En consecuencia, si uno de los objetivos de la investigación es la obtención de información desagregada a nivel de comunidad autónoma, el esquema proporcional no será el más adecuado y, en estas situaciones, el investigador deberá decidir qué grado de desproporción introduce en la distribución de la muestra de forma que la información obtenida tenga un mayor grado de comparabilidad. Así, el esquema uniforme favorece que los errores sean muy similares y que la información obtenida tenga un alto grado de comparabilidad. En el esquema mixto el nivel de error se situará entre los esquemas proporcional y uniforme[6].

6.5. Número de entrevistas por unidad primaria

A partir de la afijación de la muestra en los estratos debe fijarse el número de entrevistas a realizar en cada una de las unidades primarias y así determinar el número de unidades

[6]El empleo de esquemas desproporcionados obligará al investigador a establecer unos coeficientes de ponderación que «re-equilibren» la muestra en relación con la estructura de la población. Estos coeficientes serán más elevados conforme mayor sea la desproporción establecida.

que deberán ser seleccionadas. En el caso de que el número de entrevistas por unidad sea idéntico se estará ante un muestreo por conglomerados de igual tamaño, en otro caso se estará ante un muestreo por conglomerados de distintos tamaños.

Considerando la estratificación realizada para la encuesta mensual del CIS, en algunos casos no es necesario fijar este número de entrevistas dado que en ciertos estratos sólo existe un municipio; ésta sería la situación de las grandes urbes como Madrid o Barcelona. Cuando el número de municipios dentro del estrato permite la realización de una selección los procesos establecidos consideran un mínimo de dos municipios con el fin de no concentrar las entrevistas en un único municipio. De este modo, y con el fin de facilitar los ajustes en el reparto de entrevistas, se establecen unos intervalos según el tamaño de los municipios:

Estrato	Mínimo	Máximo
1	10	12
2	11	14
3	13	19
4,5,6	18	30

Por ejemplo, si para el cruce de comunidad autónoma y tamaño hay que realizar 30 entrevistas y los municipios pertenecen al estrato 1, se seleccionarán un total de tres municipios.

6.6. Número de entrevistas por unidad secundaria

Análogamente a lo ya comentado en el punto anterior, debe fijarse el número de entrevistas a realizar en cada unidad secundaria. En la encuesta mensual del CIS, las unidades de segunda etapa son las secciones censales que pueden considerarse unidades homogéneas con respecto de la población[7] y, por lo tanto, no se presenta de forma tan acusada el problema de la heterogeneidad comentado para los municipios. En esta situación resulta adecuado fijar un número similar de entrevistas en cada unidad, siendo el tamaño medio utilizado de diez entrevistas por sección.

Para el ejemplo considerado, en cada municipio se seleccionaría una sección donde se realizarían las 10 entrevistas.

[7] «Cada sección incluye un máximo de 2.000 electores y un mínimo de 500. Cada término municipal cuenta al menos con una sección» Artículo 23.2 de la Ley Orgánica 5/85 (Ley Orgánica del Régimen Electoral General-LOREG).

6.7. Selección de las unidades de primera y segunda etapa

Una vez determinado el número de municipios a seleccionar en cada uno de los estratos, se genera un procedimiento aleatorio de selección que puede considerar la selección en base a probabilidades iguales o desiguales.

En la encuesta mensual del CIS el proceso de selección de municipios y secciones se lleva a cabo mediante la selección proporcional al número de personas de 18 años y más existentes en cada municipio y sección, respectivamente.

6.8. Selección de las unidades de tercera etapa

A partir de la selección de las secciones se procede a la selección de las viviendas, la cual puede realizarse de diversas formas.

Si el marco empleado contiene información relativa al número total de viviendas, dentro de la sección puede realizarse una selección sistemática de las mismas; por ejemplo, si el número de viviendas es de 200 se realizará una entrevista cada 20 viviendas. Alternativamente, si no se dispone del total pueden establecerse dos estrategias:

- Realizar un barrido de la sección, donde se determina el total de viviendas y, a continuación, aplicar una selección sistemática.
- Establecer una ruta aleatoria dentro de la sección.[8]

6.9. Selección de las unidades finales

La selección de los individuos en las encuestas donde se realiza una entrevista personal puede realizarse a través de los mecanismos siguientes:

a) Selección aleatoria de individuos a partir de una lista existente.

b) Selección aleatoria de individuos a partir de una lista que se confecciona en el momento del barrido de las viviendas.

c) Selección mediante tablas que consideran la composición del hogar seleccionado en el momento de realizar la encuesta.

d) Selección mediante la aplicación de cuotas.

[8]Por ejemplo, se selecciona uno de cada tres portales en edificios con más de una vivienda y dentro de la vivienda se realiza una entrevista por cada 12 hogares, etc. Ésta es la opción actualmente utilizada.

Los métodos más empleados son c) y d) y, en particular para las encuestas de opinión, d). A pesar de que la selección de los individuos mediante cuotas no es aleatoria[9], este sistema es muy utilizado dado que si la encuesta refleja la distribución de ciertas estructuras de la población es de esperar que recoja adecuadamente la información de las variables con las que están relacionadas. Las estructura de la población empleada en la encuesta mensual del CIS son la distribución por sexo y edad (18-24; 25-34; 35-44; 45-54; 55-64 y más de 65 años).

6.10. Estimadores

Los estimadores empleados son los correspondientes a un muestreo por conglomerados; en el caso de que la afijación sea proporcional.

$$\hat{P} = \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} m_i} \tag{3}$$

donde el numerador hace referencia al número de individuos que pertenecen a una determinada clase en la sección i y el denominador al número total de entrevistas realizadas en cada una de las secciones (m_i).

De este modo,

$$a_i = \sum_{j=1}^{m_i} x_j$$

donde x_j toma valores 1 ó 0, según el individuo j pertenezca o no a la clase de interés.

En el caso de que la afijación no sea proporcional, se aplican los estimadores correspondientes al muestreo estratificado:

$$\hat{P} = \sum_{h=1} W_h \cdot \hat{P}_h$$

donde

$W_h = \frac{N_h}{N}$, es el peso del estrato h en la población y

$\hat{P}_h$ es el estimador de la proporción (3) en el estrato h.

[9]Las propiedades de los estimadores sólo quedan garantizadas a través de esquemas probabilísticos.

6.11. Comentarios

El diseño de una encuesta desarrollado en los apartados 1 al 10 de este capítulo es el aplicado para las entrevistas personales. Sin embargo, este mismo diseño puede aplicarse considerando únicamente los apartados 6.1 a 6.5.

El investigador puede establecer el diseño más conveniente, por ejemplo, distribuyendo la muestra en los estratos y procediendo a una selección aleatoria de municipios. A partir de este punto, la selección de individuos dependerá del método de recogida que se elija:

- Entrevista personal: si el investigador no dispone de una lista de la población objetivo, puede definir un conjunto de etapas intermedias, como se ha desarrollado en los apartados 6.6 a 6.9.
- Encuesta telefónica: a partir del listado telefónico del municipio se procede a la selección aleatoria de individuos.
- Encuesta por correo: a partir de un listado de direcciones se procede a una selección aleatoria de direcciones.

7. NOTAS SOBRE EL COSTE

El coste de realizar una encuesta depende de muy diversos factores, entre los cuales destacan los elementos siguientes:

- *Elaboración del marco*

  En numerosas ocasiones, como ya se ha comentado, debe elaborarse el marco dado que no existe la información adecuada para proceder al desarrollo de las distintas fases establecidas por el diseño. Además del coste económico, su elaboración suele ser lenta y, por tanto, tener un alto coste en términos de tiempo que afecte a la oportunidad de la encuesta.

- *Tamaño de la muestra*

  Es el coste principal y está relacionado con la forma de recogida de la información. En general, para un mismo cuestionario, el coste es mayor para las entrevistas personales que para las telefónicas, y para estas últimas mayor que para una por correo.

- *Dispersión geográfica de la muestra*

  Cuanto más dispersa sea la muestra en el espacio mayores serán los costes, afectando de manera especial a las entrevistas personales, tanto por los mayores gastos de desplazamiento como por aumentar las labores de coordinación de los equipos de entrevistadores.

- *Tiempo de realización*

  Como ya se ha comentado, en las encuestas de opinión este elemento puede ser clave. En este sentido, las encuestas telefónicas, para el mismo número de entrevistas y cuestionario, son más ágiles que las personales, siendo las encuestas por correo más lentas al tener que realizar uno o varios recordatorios para alcanzar un nivel mínimamente aceptable de entrevistas. Por tanto, dado el corto espacio de tiempo que existe para la recogida de la información, la aplicación de entrevistas personales obligará a disponer de un número elevado de entrevistadores, lo cual afectará significativamente al coste de la encuesta.

  En el caso de las encuestas telefónicas, así como de las personales, si éstas se desarrollan bajo los sistemas CATI y CAPI[10], se obtiene una mejora en la recogida que se combina con la rápida obtención de resultados o avances, además de facilitar los procesos de control.

8. CALIDAD DE LA INFORMACIÓN OBTENIDA

El principal problema para las encuestas de opinión se centra en la falta de respuesta dado que, en general, suele ser muy elevada y, por tanto, producir sesgos en las estimaciones que afectan a la calidad de la información obtenida.

En la modalidad de encuestas por correo la tasa de respuesta suele ser inferior a las telefónicas y éstas, a su vez, son inferiores a las personales. Por otro lado, debe distinguirse entre la falta de respuesta total por parte de los individuos, cuando se niega a colaborar, y falta de respuesta parcial en una o varias preguntas. En el caso de falta de respuesta en los individuos, la estrategia más común en las encuestas de opinión es la substitución de aquellas unidades que forman parte del grupo de no respuesta por otras que sí colaboran, de tal forma que el número de cuestionarios recogidos coincida con el número de entrevistas fijadas en el diseño muestral.

En la situación de substitución lo que se produce es que la información se recoge del grupo «colaborador» de la población, quedando el grupo «no colaborador» sin representación y, por consiguiente, la presencia de sesgos en las estimaciones puede ser importante con este tipo de práctica.

Para aquellas situaciones donde existe una falta de respuesta en la pregunta y se realice un control sobre flujo de ciertas preguntas del cuestionario, la imputación del dato que falta suele ser un método aceptable, siempre y cuando su número no sea elevado.

[10]Corresponden a las siglas inglesas de Computer Assisted Personal Interviewing.

BIBLIOGRAFÍA

Arnáiz, G. (1978). *Introducción a la estadística teórica.* Lex Nova.

Azorín, F. y Sánchez-Crespo, J.L. (1986). *Métodos y Aplicaciones del Muestreo.* AUT.

Bosch, J.L. y Torrente, D. (1993). «Encuestas telefónicas y por correo», *Colección Cuadernos Metodológicos*, 9, CIS.

Cochran, W. (1977). *Sampling Techniques.* Third Edition Wiley.

García España, E. (1974). *Diseño de la Encuesta General de Población*, I.N.E.

Grosbras, J.M. (1987). *Methodes statistiques des sondages*, Económica.

Groves, R. (1998). *Nonresponse in household interview survey*, Wiley.

Kish, L. (1965). *Survey Sampling*, Wiley.

Hansen, M., Hurwitz, W. and Madow, W. (1953). *Sample Survey Methods and Theory*, vol. I y II, Wiley.

Lessler, J. and Kalsbeek, W. (1992). *Nonsampling error in surveys*, Wiley.

Mirás, J. (1985). *Elementos de muestreo en poblaciones finitas*, I.N.E.

Rodríguez Osuna, J. (1991). «Métodos de muestreo», *Colección Cuadernos Metodológicos*, 1, CIS.

Rodríguez Osuna, J. (1993). «Métodos de muestreo. Casos Prácticos», *Colección Cuadernos Metodológicos*, 6, CIS.

Sánchez-Crespo, J.L. (1984). *Curso intensivo de muestreo en poblaciones finitas.* Tercera Edición, I.N.E.

Stephan, F.F. and McCarthy P.J. (1958). *Sampling Opinions*, Wiley.

Anexo. Tamaños de muestra según nivel de error y valores de *P*, para un 95% de intervalo de confianza

	P(%)										
Error (%)	**1**	**5**	**10**	**15**	**20**	**25**	**30**	**35**	**40**	**45**	**50**
0,1	39.600	190.000	360.000	510.000	640.000	750.000	840.000	910.000	960.000	990.000	1.000.000
0,2	9.900	47.500	90.000	127.500	160.000	187.500	210.000	227.500	240.000	247.500	250.000
0,3	4.400	21.111	40.000	56.667	71.111	83.333	93.333	101.111	106.667	110.000	111.111
0,4	2.475	11.875	22.500	31.875	40.000	46.875	52.500	56.875	60.000	61.875	62.500
0,5	1.584	7.600	14.400	20.400	25.600	30.000	33.600	36.400	38.400	39.600	40.000
0,6	1.100	5.278	10.000	14.167	17.778	20.833	23.333	25.278	26.667	27.500	27.778
0,7	808	3.878	7.347	10.408	13.061	15.306	17.143	18.571	19.592	20.204	20.408
0,8	619	2.969	5.625	7.969	10.000	11.719	13.125	14.219	15.000	15.469	15.625
0,9	489	2.346	4.444	6.296	7.901	9.259	10.370	11.235	11.852	12.222	12.346
1,0	396	1.900	3.600	5.100	6.400	7.500	8.400	9.100	9.600	9.900	10.000
1,3	253	1.216	2.304	3.264	4.096	4.800	5.376	5.824	6.144	6.336	6.400
1,5	176	844	1.600	2.267	2.844	3.333	3.733	4.044	4.267	4.400	4.444
1,8	129	620	1.176	1.665	2.090	2.449	2.743	2.971	3.135	3.233	3.265
2,0	99	475	900	1.275	1.600	1.875	2.100	2.275	2.400	2.475	2.500
2,5	63	304	576	816	1.024	1.200	1.344	1.456	1.536	1.584	1.600
3,0	44	211	400	567	711	833	933	1.011	1.067	1.100	1.111
3,5	32	155	294	416	522	612	686	743	784	808	816
4,0	25	119	225	319	400	469	525	569	600	619	625
4,5	20	94	178	252	316	370	415	449	474	489	494
5,0	16	76	144	204	256	300	336	364	384	396	400
6,0	11	53	100	142	178	208	233	253	267	275	278
7,0	8	39	73	104	131	153	171	186	196	202	204
8,0	6	30	56	80	100	117	131	142	150	155	156
9,0	5	23	44	63	79	93	104	112	119	122	123
10,0	4	19	36	51	64	75	84	91	96	99	100
15,0	2	8	16	23	28	33	37	40	43	44	44
20,0	1	5	9	13	16	19	21	23	24	25	25
25,0	1	3	6	8	10	12	13	15	15	16	16

ENGLISH SUMMARY

SURVEY DESIGN FOR OPINIONS: BAROMETER CIS

V. MARTÍNEZ
Centro de Investigaciones Sociológicas*

The survey design for opinions would be included within of the general design of population surveys, where the objective is to acquire knowledge by observing the members of the population. The basic different characteristics in a general design are the small survey size and the use of non-probabilistic schemes for individual selections. In this way, the quality of the obtained results is questioned and the extrapolations over the population must be considered cautiously. These concepts are revised and valued in the framework of monthly opinion survey elaborated by the Centro de Investigaciones Sociológicas-CIS (commonly known as «Barómetro CIS»), which initial version was presented at First International Workshop «Generation of Statistical Information: Quality and Limitations» (Barcelona 30th November-1st December 1998), held by the Xarxa Temàtica «Surveys and Statistical Information Quality».

Keywords: Frames, questionnaire, design, sample size, survey units, clusters, stratification, allocation, cluster size, quota sampling, estimators, costs, quality.

AMS Classification: 62D05, 62P25.

*Departamento de Banco de Datos. Centro de Investigaciones Sociológicas (CIS). Montalbán, 8. 28014 Madrid. E-mail: vmartinez@cis.es.

-Received December 1998.
-Accepted April 1999.

The objective of this article is to draw the guidelines that are used to elaborate the survey design of opinion samples and to describe a survey called «Barómetro» which, basically, generates monthly data about the evolution of the some sociological and political variables in Spain. «Centro de Investigaciones Sociológicas −CIS−» elaborates this survey. CIS is a public organisation dependent on the Spanish Government.

The universe of the survey is the Spanish population aged 18 and more. The frame is based on the population data elaborated by the National Statistical Office of Spain (INE). This frame forces to use a cluster sampling because there is not a public register of individuals or households.

The questionnaire is elaborated to avoid a burden respondent, that should be considered excessive, and the interviews are face to face. The objective sample size is fix, 2,500 interviews, that correspond to an error of 2% in estimating proportions with simple random sampling and the hypothesis that $P = Q$. The design is a stratified multistage cluster sampling where there are different units, sizes and probabilities. The primary units are the municipalities that are stratified by three variables: geographical, number of inhabitants and neighbourhood to a large demographic nucleus. The total number of strata is 238 (where some of these strata are empty).

After the stratification, the design uses a proportional allocation, and different cluster sizes for the selection of primary units. The selection uses a proportional probability based on the number of inhabitants aged 18 and more. The second stage units are blocks, these units are considered homogeneous and have the same number of interviews, 10. The selection uses, as in municipalities, a proportional probability based on the number of inhabitants' aged 18 and more. The third stage uses random routes for selection of households; and the selection of final units-individuals- by quota methods based on sex-age population structures. This last step makes that the global design was a non-probabilistic design.

The estimator corresponds to simple random sampling, except when the design uses a disproportionate allocation, and in those cases the estimator used correspond to a stratified random sampling. The quota method used jointly with an intensive substitution of individuals in cases of non-contact (not at home) and refusals, make that the results must be considered cautiously.

Biometria

QÜESTIIÓ, vol. 23, 2, p. 365-392, 1999

ESTUDIOS DE SUPERVIVENCIA CON DATOS NO OBSERVADOS. DIFICULTADES INHERENTES AL ENFOQUE PARAMÉTRICO

G. GÓMEZ*
C. SERRAT**
Universitat Politècnica de Catalunya

A partir de una muestra de datos de supervivencia que contiene valores no observados en las covariantes de interés, presentamos una metodología que permite extraer toda la información contenida en covariantes completamente observadas, que estén fuertemente correlacionadas con las citadas covariantes de interés. El enfoque utilizado es completamente paramétrico y se basa en el método de máxima verosimilitud. Mostramos las dificultades, tanto de índole práctica como filosófica, que aparecen en la especificación de la función de verosimilitud y en su optimización. Diseñamos una metodología que permite determinar en qué medida los estimadores resultantes dependen de la modelización del patrón de no respuesta y la implementamos sobre S-PLUS. Dicha metodología, a su vez, permite el estudio de la tipología de datos no observados y proporciona un análisis de sensibilidad de los resultados obtenidos. Ilustramos la metodología presentada y sus dificultades con una aplicación a una cohorte de pacientes con tuberculosis pulmonar infectados por el virus de la inmunodeficiencia humana.

Survival studies with non observed data. Difficulties concerning the parametric approach.

Palabras clave: Análisis de supervivencia, estudio de validación, máxima verosimilitud, modelos de datos incompletos, patrón de no respuesta aleatorio, patrón de no respuesta no ignorable.

Clasificación AMS: 62F10, 62H99, 92C60.

Este trabajo ha sido parcialmente financiado por el proyecto DGICYT PB95–0776.

* Departament d'Estadística i Investigació Operativa. Universitat Politècnica de Catalunya. Pau Gargallo, 5. 08028 Barcelona. E-mail: ggg@eio.upc.es.

** Departament de Matemàtica Aplicada I. Universitat Politècnica de Catalunya. Av. Dr. Gregorio Marañón, 44–50. 08028 Barcelona. E-mail: carles@mal.upc.es.

– Recibido en febrero de 1998.
– Aceptado en febrero de 1999.

1. INTRODUCCIÓN

El llamado *missing data problem*, y al que nos podemos referir como el problema de los datos no observados, es un problema clásico pero que continúa sin ser resuelto de forma satisfactoria. Dicha problemática contiene, en particular, aquellas situaciones en que la variable dependiente sí es observada totalmente, pero la información sobre las covariantes o variables independientes es incompleta. Dentro de ésta, y si uno se encuentra en un contexto de análisis de supervivencia, la variable dependiente, o tiempo de supervivencia, es parcialmente observada debido al llamado problema de censura.

La mayoría de las metodologías estadísticas existentes basan sus conclusiones en la suposición, no siempre explícita, de que los datos no observados son completamente aleatorios o, en el mejor de los casos, aleatorios (dichas definiciones se definirán de forma precisa en la sección 2 de este artículo). Un breve repaso a los diferentes enfoques, dentro del contexto del análisis de la supervivencia, así como a sus ventajas e inconvenientes, nos lleva a considerar al menos las siguientes cuatro posibilidades.

Un primer enfoque consiste en basar todos los análisis solamente en aquellos individuos que tienen todas las covariantes observadas. En este caso la inferencia conllevará resultados sesgados e inconsistentes, puesto que los individuos observados no tienen porqué ser representativos de toda la muestra parcialmente observada. Por otro lado, los estimadores basados en dichos análisis serán menos precisos debido a la reducción del tamaño de la muestra. Este tipo de análisis, desafortunadamente ampliamente usado, es obviamente práctico y sencillo de ser utilizado puesto que puede llevarse a cabo mediante el software existente. Además, aún en el caso en que la no observación de los datos se hubiera producido de forma completamente aleatoria, los estimadores deducidos por este método serían ineficientes.

Otra metodología posible consiste en la imputación de los valores no observados (Glynn, Laird y Rubin (1993), Efron (1994), Serrat y Gómez (1995)). El principal problema con dicho enfoque reside en la suposición de que los datos no observados siguen el mismo patrón que los observados. Obviamente, como dicha hipótesis no tiene porqué ser cierta, su uso conlleva un gran riesgo en la inferencia que se realiza e implica un sesgo evidente en los estimadores. Otro de los enfoques clásicos consiste en la modelización paramétrica del problema y en su resolución via la maximización de la función de verosimilitud. Este tipo de soluciones nace con los trabajos de Little y Rubin en la década de los 70; éstos proponen el método de la máxima verosimilitud con el objetivo de extraer la máxima información contenida en los datos observados (Little y Rubin (1987), Glynn, Laird y Rubin (1986)). La inferencia basada en este tipo de análisis goza de buenas propiedades asintóticas y es ampliamente usado; sin embargo, su implementación no es en absoluto sencilla y los estimadores resultantes dependen fuertemente del gran número de hipótesis que tienen que asumirse, lo que condiciona la credibilidad de los mismos. Últimamente, algunos autores, entre ellos Baker, proponen una metodo-

logía jerárquica basada también en el método de la máxima verosimilitud, que permite combinar distintos modelos de patrón de no respuesta y estudiar las variaciones de los estimadores bajo estos supuestos (Baker (1994)).

El enfoque semiparamétrico es una cuarta vía que permite modelar sólo aquello que es estrictamente necesario; en concreto, la relación entre la supervivencia y las covariantes, y la relación entre la supervivencia y el patrón de no respuesta. Los estimadores semiparamétricos son insesgados, consistentes y asintóticamente normales (Newey (1990), Rotnitzky y Wypij (1994), Robins, Rotnitzky y Zhao (1994)).

Este trabajo parte, por un lado, de las ideas desarrolladas por la Dra. Andrea Rotnitzky, de la Universidad de Harvard, en el curso de especialización en *Datos No Observados (Missing Data)* impartido en la Universitat Politècnica de Catalunya en julio de 1996 y, por otro, de las numerosas discusiones con ella mantenidas posteriormente. En el artículo, como punto previo a un análisis semiparamétrico, los autores abordan el problema de los datos no observados en un contexto de análisis de la supervivencia y desde una perspectiva totalmente paramétrica con un doble objetivo: a) mostrar las dificultades tanto de índole práctica como filosófica en la especificación de la función de verosimilitud y en la optimización de la misma y b) diseñar una metodología que permita determinar en qué medida los estimadores resultantes dependen del patrón de no respuesta.

La motivación de la metodología que pretendemos abordar surge de los estudios clínicos y epidemiológicos. En estos estudios disponemos de una cohorte de pacientes y queremos estudiar por un lado su supervivencia y por otro determinar aquellas variables que puedan ser predictoras de la misma. En particular, la colaboración, desde 1994, con los epidemiólogos del Institut Municipal de la Salut de Barcelona, ha motivado la necesidad de encontrar una metodología que permita «tratar» los valores no observados en las covariantes de interés.

El desarrollo del trabajo es como sigue. En la siguiente sección introducimos la notación así como las definiciones relevantes. En la sección 3 planteamos el problema y lo resolvemos paramétricamente. En la sección 4 presentamos como ilustración el análisis de los datos que dieron lugar a dichas reflexiones. El artículo finaliza con una discusión.

2. NOTACIÓN Y DEFINICIONES

Llamamos **vector de datos potenciales** $L = (L_1, L_2, \cdots, L_K)$ de un individuo arbitrario al vector de dimensión K formado por los datos observados y los datos no observados de dicho individuo. Asimismo, definimos el **vector de respuesta** de dicho individuo, y lo representamos por $R = (R_1, R_2, \cdots, R_K)$, como el vector cuya com-

ponente k-ésima ($k = 1, \cdots, K$) es igual a 1 si la variable k-ésima ha sido observada y es 0 en caso contrario.

Subdividimos el vector de datos potenciales L en dos subvectores $L_{(R)}$ y $L_{(\bar{R})}$ correspondientes a los datos observados y a los no observados, respectivamente. El subvector $L_{(R)}$ está formado por aquellas componentes l del vector L para las cuales $R_l = 1$, mientras que el subvector de las variables no observadas, $L_{(\bar{R})}$, consiste en aquellas componentes l del vector L para las cuales $R_l = 0$. Si por ejemplo $K = 5$ y $R = (1, 0, 1, 1, 0)$ entonces $L_{(R)} = (L_1, L_3, L_4)$ mientras que $L_{(\bar{R})} = (L_2, L_5)$.

Denotemos por $r = (r_1, r_2, \cdots, r_K)$, $r_k \in \{0, 1\}$, $k = 1, \cdots, K$, una realización del vector de respuesta de un individuo arbitrario. La probabilidad condicional de r dado el vector de datos potenciales L, la denotamos por $\pi_L(r) = P(R = r|L)$, y los diferentes tipos de procesos de no respuesta dependerán de los valores que ésta tome.

El proceso de no respuesta se denomina **completamente aleatorio** y se abrevia por MCAR (Missing Completely at Random) si, y sólo si, la probabilidad de una realización r es constante, es decir, si $\pi_L(r)$ no depende de L. El proceso es **aleatorio** o Missing at Random (MAR) si, y sólo si, la probabilidad de una realización r depende únicamente de las variables que han sido observadas, es decir, $\pi_L(r)$ depende sólo de $L_{(r)}$. Por último, el proceso de no respuesta es **no ignorable** si, y sólo si, la probabilidad condicional $\pi_L(r)$ depende del subvector de datos no observados $L_{(\bar{r})}$.

Si suponemos que las componentes del vector L siguen un orden establecido, el proceso de no respuesta se denomina **monótono** cuando la **no** observación de una variable implica la **no** observación de las siguientes; es decir, cuando $R_l = 0$ implica $R_m = 0$ para todo $m > l$.

En un estudio de supervivencia la variable de interés, que denotaremos por T, es usualmente el tiempo transcurrido desde un origen (*i.e.* aleatorización en un ensayo clínico, inicio de un tratamiento, etc.) hasta la realización de un cierto suceso (muerte, diagnóstico de SIDA, recaída de una enfermedad, etc.). Frecuentemente, en dichos estudios la realización de dicho suceso no es siempre observada, debido, bien a la finalización del estudio, bien a la pérdida de algunos de los individuos, dando lugar a la denominada censura por la derecha. Si denotamos por C el tiempo de censura, en realidad sólo observamos la variable $Y = min\{T, C\}$ y un indicador de censura $\delta = \mathbf{1}\{T \leq C\} = \mathbf{1}\{Y = T\}$ que indica si el dato ha sido o no censurado.

Para cada individuo, además de recoger los valores de (Y, δ), se recogen los valores de otras variables, llamadas covariantes, tales como indicadores sociodemográficos, variables de tipo clínico, resultados de análisis, etc. El objetivo es, en general, la modelización del tiempo de supervivencia en función de las covariantes de interés. Denotemos por X el vector formado por las covariantes de interés. Si X_* es una componente de X y V_* es otra covariante (no necesariamente componente de X), diremos que V_* es **subro-**

gante de X_* si X_* y V_* están fuertemente correlacionadas. Por último, designemos por V el vector de covariantes subrogantes (no contenidas en X) para las componentes de interés.

Como es habitual en los análisis de supervivencia, supondremos que la distribución del tiempo de censura, C, es independiente de T, dado el vector de covariantes $(V^t, X^t)^t$.

Supongamos que disponemos de una muestra de individuos de tamaño n; de acuerdo con la notación introducida, para cada individuo i $(i = 1, \ldots, n)$ representamos por $L_i = (Y_i, \delta_i, V_i^t, X_i^t)^t$ el vector de datos potenciales, por R_i el vector de respuesta y por $L_{(R_i)}$ y $L_{(\bar{R}_i)}$ los subvectores correspondientes a los datos observados y a los no observados, respectivamente. En nuestro estudio tenemos, por construcción del vector L_i, que $L_{i1} = Y_i$ y $L_{i2} = \delta_i$ y por consiguiente $R_{i1} = R_{i2} = 1$ para todo $i = 1, \ldots, n$.

3. PRUEBAS DE HIPÓTESIS PARA EL ESTUDIO DEL PROCESO DE NO RESPUESTA

En esta sección nos proponemos el estudio y validación del proceso de no respuesta subyacente en los datos. En primer lugar, planteamos pruebas sobre la hipótesis de que el proceso de no respuesta sea completamente aleatorio. En segundo lugar, y bajo una perspectiva totalmente paramétrica, introducimos un esquema jerárquico de validación a partir del cual realizamos un análisis de sensibilidad de la adecuación del modelo bajo los distintos patrones de no respuesta. Esta metodología permite, en particular, elucidar sobre la no ignorabilidad del proceso de no respuesta.

3.1. Validación del modelo MCAR

La validación de un patrón de no respuesta completamente aleatorio, dada la muestra observada, $(R_i, L_{(R_i)})_{i=1,2,\cdots,n}$, se basa en la comparación de las probabilidades $P(R_{ik} = 1), i = 1, \cdots, n$ y $P(R_{ik} = 1|L_i), i = 1, \cdots, n$ para cada $k = 1, \cdots, K$. Para la resolución de la prueba de hipótesis

$$H_0 : \text{Proceso de no respuesta completamente aleatorio}$$
$$H_A : \text{Proceso de no respuesta aleatorio o no ignorable}$$

desarrollamos dos procedimientos en función de que el patrón de no respuesta sea monótono o que no lo sea.

Si el patrón de no respuesta es monótono la comparación anterior queda reducida a la comparación de las probabilidades $P(R_{ik} = 1|R_{i(k-1)} = 1), i = 1, \cdots, n$ y $P(R_{ik} = 1|R_{i(k-1)} = 1, L_{i1}, \cdots, L_{i(k-1)}), i = 1, \cdots, n$ para cada $k = 1, \cdots, K$

(Apéndice, apartado a)). Si expresamos el *logit* de $P(R_{ik} = 1|R_{i(k-1)} = 1, L_{i1}, \cdots, L_{i(k-1)})$ como $\alpha_{k1} + \alpha_{k2}^t h_k(L_{i1}, \cdots, L_{i(k-1)})$ donde $h_k(L_{i1}, \cdots, L_{i(k-1)})$ es una función vectorial arbitraria de los datos $L_{i1}, \cdots, L_{i(k-1)}$, la prueba de hipótesis H_0 contra H_A puede plantearse como las siguientes K pruebas simultáneas:

$$
\begin{aligned}
H_{0k} &: logit(P(R_{ik} = 1|R_{i(k-1)} = 1, L_{i1}, \cdots, L_{i(k-1)})) = \alpha_{k1} \\
H_{Ak} &: logit(P(R_{ik} = 1|R_{i(k-1)} = 1, L_{i1}, \cdots, L_{i(k-1)})) = \\
&\quad = \alpha_{k1} + \alpha_{k2}^t h_k(L_{i1}, \cdots, L_{i(k-1)}) \qquad\qquad k = 1, \cdots, K.
\end{aligned}
$$

Para cada k, $k = 1, \cdots, K$, el estadístico basado en la razón de verosimilitud de H_{0k} contra H_{Ak} nos proporciona un p-valor, p_k. En el apartado b) del Apéndice demostramos que si los datos son MCAR, *i.e.* las hipótesis $H_{0k}, k = 1, \cdots, K$, son todas verdaderas, entonces los p-valores resultantes siguen una distribución uniforme en (0,1) y son independientes entre sí.

Para la interpretación conjunta de los K p-valores resultantes, $p_1, \cdots, p_K$, utilizamos el estadístico combinado $S = -2\sum_{k=1}^{K} \log p_k$, que sigue bajo H_0 una distribución χ^2 con $2K$ grados de libertad.

Si el patrón de no respuesta es no monótono la comparación de H_0 versus H_A debe resolverse a partir de la comparación, para cada $i = 1, \cdots, n$ y para cada valor de k, $k = 1, \cdots, K$, de las probabilidades $P(R_{ik} = 1)$ y $P(R_{ik} = 1|$ datos observados para $k' \neq k)$. En este caso los p-valores obtenidos no son independientes y en el diseño de las correspondientes pruebas de hipótesis se deben utilizar técnicas de inferencia simultánea (Miller, 1980), como por ejemplo el estadístico t de Bonferroni. En general, estas técnicas son más conservadoras y en consecuencia reducirán la potencia de las pruebas de hipótesis resultantes. En la sección 4.2 presentamos detalladamente la aplicación de esta metodología.

3.2. Planteamiento paramétrico del problema

Planteamos el problema de la estimación del modelo de supervivencia de T desde una perspectiva completamente paramétrica y a partir de la muestra de datos potenciales $L_i = (Y_i, \delta_i, V_i^t, X_i^t)^t$, $i = 1, 2, \cdots, n$, donde X es un vector de covariantes parcialmente observadas y V es un vector de covariantes completamente observadas. Dicha perspectiva nos va a permitir introducir la modelización de las probabilidades de no respuesta y, en consecuencia, estudiar la bondad del ajuste para dichas modelizaciones. Dicha estimación se lleva a cabo mediante un análisis de máxima verosimilitud. Con el objetivo de especificar la función de verosimilitud, L, para dicha muestra, denotamos por $f_c(l;\theta)$ la función de densidad de los datos completos, y por $P(R_i = r|L_i;\psi)$ las

probabilidades de observar exactamente ciertas componentes de L_i. Nótese que dichas funciones dependen, respectivamente, de sendos parámetros θ y ψ, que se suponen de variación independiente.

La contribución del individuo i-ésimo a la función de verosimilitud $L(\theta, \psi)$ es: $f_c(L_i; \theta) \cdot P(R_i = \mathbf{1}|L_i; \psi)$ si el individuo ha sido completamente observado (*i.e.*, $R_i = \mathbf{1}$), y $\int f_c(L_i; \theta) \cdot P(R_i = r|L_i; \psi) dL_{(\bar{r})i}$ si el individuo ha sido parcialmente observado y su vector de respuesta es $R_i = r$ con $r \neq \mathbf{1}$. Esta segunda expresión corresponde a la marginalización de la primera respecto a los datos no observados. Así, la función de verosimilitud $L(\theta, \psi)$ a partir de los datos observados es

$$L(\theta, \psi) = \prod_{i=1}^{n} \left\{ [f_c(L_i; \theta) \cdot P(R_i = \mathbf{1}|L_i; \psi)]^{I(R_i = \mathbf{1})} \right.$$
$$\left. \prod_{r \neq \mathbf{1}} \left[\int f_c(L_i; \theta) \cdot P(R_i = r|L_i; \psi) dL_{(\bar{r})i} \right]^{I(R_i = r)} \right\}.$$

En el apartado c) del Apéndice demostramos que, cuando el patrón de no respuesta es MCAR o MAR, es decir, cuando las probabilidades $P(R_i = r|L_i)$ no dependen de $L_{(\bar{r})i}$, dichas probabilidades se pueden factorizar en la expresión de la verosimilitud, $L(\theta, \psi)$, y, por consiguiente, la estimación máximo verosímil del parámetro θ es independiente del patrón de no respuesta utilizado.

En nuestro estudio la función de densidad de los datos, $f_c(L_i; \theta)$, toma la expresión

$$f_c(L_i; \theta) = f_c((Y_i, \delta_i, V_i^t, X_i^t)^t; \theta) =$$
$$= f_c(X_i; \theta) \cdot f_c(Y_i, \delta_i|X_i; \theta) \cdot f_c(V_i|Y_i, \delta_i, X_i; \theta).$$

La modelización paramétrica del problema exige la correcta especificación de: a) la distribución de las covariantes de interés, X; b) la distribución de los tiempos observados, Y, condicionada a las covariantes X; c) la distribución de las covariantes subrogantes, V, condicionadas a Y y a X, y d) las probabilidades de respuesta condicionadas a los datos potenciales. Hemos de tener en cuenta que dichas especificaciones pueden ser hasta cierta medida arbitrarias y que, además, ninguna de las citadas distribuciones es validable a partir de los datos observados.

Supongamos en lo que sigue que el vector de covariantes del i–ésimo individuo, X_i, está formado por p variables aleatorias discretas, $X_{i1}, X_{i2}, \cdots, X_{ip}$ y que el vector V_i de covariantes subrogantes contiene asimismo q variables aleatorias discretas, $V_{i1}, V_{i2}, \cdots, V_{iq}$.

La especificación de las funciones de densidad $f_c(X_i; \theta)$ y $f_c(V_i|Y_i, \delta_i, X_i; \theta)$, así como la determinación de las probabilidades $P(R_i = r|L_i; \psi)$, puede hacerse en términos de los logaritmos de las *odds ratio* condicionadas de cada categoría respecto al grupo de

referencia (regresiones logísticas condicionadas, en el caso binario), es decir, dando una modelización para cada una de las siguientes expresiones:

$$\log \frac{P(X_{ij}=k|X_{i1},\cdots,X_{i(j-1)})}{P(X_{ij}=0|X_{i1},\cdots,X_{i(j-1)})} \quad j=1,\cdots,p \;\; k\neq 0,$$

$$\log \frac{P(V_{ij}=k|V_{i1},\cdots,V_{i(j-1)})}{P(V_{ij}=0|V_{i1},\cdots,V_{i(j-1)})} \quad j=1,\cdots,q \;\; k\neq 0 \quad \text{y}$$

$$logit(P(R_{ij}=1|R_{i1},\cdots,R_{i(j-1)})) \;\; j=1,\cdots,p.$$

Como veremos en la ilustración de la siguiente sección, el uso de distintos modelos en la expresión de las probabilidades $P(R_{ij}=1|R_{i1},\cdots,R_{i(j-1)})$, $j=1,\cdots,p$, en función de los datos L_i permite analizar la sensibilidad de los estimadores al patrón de no respuesta utilizado.

La modelización de los tiempos de supervivencia y su relación con las covariantes X, queda restringida a la especificación de la función de densidad $f_c(Y_i,\delta_i|X_i;\theta)$. En este caso, pueden ser útiles análisis previos basados en la submuestra completamente observada. Sin embargo, una vez más, estas suposiciones no se pueden validar a partir de los datos observados.

Una limitación añadida a las anteriores modelizaciones es el crecimiento rápido de la dimensión de los parámetros θ y ψ. Este hecho puede repercutir, de manera importante, en la ejecución de las correspondientes implementaciones, a la vez que reduce el tamaño muestral relativo. Por ejemplo, es fácil calcular que en el simple supuesto de que todas las covariantes fueran binarias y usando sólo modelos lineales que no incluyesen interacciones, tendríamos

$$\begin{aligned} dim(\theta) &= (2^p-1)+(p+1)+(2^q-1)(p+3), \\ dim(\psi) &= (2^p-1)(p+q+3), \end{aligned}$$

y para una situación sencilla en la que $p=q=3$, $dim(\theta)=53$, $dim(\psi)=63$ y por consiguiente tendríamos un total de 116 parámetros en la función de verosimilitud.

4. ILUSTRACIÓN

Como ilustración de la metodología de la sección anterior presentamos una aplicación en un estudio de supervivencia en una cohorte de pacientes con tuberculosis pulmonar infectados por el virus de la inmunodeficiencia humana (VIH). Dicha aplicación forma parte de la colaboración que los autores mantienen con el Servicio de Epidemiología del Institut Municipal de la Salut de Barcelona. Los datos proceden de distintos registros de pacientes del Programa de Prevención y Control de la Tuberculosis. Entre los objetivos

epidemiológicos del citado Programa destacan: a) el estudio de la progresión del SIDA en pacientes tuberculosos y b) la determinación de indicadores de supervivencia en pacientes VIH+ y con tuberculosis.

4.1. Material y métodos

Se dispone de una muestra de 418 pacientes VIH+ con tuberculosis pulmonar residentes en Barcelona que fueron diagnosticados con tuberculosis en el período 1992–1994. El cierre del estudio se realizó en fecha 30 de septiembre de 1995. El tiempo de supervivencia de interés es el transcurrido entre el diagnóstico de la enfermedad (con el consiguiente immediato inicio del tratamiento antituberculosis) y la muerte. Para cada individuo de la muestra se dispone, potencialmente, de las siguientes variables de tipo sociológico y clínico, recogidas a su entrada en el estudio: sexo, edad, distrito de residencia, antecedentes de prisión, seguimiento anterior de un tratamiento antituberculosis, pertenencia a un grupo de prácticas de riesgo, localización de la tuberculosis, resultados radiológicos, resultados bacteriológicos, porcentajes de subpoblaciones linfocitarias T-CD4+ y T-CD8+, resultado de la prueba de la tuberculina, etc.

Estudios previos (Caylà *et al.*, 1993; Serrat y Gómez, 1995) realizados con datos completamente observados demuestran que para la estimación del mencionado tiempo de supervivencia las covariantes de interés son el porcentaje de T-CD4+ (y en particular su dicotomización en un nivel bajo y uno alto de inmunodepresión) y el resultado de la prueba de la tuberculina. Nos referiremos a estas variables por CD4 y PPD, respectivamente, y serán las componentes del vector X introducido en la sección 2. Distinguiremos los pacientes por el nivel de inmunodepresión según si la variable CD4 toma valores superiores al 14% o no. La variable PPD toma valor 1 si el resultado de la prueba de la tuberculina es positivo, y 0 en caso contrario. El problema metodológico viene motivado por el hecho de que en nuestra muestra dichas variables presentan un 37.5 % y 50.5 % de valores no observados, respectivamente. Más concretamente, sólo se dispone de ambas en un 31.3 % de los individuos de la muestra y hay un 19.1 % de la muestra que no tiene recogido el valor de ninguna de las dos variables.

El objetivo es determinar el carácter predictivo de los indicadores CD4 y PPD haciendo uso de toda la información contenida en la muestra y, en particular, estudiar el resultado de la prueba de la tuberculina como medida de calidad complementaria al nivel de respuesta inmunológica que proporciona el recuento de T-CD4+. Para todo ello utilizaremos la metodología descrita en la sección anterior.

Después de estudiar, conjuntamente con el equipo de epidemiólogos, qué variables podrían proporcionar información cualitativa sobre el patrón de no respuesta o sobre las variables CD4 y PPD, se han elegido: a) el haber seguido un tratamiento anterior (TR: 1=Sí y 2=No); b) la radiología (RA: 0 = Normal, 1 = Anormal con patrón cavi-

tario y 2 = Anormal sin patrón cavitario), y c) la bacteriología (BA: 0 = Negativa, 1 = Positiva y 2 = Positiva con cultivo bacteriológico) como covariantes subrogantes, V, de las covariantes de interés, X. Así pues, de acuerdo con la notación introducida, nuestro vector de datos es:

$$L_i = (Y_i, \delta_i, TR_i, RA_i, BA_i, \mathbf{1}\{CD4_i > 14\}, PPD_i)^t.$$

Para simplificar la notación omitiremos, siempre que no se preste a confusión, el subíndice i.

Observemos que, dado que las covariantes subrogantes son completamente observadas, $R_3 = R_4 = R_5 \equiv 1$. Así pues, el vector de observaciones $r \in \{0,1\}^7$, tomará únicamente los valores (1,1,1,1,1,1,1), (1,1,1,1,1,1,0), (1,1,1,1,1,0,1) y (1,1,1,1,1,0,0).

4.2. Validación del modelo MCAR

Para la validación del modelo MCAR aplicamos la metodología presentada en la sección 3.1. Como los datos no responden a un patrón de no respuesta monótono, hemos de comparar la probabilidad $P(R_{CD4} = 1)$ con $P(R_{CD4} = 1|Y, \delta, TR, RA, BA, PPD)$, y $P(R_{PPD} = 1)$ con $P(R_{PPD} = 1|Y, \delta, TR, RA, BA, \mathbf{1}\{CD4 > 14\})$.

Para el primer caso, si modelamos la probabilidad de respuesta a la variable $CD4$ como función de las otras variables observadas según el modelo logístico

$$\begin{aligned} &logit(P(R_{CD4} = 1|Y, \delta, TR, RA, BA, PPD)) = \\ &= \alpha_0 + \alpha_1 Y + \alpha_2 \delta + \alpha_3 TR + \alpha_4 RA + \alpha_5 BA + \alpha_6 PPD, \end{aligned}$$

la comparación entre las probabilidades $P(R_{CD4} = 1)$ y $P(R_{CD4} = 1|Y, \delta, TR, RA, BA, PPD)$ es equivalente a la siguiente prueba de hipótesis

$$\begin{aligned} H_0 &: \alpha_i = 0 \quad \forall i \in \{1, \cdots, 6\} \\ H_A &: \exists i \in \{1, \cdots, 6\} | \alpha_i \neq 0. \end{aligned}$$

Si es verdadera la hipótesis nula, la probabilidad de respuesta a la variable $CD4$ no dependerá de las otras variables observadas y su contribución en la prueba de hipótesis MCAR del proceso de no respuesta será en el sentido de no rechazar este supuesto. En caso contrario, de no ser la hipótesis nula cierta, tendremos evidencia para rechazar la hipótesis de no respuesta MCAR.

Análogamente, aplicamos también esta metodología para la probabilidad de respuesta a la variable PPD, con el modelo

$$\begin{aligned} &logit(P(R_{PPD} = 1|Y, \delta, TR, RA, BA, \mathbf{1}\{CD4 > 14\})) = \\ &= \beta_0 + \beta_1 Y + \beta_2 \delta + \beta_3 TR + \beta_4 RA + \beta_5 BA + \beta_6 \mathbf{1}\{CD4 > 14\}. \end{aligned}$$

Para combinar ambas pruebas utilizamos la corrección de Bonferroni. Esta corrección consiste en utilizar como nivel de significación en cada prueba de hipótesis parcial el nivel de significación global dividido por el número de pruebas que se desean combinar. El criterio de decisión que se utiliza es el siguiente: se rechaza la hipótesis nula cuando alguna prueba parcial rechaza su correspondiente hipótesis nula. Observemos que, por una parte, puede parecer más immediato rechazar la hipótesis nula por el simple hecho de utilizar más de una prueba; ahora bien, el hecho de utilizar una fracción del nivel de significación en cada prueba parcial exige más evidencia en contra de la respectiva hipótesis nula parcial. El efecto combinado es una prueba, en general, más conservadora.

En nuestro caso, las pruebas de hipótesis para las variables R_{CD4} y R_{PPD} tienen un p-valor 0.03766 y 0.1733, respectivamente. Si utilizamos un nivel de significación global del 5%, ambos resultados no resultan significativos (son mayores de 0.025) con lo que en ambas variables no se detecta evidencia en contra de la hipótesis nula. En consecuencia, no podemos rechazar la hipótesis que el proceso de no respuesta sea MCAR.

4.3. Planteamiento paramétrico del problema

Siguiendo los pasos indicados en la sección 3.2, la contribución de un individuo a la función de verosimilitud $L(\theta, \psi)$ se puede calcular a partir de las cuatro expresiones que a continuación se detallan.

a) Para las distribuciones de las covariantes de interés, $CD4$ y PPD, se han usado modelos logísticos para la probabilidad de $\mathbf{1}\{CD4 > 14\}$ y para las probabilidades de $PPD = 1$ condicionadas a los valores de $\mathbf{1}\{CD4 > 14\}$, es decir

$$\begin{aligned} logit(P(\mathbf{1}\{CD4 > 14\} = 1)) &= \alpha_1, \\ logit(P(PPD = 1|\mathbf{1}\{CD4 > 14\} = 1)) &= \alpha_2 \quad \text{y} \\ logit(P(PPD = 1|\mathbf{1}\{CD4 > 14\} = 0)) &= \alpha_3. \end{aligned}$$

b) En un estudio inicial realizado sobre la misma cohorte de pacientes (Serrat *et al.*, 1998) se observó que el tiempo de supervivencia podía ser modelado satisfactoriamente mediante una distribución de Weibull con covariantes CD4 y PPD. Esta misma modelización es la utilizada en esta aplicación y por tanto la función $f_c(Y, \delta|X; \theta)$ se expresa como

$$f_c(Y, \delta|CD4, PPD; \theta) = \left(\frac{1}{\sigma Y}(e^{-\beta}Y)^{\frac{1}{\sigma}}\right)^{\delta} \cdot e^{-(e^{-\beta}Y)^{\frac{1}{\sigma}}},$$

donde $\beta = \beta_0 + \beta_1 \mathbf{1}\{CD4 > 14\} + \beta_2 PPD$ y $\sigma = \sigma_0 + \sigma_1 \mathbf{1}\{CD4 > 14\} + \sigma_2 PPD$.

c) Las distribuciones de TR, RA y BA dados $Y, \delta, CD4, PPD$ se modelizan a partir de los logaritmos de las *odds ratio* respecto del grupo de referencia, condicionadas a las variables previas. Si por $\lambda_.$ denotamos de manera genérica una de las siguientes *odds ratio*

$$\lambda_1 = \frac{P(TR=1)}{P(TR=0)},$$

$$\lambda_{2ij} = \frac{P(RA=j|TR=i)}{P(RA=0|TR=i)} \quad i=0,1 \quad j=1,2,$$

$$\lambda_{3ijk} = \frac{P(BA=k|TR=i,RA=j)}{P(BA=0|TR=i,RA=j)} \quad i=0,1 \quad j=0,1,2 \quad k=1,2,$$

entonces su logaritmo, $\log(\lambda_.)$, queda especificado por

$$\log(\lambda_.) = \gamma_{.0} + \gamma_{.1}\log(Y) + \gamma_{.2}\delta + \gamma_{.3}\log(Y)\delta + \gamma_{.4}\mathbf{1}\{CD4 > 14\} + \gamma_{.5}PPD,$$

y está en función de las variables Y, δ y X, de algunas posibles interacciones entre ellas, y del parámetro $(\gamma_{.0}, \cdots, \gamma_{.5})$. Obsérvese que para reducir el efecto debido a los valores extremos en los tiempos de supervivencia observados, hemos reescalado los tiempos a escala logarítmica.

d) Los distintos patrones de no respuesta los diseñamos a partir de la modelización de las probabilidades $P(R_{ij} = 1|R_{i1}, \cdots, R_{i(j-1)})$ en términos de las covariantes V y X. Definimos modelos logísticos para la probabilidad de observación de la variable $CD4$ y para las probabilidades de observación de la variable PPD condicionadas a los valores de la variable $CD4$ de acuerdo con el siguiente esquema:

$$\begin{aligned}
logit(P(R_{CD4}=1)) &= \alpha_{10} + \\
&+ \alpha_{11}TR + \alpha_{12}RA1 + \alpha_{13}RA2 + \alpha_{14}BA1 + \alpha_{15}BA2 \\
&+ \alpha_{16}\mathbf{1}\{CD4 > 14\} + \alpha_{17}PPD, \\
logit(P(R_{PPD}=1|R_{CD4}=1)) &= \alpha_{20} + \\
&+ \alpha_{21}TR + \alpha_{22}RA1 + \alpha_{23}RA2 + \alpha_{24}BA1 + \alpha_{25}BA2 \\
&+ \alpha_{26}\mathbf{1}\{CD4 > 14\} + \alpha_{27}PPD \quad \text{y} \\
logit(P(R_{PPD}=1|R_{CD4}=0)) &= \alpha_{30} + \\
&+ \alpha_{31}TR + \alpha_{32}RA1 + \alpha_{33}RA2 + \alpha_{34}BA1 + \alpha_{35}BA2 \\
&+ \alpha_{36}\mathbf{1}\{CD4 > 14\} + \alpha_{37}PPD,
\end{aligned}$$

donde $RAi = \mathbf{1}\{RA = i\}, i = 1,2$ y $BAi = \mathbf{1}\{BA = i\}, i = 1,2$ denotan las variables binarias que recogen los efectos de las categorías en radiología y bacteriología, respectivamente.

El parámetro θ de la densidad $f_c(L_i;\theta)$ en la verosimilitud $L(\theta,\psi)$ es

$$\theta = (\alpha_1, \alpha_2, \alpha_3, \beta_0, \beta_1, \beta_2, \sigma_0, \sigma_1, \sigma_2, \gamma_{.0}, \cdots, \gamma_{.5})$$

y tiene dimensión 111. El parámetro complementario resultante de la modelización de las probabilidades de respuesta es

$$\psi = (\alpha_{10}, \cdots, \alpha_{37})$$

y tiene dimensión 24. Observemos que el parámetro de interés, a efectos de estimación, está formado únicamente por las componentes $(\beta_0, \beta_1, \beta_2, \sigma_0, \sigma_1, \sigma_2)$ del vector θ y es de dimensión 6.

La jerarquización de las probabilidades del apartado d) permite simular, de forma anidada, los distintos patrones de no respuesta definidos en la sección 2. En nuestro estudio optimizamos la función de verosimilitud para los siguientes cinco casos:

$$\begin{aligned}
\alpha_{ij} = 0 \quad i = 1,2,3 \quad j = 1,\cdots,7 &\Rightarrow \text{MCAR} \\
\alpha_{ij} = 0 \quad i = 1,2,3 \quad j = 6,7 &\Rightarrow \text{MAR} \\
\alpha_{ij} = 0 \quad i = 1,2,3 \quad j = 7 &\Rightarrow \text{No ignorable (1r caso) NI1} \\
\alpha_{ij} = 0 \quad i = 1,2,3 \quad j = 6 &\Rightarrow \text{No ignorable (2o caso) NI2} \\
\text{Sin restricciones sobre los valores de } \alpha_{ij} &\Rightarrow \text{No ignorable (3r caso) NI3}
\end{aligned}$$

La tabla 1 muestra la estimación de los cuartiles relativos del grupo tuberculín positivo ($PPD = 1$) respecto del grupo tuberculín negativo ($PPD = 0$), bajo los distintos supuestos de covariantes subrogantes y, en cada uno de ellos, considerando los distintos patrones de no respuesta citados.

El uso de otros modelos basados en las mismas covariantes permitiría llevar a cabo un análisis de sensibilidad más exhaustivo de los estimadores resultantes y podría dar una respuesta global al problema en función del patrón de no respuesta subyacente en los datos.

La implementación de esta metodología se ha realizado en S-PLUS y ejecutado en un ordenador PC-Pentium Pro, 200 Mhz con 32 Mb RAM, en entorno Windows 95.

4.4. Interpretación de los resultados

En cuanto a la validación del modelo MCAR, hemos visto que, si bien existen técnicas eficientes para su validación bajo un patrón de no respuesta monótono, las técnicas existentes para el caso no monótono no gozan de potencia suficiente. En otros términos, si se utiliza dicha metodología se hace necesaria una mayor evidencia en los datos para

rechazar la hipótesis MCAR sobre el modelo de no respuesta. En nuestro caso sería necesario utilizar un nivel de confianza máximo del 92.4% para rechazar la hipótesis MCAR.

Tabla 1. Estimación de cuartiles relativos para el grupo tuberculín positivo respecto del grupo tuberculín negativo, bajo los distintos supuestos de covariantes subrogantes (V) y modelos de patrón de no respuesta (M_i).

Covariantes Subrogantes V	Patrón de no respuesta M_i	Primer cuartil	Mediana	Tercer cuartil	*Deviance* D_{M_i}	Número de parámetros n_i
Ninguna	MCAR=MAR	2.983	1.630	1.012	4102.055	12
	NI1	2.293	1.384	0.929	4092.145	15
	NI2	2.675	1.479	0.927	4092.993	15
	NI3	1.784	1.125	0.782	4085.622	18
TR	MCAR/MAR	2.674	1.489	0.939	4522.524/4513.648	18/21
	NI1	1.973	1.270	0.897	4504.61	24
	NI2	2.640	1.474	0.931	4504.236	24
	NI3	1.780	1.121	0.779	4496.693	27
RA	MCAR/MAR	2.963	1.592	0.976	4688.37/4683.306	24/30
	NI1	2.251	1.341	0.892	4672.024	33
	NI2	2.652	1.483	0.938	4674.658	33
	NI3	1.178	1.122	0.781	4664.661	36
BA	MCAR/MAR	2.912	1.644	1.048	4962.181/4957.538	24/30
	NI1	2.262	1.391	0.948	4947.412	33
	NI2	2.657	1.484	0.937	4947.574	33
	NI3	1.776	1.124	0.783	4939.764	36
TR, RA	MCAR/MAR	2.751	1.496	0.926	5102.458/5087.667	42/51
	NI1	2.013	1.265	0.877	5077.458	54
	NI2	2.623	1.479	0.942	5077.824	54
	NI3	2.039	1.284	0.892	5070.985	57
TR, BA	MCAR/MAR	2.102	1.264	0.847	5356.891/5343.023	42/51
	NI1	2.164	1.229	0.787	5334.777	54
	NI2	2.182	1.286	0.848	5340.131	54
	NI3	1.245	1.423	1.580	5326.711	57
RA, BA	MCAR/MAR	2.890	1.641	1.050	5513.294/5504.302	60/72
	NI1	2.316	1.426	0.974	5492.744	75
	NI2	2.740	1.524	0.960	5494.363	75
	NI3	2.273	1.415	0.973	5487.814	78
TR, RA, BA	MCAR/MAR	2.739	1.451	0.880	5888.679/5863.155	114/129
	NI1	2.691	1.442	0.881	5853.455	132
	NI2	2.773	1.456	0.877	5854.811	132
	NI3	1.898	1.177	0.807	5838.814	135

Tabla 2. Comparación de los ajustes obtenidos en la estimación paramétrica, bajo distintos supuestos de covariantes subrogantes (V) y modelos de patrón de no respuesta (M_i). $\Delta_{M_i M_j} = (-2\log L_{M_i}) - (-2\log L_{M_j}) = D_{M_i} - D_{M_j}$.
(*) Resultado significativo al 95%,
(**) Resultado significativo al 99%

Covariantes Subrogantes V	Patrón de no respuesta M_i	Patrón de no respuesta M_j	$\Delta_{M_i M_j}$	Grados de libertad $n_j - n_i$	$p-$value
Ninguna	MCAR=MAR	NI1	9.91	3	0.019(*)
	MCAR=MAR	NI2	9.062	3	0.029(*)
	NI1	NI3	6.523	3	0.089(*)
	NI2	NI3	7.371	3	0.061(*)
TR	MCAR	MAR	8.876	3	0.031(*)
	MAR	NI1	9.038	3	0.029(*)
	MAR	NI2	9.412	3	0.024(*)
	NI1	NI3	7.917	3	0.048(*)
	NI2	NI3	7.543	3	0.057
RA	MCAR	MAR	5.064	6	0.536
	MAR	NI1	11.282	3	0.010(*)
	MAR	NI2	8.648	3	0.034(*)
	NI1	NI3	7.363	3	0.061
	NI2	NI3	9.997	3	0.019(*)
BA	MCAR	MAR	4.643	6	0.59
	MAR	NI1	10.126	3	0.018(*)
	MAR	NI2	9.964	3	0.019(*)
	NI1	NI3	7.648	3	0.054
	NI2	NI3	7.81	3	0.050
TR, RA	MCAR	MAR	14.791	9	0.097
	MAR	NI1	10.209	3	0.017(*)
	MAR	NI2	9.843	3	0.02(*)
	NI1	NI3	6.473	3	0.091
	NI2	NI3	6.839	3	0.077
TR, BA	MCAR	MAR	13.868	9	0.127
	MAR	NI1	8.246	3	0.041(*)
	MAR	NI2	2.892	3	0.409
	NI1	NI3	8.066	3	0.045(*)
	NI2	NI3	13.42	3	0.004(**)
RA, BA	MCAR	MAR	8.992	12	0.704
	MAR	NI1	11.558	3	0.009(**)
	MAR	NI2	9.939	3	0.019(*)
	NI1	NI3	4.93	3	0.177
	NI2	NI3	6.549	3	0.088
TR, RA, BA	MCAR	MAR	25.524	15	0.043(*)
	MAR	NI1	9.7	3	0.021(*)
	MAR	NI2	8.344	3	0.039(*)
	NI1	NI3	14.641	3	0.002(**)
	NI2	NI3	15.997	3	0.001(**)

La sensibilidad de la estimación de los parámetros a la modelización del patrón de no respuesta, como queda reflejado en la tabla 1, demuestra que el patrón de no respuesta no es MCAR e ilustra la baja potencia de dicha metodología.

Respecto a la estimación paramétrica, para la comparación de los resultados obtenidos, para un cierto subconjunto de covariantes subrogantes, según los distintos modelos anidados del patrón de no respuesta (MCAR a NI3), utilizaremos el estadístico resultante de la diferencia entre las *deviances* $(-2 \log L)$ de las dos modelizaciones.

Si denotamos por D_{M_i} la *deviance* obtenida en la modelización M_i, $i = 1, \cdots, 5$, el estadístico $\Delta_{M_i M_j} = D_{M_i} - D_{M_j}$ sigue una distribución $\chi^2_{n_j - n_i}$ donde n_i y n_j $(n_i < n_j)$ son el número de parámetros a estimar bajo cada una de las dos modelizaciones anidadas M_i y M_j, respectivamente. La tabla 2 presenta los p-valores obtenidos en la comparación de los modelos paramétricos resultantes, bajo distintos supuestos de covariantes subrogantes (V) y modelos de patrón de no respuesta (M_i).

Del análisis de las tablas 1 y 2 podemos concluir:

1. En nuestra aplicación, en general, el modelo MAR no resulta significativo respecto al correspondiente MCAR. La interpretación inmediata es pues que la probabilidad de observación de las variables $CD4$ y PPD poco depende de los valores observados en las variables subrogantes.

   Concretamente, sólo obtenemos diferencias significativas cuando utilizamos como variable subrogante el haber seguido recientemente tratamiento antituberculosis. En este caso las estimaciones de los coeficientes correspondientes a la variable TR resultan de signo positivo; esto nos permite concluir, como parece lógico, que aquellos pacientes que tienen antecedentes de tratamiento antituberculosis tienen mayor probabilidad de tener informadas las variables $CD4$ y PPD.

2. En todos los supuestos, los modelos No Ignorables resultan significativos respecto a los MCAR y MAR. Esto nos indica que efectivamente la probabilidad de observación de las variables CD4 y PPD depende de los valores potenciales de dichas variables. Observemos que este resultado es imposible de obtener a partir de pruebas de hipótesis basadas en los datos observados (Sección 3.1)

3. La positividad en la prueba de la tuberculina es un mejor pronóstico de supervivencia *a corto plazo*; sin embargo, a *largo plazo* la supervivencia es peor. Para ilustrar este hecho presentamos en la figura 1 la estimación de la función de supervivencia para el grupo tuberculín positivo y para el grupo tuberculín negativo, según el nivel de inmunodepresión, cuando utilizamos como covariantes subrogantes las variables TR y RA, y como patrón de no respuesta el NI2. Las estimaciones correspondientes a los parámetros β y σ de la distribución de Weibull son $\beta = 6.878 + 1.362 \cdot \mathbf{1}\{CD4 > 14\} + 0.153 \cdot PPD$ y $\sigma = 1.426 + 0.244 \cdot \mathbf{1}\{CD4 > 14\} - 0.651$

$\cdot PPD$, respectivamente. Esto nos permite estimar la distribución de cuartiles para el grupo más inmunodeprimido en 164, 576 y 1547 días si el resultado de la prueba de la tuberculina es negativo o en 431, 851 y 1457 días si el resultado de dicha prueba es positivo. Análogamente, para el grupo menos inmunodeprimido, podemos estimar los cuartiles en 473, 2055 y 6539 días o 1241, 3040 y 6160 días, según el resultado de la prueba de la tuberculina. Efectivamente a corto y medio plazo (hasta aproximadamente el percentil del 72%) la supervivencia del grupo tuberculín positivo es mejor —independientemente del nivel de inmunodepresión— que la del grupo tuberculín negativo.

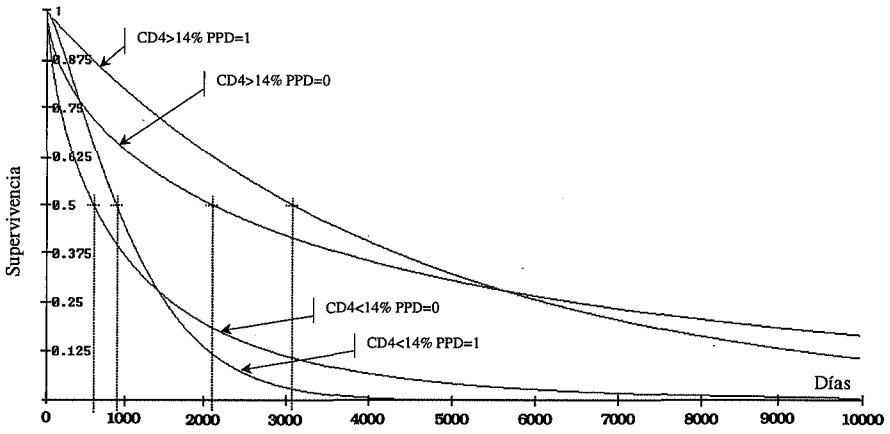


Figura 1. Estimación de la función de supervivencia para el grupo tuberculín positivo ($PPD = 1$) y para el grupo tuberculín negativo ($PPD = 0$), según el nivel de inmunodepresión (alto $\equiv$ $CD4 \leqslant 14\%$, bajo $\equiv$ $CD4 > 14\%$), cuando se utilizan las covariantes TR y RA como subrogantes y el modelo NI2 como patrón de no respuesta.

5. DISCUSIÓN

La metodología presentada adolece de ciertas limitaciones que hace falta conocer y que condicionan su aplicabilidad. Por un lado, es realmente difícil especificar correctamente tanto los modelos usados para las distintas funciones de densidad, como para las probabilidades de respuesta. Una consecuencia inmediata de dicha dificultad es la introducción de sesgo en los estimadores o, equivalentemente, la fuerte dependencia de los estimadores resultantes respecto de las hipótesis consideradas.

En segundo lugar, es imprescindible elegir adecuadamente las covariantes subrogantes. Para ello hará falta tener en cuenta de manera especial, además de criterios estadísticos, criterios inherentes al contenido de las variables (clínicos, epidemiológicos y del mismo proceso de recogida de datos) en la medida en que éstas permitan «aportar» información bien sobre el proceso de no respuesta, bien sobre las covariantes parcialmente observadas. Por estos motivos se hace necesario un análisis de sensibilidad complementario que permita dar interpretaciones razonables de las estimaciones bajo los diferentes supuestos de patrón de no respuesta.

Otra de las limitaciones con las que se encuentra este tipo de análisis se debe al crecimiento geométrico de la dimensión de los parámetros que intervienen en las distintas modelizaciones. Como única alternativa a dicha dificultad se propone una restricción en el número de covariantes de interés y de covariantes subrogantes. También la simplificación de las modelizaciones (*i.e.*, estableciendo relaciones funcionales entre las covariantes) puede reducir la dimensión de los parámetros que intervienen en la optimización. Una vez más, estas relaciones funcionales no podrán ser validadas a partir de los datos observados.

Por último, cabe destacar el elevado coste computacional que supone la ejecución de un análisis completo. El tiempo de estimación de un modelo oscila, según la complejidad de éste, de minutos en los más simples a horas —e incluso días— en los más complejos. Todo ello condiciona fuertemente su utilización y se hace necesario el uso de metodologías menos restrictivas, como por ejemplo la metodología semiparamétrica mencionada en la introdución de este trabajo y en la que los autores están trabajando actualmente.

En resumen, los puntos considerados en esta discusión, y a lo largo del artículo, ponen de manifiesto las dificultades subyacentes en el diseño, implementación e interpretación de la metodología paramétrica en análisis de supervivencia con datos no observados en las covariantes y justifican la investigación y el desarrollo de nuevos procedimientos.

6. AGRADECIMIENTOS

Este trabajo ha sido parcialmente subvencionado por el proyecto DGICYT PB95-0776 del Ministerio de Educación y Ciencia. Los autores agradecen a la Dra. Rotnitzky su colaboración durante los últimos años, al Dr. Caylà y a su equipo del Institut Municipal de la Salut la cesión de los datos que ilustran este trabajo, así como su colaboración, y a los evaluadores anónimos del artículo sus valiosos comentarios y sugerencias en la fase de revisión del mismo.

REFERENCIAS

Baker, S.G. (1994). «Regression Analysis of Grouped Survival Data with Incomplete Covariates: Nonignorable Missing-Data and Censoring Mechanisms». *Biometrics*, 50, 821-826.

Caylà, J.A., Jansà, J.M., Artacoz, L., Plasència, A. and AIDS–TB Group (1993). «Predictors of AIDS in a cohort of HIV–infected patients with pulmonary or pleural tuberculosis». *Tubercle and Lung Disease*, 74, 113–120.

Efron, B. (1994). «Missing Data, Imputation and the Bootstrap». *Journal of the American Statistical Association*, 89, 463–479.

Glynn, R.J., Laird, N.M. and Rubin, D.B. (1993). «Multiple Imputation in Mixture Models for Nonignorable Nonresponse With Follow-ups», *Journal of the American Statistical Association*, 88, 423, 984–993.

Glynn, R.J., Laird, N.M. and Rubin, D.B. (1986). «Selection Modeling Versus Mixture Modeling with Nonignorable Nonresponse», in *Drawing Inferences from Self Selected Samples*, Howard Wainer (editor). Springer-Verlag, New York, 115–157.

Little, R.J.A. and Rubin, D.B. (1987). *Statistical Analysis With Missing Data.* John Wiley, New York.

Miller, R.G. (1980). *Simultaneous Statistical Inference.* Springer–Verlag, New York.

Newey, W.K. (1990). «Semiparametric Efficiency Bounds». *Journal of Applied Econometrics*, 5, 99–135.

Robins, J.M., Rotnitzky, A. and Zhao, L.P. (1994). «Estimation of Regression Coefficients When Some Regressors Are Not Always Observed». *Journal of the American Statistical Association*, 89, 427, 846–866.

Rotnitzky, A. and Wypij, D. (1994). «A Note on the Bias of Estimators with Missing Data». *Biometrics*, 50, 1163–1170.

Serrat, C. and Gómez, G. (1995). «Estimació de paràmetres amb dades absents en un estudi d'anàlisi de supervivència». *Document de Recerca DR95/10.* Dept. Estadística i Investigació Operativa. Universitat Politècnica Catalunya.

Serrat, C., Gómez, G., García, P. and Caylà, J. (1998). «CD4+ Lymphocytes and Tuberculin Skin Test as Survival Predictors in Pulmonary Tuberculosis HIV-Infected Patients». *International Journal of Epidemiology*, 27, 703-712.

APÉNDICE

a) *Si el patrón de no respuesta es monótono, la condición de MAR*

$$P(R_i = r|L_i) = P(R_i = r|L_{(r)i}), i = 1, \cdots, n \qquad [A]$$

donde $r \in \{0,1\}^K$, *es equivalente a la condición*

$$P(R_{ik} = 1|R_{i(k-1)} = 1, L_i) = P(R_{ik} = 1|R_{i(k-1)} = 1, L_{i1}, \cdots, L_{i(k-1)}), i = 1, \cdots, n \qquad [B]$$

para cada $k = 1, \cdots, K$.

En particular, el proceso de no respuesta es MCAR si, y sólo si, para cada $k = 1, \cdots, K$ *la probabilidad de observación de la k-ésima variable condicionada a la observación de la variable* $(k-1)$*-ésima y a los datos potenciales es constante, es decir*

$$P(R_{ik} = 1|R_{i(k-1)} = 1, L_i) = P(R_{ik} = 1|R_{i(k-1)} = 1), i = 1, \cdots, n$$

para cada $k = 1, \cdots, K$.

Demostración: Cuando el patrón de no respuesta es monótono el espacio muestral del vector de respuesta, R_i -de dimensión K-, queda reducido al conjunto

$$\Omega_{R_i} = \{r_0 = (0, \cdots, 0), r_1 = (1, 0, \cdots, 0), r_2 = (1, 1, 0, \cdots, 0), \cdots, r_K = (1, \cdots, 1)\}$$

y la condición [A] se puede reescribir

$$P(R_i = r_k|L_i) = P(R_i = r_k|L_{i1}, \cdots, L_{ik}), i = 1, \cdots, n \qquad [A']$$

para cada $k = 0, \cdots, K$.

– Para probar la implicación [A] $\Rightarrow$ [B], demostraremos por inducción sobre el índice k que, bajo la hipótesis [A'], las probabilidades

$$P(R_{ik} = 1|L_i) \text{ y } P(R_{ik} = 1|R_{i(k-1)} = 1, L_i)$$

sólo dependen de los datos $L_{i1}, \cdots, L_{i(k-1)}$.

En efecto, si $k = 1$, tenemos

$$P(R_{i1} = 1|L_i) = 1 - P(R_{i1} = 0|L_i) = 1 - P(R_i = r_0|L_i)$$

que por la condición [A'] no depende de los datos.

Supongamos demostrado el enunciado hasta un índice j, $(1 < j < K)$. Para demostrar el resultado para el índice $j+1$, es suficiente comprobar las relaciones

(1) $$P(R_{i(j+1)} = 1|L_i) = P(R_{ij} = 1|L_i) - P(R_i = r_j|L_i)$$

(2) $$P(R_{i(j+1)} = 1|R_{ij} = 1, L_i) = 1 - \frac{P(R_i = r_j|L_i)}{P(R_{ij} = 1|L_i)}.$$

Como los segundos términos de las igualdades anteriores dependen como mucho, por la condición [A'] y la hipótesis de inducción, solamente de los datos $L_{i1}, \cdots, L_{ij}$, queda demostrado que la condición [A] implica la condición [B].

La expresión (1) se deduce directamente ya que, por la monotonía del patrón de no respuesta, $\{R_{ij}\}$ es unión disjunta de $\{R_{i(j+1)}\}$ y $\{R_i = r_j\}$.

De manera análoga, la igualdad (2) se obtiene a partir de

$$\begin{aligned} P(R_{i(j+1)} = 1|R_{ij} = 1, L_i) &= P(R_{ij} = 1|R_{ij} = 1, L_i) - P(R_i = r_j|R_{ij} = 1, L_i) = \\ &= 1 - P(R_i = r_j|R_{ij} = 1, L_i) = \\ &= 1 - \frac{P(R_i = r_j, R_{ij} = 1|L_i)}{P(R_{ij} = 1|L_i)} = \\ &= 1 - \frac{P(R_i = r_j|L_i)}{P(R_{ij} = 1|L_i)}. \end{aligned}$$

- Para ver que la condición [B] es suficiente para que el proceso de no respuesta sea MAR basta comprobar que se cumple la condición [A'].

  Si $k = 0$, la condición [A'] es cierta ya que $P(R_i = r_0|L_i) = P(R_{i1} = 0|L_i) = 1 - P(R_{i1} = 1|L_i)$, bajo la hipótesis [B], no depende de los datos.

  Para $k = 1, \cdots, K$, y utilizando que el patrón de no respuesta es monótono, tenemos

  $$\begin{aligned} P(R_i = r_k|L_i) &= P(R_{i1} = 1|L_i) \cdot P(R_{i2} = 1|R_{i1} = 1, L_i) \cdots \\ &\cdots P(R_{ik} = 1|R_{i(k-1)} = 1, L_i) \cdot P(R_{i(k+1)} = 0|R_{ik} = 1, L_i) \end{aligned}$$

  y si se cumple [B] todos los factores del segundo término de la igualdad anterior dependen, a lo sumo, de los datos $L_{i1}, \cdots, L_{ik}$, lo que prueba la implicación que queríamos demostrar.

En particular, que las probabilidades $P(R_{ik} = 1|R_{i(k-1)} = 1, L_i)$, $k = 1, \cdots, K$, sean constantes es la condición para que el proceso de no respuesta sea MCAR. Si denotamos por π_k la probabilidad condicionada $P(R_{ik} = 1|R_{i(k-1)} = 1, L_i)$ entonces las probabilidades del espacio muestral Ω_{R_i} son

$$P(R_i = r_k|L_i) = \begin{cases} 1-\pi_1 & \text{si } k=0 \\ \pi_1\pi_2\cdots\pi_k(1-\pi_{k+1}) & \text{si } k=1,\cdots,K-1, \\ \prod_{k=1}^{K}\pi_k & \text{si } k=K \end{cases}$$

que no dependen de los datos L_i.

■

b) *Si el patrón de no respuesta es monótono y los datos son MCAR, los p-valores, $p_1, p_2, \cdots, p_K$, resultantes del estadístico basado en la razón de verosimilitud de H_{0k} contra H_{Ak} en las K pruebas*

$$H_{0k} : logit(P(R_{ik} = 1|R_{i(k-1)} = 1, L_{i1}, \cdots, L_{i(k-1)})) = \alpha_{k1}$$

$$H_{Ak} : logit(P(R_{ik} = 1|R_{i(k-1)} = 1, L_{i1}, \cdots, L_{i(k-1)})) =$$

$$= \alpha_{k1} + \alpha_{k2}^t h_k(L_{i1}, \cdots, L_{i(k-1)}) \qquad k = 1, \cdots, K,$$

siguen una distribución uniforme en $(0,1)$ y son independientes entre sí.

Demostración: Fijado un valor de k, $k = 1, \cdots, K$, como el patrón de no respuesta es MCAR la hipótesis H_{0k} es verdadera y $logit(P(R_{ik} = 1|R_{i(k-1)} = 1)) = \alpha_{k1}$, y por lo tanto

$$P(R_{ik} = 1|R_{i(k-1)} = 1) = \frac{\exp(\alpha_{k1})}{1+\exp(\alpha_{k1})} = \pi_k.$$

En consecuencia, la variable aleatoria resultante de calcular las frecuencias relativas del suceso «observación de la variable k-ésima condicionada a la observación de la variable $(k-1)$-ésima», $\{R_{ik} = 1|R_{i(k-1)} = 1\}$, y que denotamos por f_{ik}, se comporta, asintóticamente, como una variable aleatoria $N\left(\pi_k, \sqrt{\frac{\pi_k(1-\pi_k)}{n_k}}\right)$, donde n_k es el número de individuos para los cuales se ha observado la variable $L_{i(k-1)}$ (por construcción, n_1 es el tamaño muestral).

Si p_k es el p-valor resultante del estadístico basado en la razón de verosimilitud de H_{0k} contra H_{Ak}, tenemos $p_k = P(|f_{ik} - \pi_k| > |f_{ik,datos} - \pi_k|)$.

Para demostrar que $p_k \sim U(0,1)$ es suficiente demostrar que $\forall p \in [0,1]$, $P(p_k \leq p) = p$. Y en efecto, fijado p, $0 \leq p \leq 1$, si denotamos por I_{1-p} el correspondiente intervalo con probabilidad $1-p$ centrado en π_k para la distribución $N\left(\pi_k, \sqrt{\frac{\pi_k(1-\pi_k)}{n_k}}\right)$, obtenemos $P(p_k \leq p) = P(f_{ik} \notin I_{1-p}) = p$, como queríamos probar.

Los K p-valores, $p_1, p_2, \cdots, p_K$, son independientes entre sí por el hecho de que las probabilidades condicionadas π_k no dependen de los datos L_i.

■

c) *Si el patrón de no respuesta es MAR, la maximización de la función de verosimilitud no depende de las probabilidades de no respuesta.*

Demostración: De acuerdo con la notación introducida, es suficiente demostrar que la expresión $L(\theta, \psi)$ se puede descomponer en producto de dos funciones $L_1(\theta)$ y $L_2(\psi)$.

En efecto, si las probabilidades de no respuesta $P(R_i = r|L_i; \psi)$ no dependen de los datos no observados $L_{(\bar{r})i}$ entonces

$$\int f_c(L_i;\theta) \cdot P(R_i = r|L_i;\psi) dL_{(\bar{r})i} = P(R_i = r|L_i;\psi) \cdot \int f_c(L_i;\theta) dL_{(\bar{r})i}$$

y la expresión

$$\begin{aligned} L(\theta,\psi) = \prod_{i=1}^{n} \{ & [f_c(L_i;\theta) \cdot P(R_i = \mathbf{1}|L_i;\psi)]^{I(R_i=\mathbf{1})} \\ & \prod_{r\neq\mathbf{1}} \left[\int f_c(L_i;\theta) \cdot P(R_i = r|L_i;\psi) dL_{(\bar{r})i}\right]^{I(R_i=r)} \Big\} \end{aligned}$$

admite la descomposición $L(\theta, \psi) = L_1(\theta) \cdot L_2(\psi)$ donde

$$\begin{aligned} L_1(\theta) &= \prod_{i=1}^{n} \left\{ f_c(L_i;\theta)^{I(R_i=\mathbf{1})} \cdot \prod_{r\neq\mathbf{1}} \left[\int f_c(L_i;\theta) dL_{(\bar{r})i}\right]^{I(R_i=r)} \right\} \\ L_2(\psi) &= \prod_{i=1}^{n} \left\{ \prod_r P(R_i = r|L_i;\psi)^{I(R_i=r)} \right\}. \end{aligned}$$

■

ENGLISH SUMMARY

SURVIVAL STUDIES WITH NON OBSERVED DATA. DIFFICULTIES CONCERNING THE PARAMETRIC APPROACH

G. GÓMEZ*
C. SERRAT**
Universitat Politècnica de Catalunya

We analyze a parametric approach in survival studies when some of the covariates have not been observed in some individuals. Difficulties in the design, in the computations, as well as philosophicals, that arise in the specification of the likelihood function and in its optimization are discussed. A sensitivity analysis of the estimators, implemented in S-PLUS, is presented. As illustration, the introduced methodology is applied in a survival study in a cohort of pulmonary tuberculosis HIV–infected patients.

Keywords: Incomplete data models, maximum likelihood, missing at random, non-ignorable non-response, survival analysis, validation study.

AMS Classification: 62F10, 62H99, 92C60.

This work has been partially supported by DGICYT project PB95–0776.

* Departament d'Estadística i Investigació Operativa. Universitat Politècnica de Catalunya. Pau Gargallo, 5. 08028 Barcelona. E-mail: ggg@eio.upc.es.

** Departament de Matemàtica Aplicada I. Universitat Politècnica de Catalunya. Av. Dr. Gregorio Marañón, 44–50. 08028 Barcelona. E-mail: carles@mal.upc.es.

– Received February 1998.
– Accepted February de 1999.

The missing data problem is already a classical problem that has not been yet solved satisfactorily. Most of the existing methodologies are based on the assumption that non observed data are missing completely at random or at most missing at random. This problem includes those situations where the dependent variable is the survival time which itself could be censored.

A brief review to the different approaches and their advantages and inconvenients follows. First, we could base the analysis on those individuals with all the observed covariates. The inference based on the so-called complete sample is biased and not consistent because the observed individuals are not necessarily a good representation of the overall sample. Furthermore, the analysis based on the complete sample would be inefficient due to reduction in the sample size. Therefore, although this approach is appealing because can be handled with the existing software, its use has to be avoided.

A second approach consists in the imputation of the non observed values. The main problem here relies in that the observed values are used to model the non observed ones and this assumption might not be true. The resulting estimators could then be seriously biased.

A parametric modelization and the solution via maximum likelihood is another possible approach. As it is known this methodology is asymptotically efficient. This is a widely used method; however, its implementation is not straightforward and the corresponding estimators rely heavily on a large number of assumptions that cannot be validated.

The semiparametric approach is a fourth way of handling the missing data problem that allows to model only what is strictly necessary; in our situation we will have to model the relationship between survival and the covariates and the non-response pattern.

In this paper, and previously to an analysis based on a semiparametric approach, we use a completely parametric point of view. This parametric analysis has two main goals: on one hand to show the inconvenients, practical and philosophical, in the specification of the likelihood function and in its optimization, and on the other to design a methodology that would allow to determine how the resulting estimators depend on the non-response pattern.

We start introducing some needed terminology. The potential data vector $L = (L_1, L_2, \cdots, L_K)$ for an arbitrary individual is defined as the vector that contains his/her observed and not observed data. The response vector $R = (R_1, R_2, \cdots, R_K)$ for this individual has k-th ($k = 1, \cdots, K$) component equal to 1 if the k-th variable has been observed and 0 otherwise. L is split into the subvectors $L_{(R)}$ and $L_{(\bar{R})}$ corresponding to the observed and unobserved data, respectively.

Different non-response patterns will depend on the values that $\pi_L(r) = P(R = r|L)$ takes for an arbitrary individual, where $r = (r_1, r_2, \cdots, r_K), r_k \in \{0, 1\}, k = 1, \cdots, K$.

The non-response process is Missing Completely at Random (MCAR), if and only if, $\pi_L(r)$ is independent of L. The process is Missing at Random (MAR), if and only if, $\pi_L(r)$ depends at most of $L_{(r)}$. The non-response process is Non-Ignorable (NI) if $\pi_L(r)$ depends on $L_{(\bar{r})}$.

In a survival study the main variable T, is usually the elapsed time between an origin (randomization date in a clinical trial, treatment initiation, etc) and the realization of an event (death, diagnosis of AIDS, etc). Data in these studies is often right censored and the observed data are $Y = min\{T, C\}$ and $\delta = \mathbf{1}\{T \leq C\} = \mathbf{1}\{Y = T\}$ where C is the censoring time. For each individual we also have the values of certain covariates. Let X be the covariates vector. If X_* is a subvector of X and V_* is another set of covariates, we will say that V_* is a surrogate of X_* if X_* and V_* are strongly correlated. The goal in a survival study, with missing data in the covariates, is to model T in terms of the covariates vector X, using the information provided by X and by the surrogate vector V_*.

The sequence of the paper is now the following. We start testing whether the non-response process is MCAR. Then, we introduce a hierarchical scheme to check the sensitivity of the model parameters under different non-response patterns with the objective to elucidate about the non-ignorability of the non-response process. The paper ends with an exhaustive illustration where the methodology is applied and the main disadvantages of the parametric approach are shown.

The test to check whether or not the non-response process is MCAR is based on the comparison of the probabilities $P(R_{ik} = 1), i = 1, \cdots, n$ and $P(R_{ik} = 1|L_i), i = 1, \cdots, n$ for every $k = 1, \cdots, K$. The hypothesis test is formulated as H_0 : MCAR non-response process versus H_A : MAR or NI non-response process, and two different procedures are developed depending on whether or not the non-response process is monotonous.

To develop the parametric approach, let $f_c(l;\theta)$ be the density function for the complete data and $P(R_i = r|L_i;\psi)$ be the probability of observing certain L_i components. We wish to estimate the parameters θ and ψ by maximum likelihood when the likelihood function from the observed data is given by

$$L(\theta,\psi) = \prod_{i=1}^{n} \left\{ [f_c(L_i;\theta) \cdot P(R_i = \mathbf{1}|L_i;\psi)]^{I(R_i=\mathbf{1})} \right.$$
$$\left. \prod_{r \neq \mathbf{1}} \left[\int f_c(L_i;\theta) \cdot P(R_i = r|L_i;\psi) dL_{(\bar{r})i} \right]^{I(R_i=r)} \right\}.$$

It can be proved that if the non-response process is either MCAR or MAR, the maximum likelihood estimator of θ is independent of the non-response process.

In our survival study the density function can be expressed as

$$f_c(L_i;\theta) = f_c((Y_i, \delta_i, V_i^t, X_i^t)^t;\theta) = f_c(X_i;\theta) \cdot f_c(Y_i, \delta_i|X_i;\theta) \cdot f_c(V_i|Y_i, \delta_i, X_i;\theta).$$

To solve the problem parametrically we have to correctly specify: a) the distribution of the covariates, X, b) the distribution of the observed times, Y, conditioned to X, c) the distribution of the surrogate covariates, V, conditioned to Y and to X and d) the non-response probabilities conditioned to the potential data.

To illustrate the above methodology we study survival time from tuberculosis in patients who have contracted the human immunodeficiency virus. The following variables are collected for each individual in the sample: sex, age, risk group, previous tuberculosis treatment, tuberculin skin test (PPD for short), radiological test, bacteriological test, T-CD4+ levels (CD4 for short), ... among others. The main goal in this study is to determine the survival predictive role of CD4 and PPD. The methodological problem relies in the fact that the percentage of missing data in CD4 and PPD are respectively 37.5% and 50.5%. Following some discussions with the epidemiologists, the following variables are decided to use as surrogates of CD4 and PPD: previous treatment (TR), radiology (RA) and bacteriology (BA). The MCAR validation is carried out and it is concluded that the null hypothesis cannot be rejected. The parametric problem is settled up after modelling: a) the distribution of the main variables, CD4 and PPD, through conditional logistic models, b) the observed survival time, conditioned to CD4 and PPD, via a Weibull model, c) the distributions of the surrogate covariates, TR,RA and BA, conditioned to CD4, PPD and the observed survival time, through the logarithms of the odds ratios, d) the non-response pattern is modelled as conditional logistic probabilities with respect to the covariates in the following way:

$$\begin{aligned}
logit(P(R_{CD4}=1)) &= \alpha_{10}+\alpha_{11}TR+ \\
&+ \alpha_{12}RA1+\alpha_{13}RA2+\alpha_{14}BA1+\alpha_{15}BA2 \\
&+ \alpha_{16}\mathbf{1}\{CD4>14\}+\alpha_{17}PPD, \\
logit(P(R_{PPD}=1|R_{CD4}=1)) &= \alpha_{20}+\alpha_{21}TR+ \\
&+ \alpha_{22}RA1+\alpha_{23}RA2+\alpha_{24}BA1+\alpha_{25}BA2 \\
&+ \alpha_{26}\mathbf{1}\{CD4>14\}+\alpha_{27}PPD \quad \text{y} \\
logit(P(R_{PPD}=1|R_{CD4}=0)) &= \alpha_{30}+\alpha_{31}TR+ \\
&+ \alpha_{32}RA1+\alpha_{33}RA2+\alpha_{34}BA1+\alpha_{35}BA2 \\
&+ \alpha_{36}\mathbf{1}\{CD4>14\}+\alpha_{37}PPD.
\end{aligned}$$

Thus, if θ stands for all the parameters involved in the a) through c) modelization and $\psi=(\alpha_{10},\cdots,\alpha_{37})$ corresponds to those in d), the likelihood $L(\theta,\psi)$ will depend, at most, on a 135 dimension parameter vector.

The way that the different non-response patterns are designed in d) allows us to establish the following five cases:

$$\alpha_{ij} = 0 \quad i = 1,2,3 \quad j = 1,\cdots,7 \Rightarrow \text{MCAR}$$

$$\alpha_{ij} = 0 \quad i = 1,2,3 \quad j = 6,7 \Rightarrow \text{MAR}$$

$$\alpha_{ij} = 0 \quad i = 1,2,3 \quad j = 7 \Rightarrow \text{NI1}$$

$$\alpha_{ij} = 0 \quad i = 1,2,3 \quad j = 6 \Rightarrow \text{NI2}$$

$$\text{Without restrictions on the values of } \alpha_{ij} \Rightarrow \text{NI3}$$

Tables 1 and 2 summarize the results. It can be concluded that in general the MAR model is not significatively different from the MCAR model. Therefore, the probability of observing CD4 and PPD does not depend on the surrogate observed values. In all the settings, the NI models are significatively different from the MAR. Thus, the probabilty of observing CD4 and PPD depends on their potential values. Finally, it can be concluded that the positivity in the tuberculin skin test is a better predictor for short-time survival.

The parametric approach developed in this paper is of limited use due to the following considerations. First, the methodology depends on a large number of assumptions that can be quite arbitrary and cannot be validated from the observed data. As a consequence, the estimators might be strongly assumption-dependent. Secondly, the choice of the surrogate covariates is crucial in the sense that they have to be based on clinical and epidemiological considerations. Thirdly, the geometrical growth of the parameter dimension is another limitation of this approach. Finally, the execution of a complete analysis is extremely computationally costly –from minutes to days–.

Summarizing, alternative ways to analyze survival models with missing covariates have to be developed. The authors are presently working on the semiparametric approach.

Secció Docent i Problemes

SECCIÓ DOCENT I PROBLEMES

La «Secció docent i problemes» té l'objectiu de publicar articles de caire docent, difícilment publicables en revistes de recerca. A cada número de *Qüestiió* s'inclouen d'un a tres problemes i les solucions es donen en el número següent.

Els lectors poden 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.

SOLUCIONS ALS PROBLEMES PROPOSATS AL VOLUM 23 N. 1

PROBLEMA N. 74

Llamando

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2, \tag{1}$$

De Vries (1986, p. 8) ha demostrado que

$$S^2 = \frac{n-1}{n^2}\sum_{i=1}^{n}x_i^2 - \frac{2}{n^2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}x_i x_j. \tag{2}$$

Veamos ahora que

$$S^2 = \frac{1}{n^2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(x_i - x_j)^2 \tag{3}$$

por inducción en n.

Para $n = 1$, S^2 no está definida o vale cero, lo que supone un caso obviamente sin interés práctico especial.

Partimos del caso $n = 2$. Según la fórmula (1),

$$S^2 = \frac{1}{2}\sum_{i=1}^{2}\left[x_i - \frac{1}{2}(x_1 + x_2)\right]^2 = \frac{1}{2}\left\{\left[\frac{1}{2}(x_1 - x_2)\right]^2 + \left[\frac{1}{2}(-x_1 + x_2)\right]^2\right\} =$$

$$= \frac{1}{8}\left(2x_1^2 + 2x_2^2 - 4x_1x_2\right) = \frac{1}{2^2}(x_1 - x_2)^2 = \frac{1}{2^2}\sum_{i=1}^{1}\sum_{j=i+1}^{2}(x_i - x_j)^2,$$

que es también la fórmula (3) para $n = 2$. Con esto queda demostrada la fórmula (3) en el caso $n = 2$.

Ahora suponemos que la fórmula (3) es cierta para n ($n \geq 2$), y vamos a demostrar o justificar que la fórmula (3) sigue siendo cierta para $n + 1$. La hipótesis de inducción es que (3) es válida para n, y en lo que resta demostraremos que (3) será válida cuando tenemos un dato más, x_{n+1}, basándonos en la hipótesis de inducción (que consiste en que las relaciones (1) y (3) son iguales para n) y en que la fórmula (2) es igual a la (1),

ya demostrado para cualquier valor entero de n ($n \geq 1$) por De Vries (1986). Usaremos como subíndice en la notación de S^2 ó $\bar{x}$ al número de datos de la variable, por lo que

$$S^2_{n+1} = \frac{1}{n+1} \sum_{i=1}^{n+1} (x_i - \bar{x}_{n+1})^2 =$$

$$= \frac{1}{n+1} \left[\sum_{i=1}^{n} (x_i - \bar{x}_n + \bar{x}_n - \bar{x}_{n+1})^2 + (x_{n+1} - \bar{x}_{n+1})^2 \right] =$$

$$= \frac{1}{n+1} \left[\sum_{i=1}^{n} (x_i - \bar{x}_n)^2 + n (\bar{x}_n - \bar{x}_{n+1})^2 + (x_{n+1} - \bar{x}_{n+1})^2 \right] =$$

por la fórmula (1),

$$= \frac{1}{n+1} \left[n S^2_n + \left(n \bar{x}^2_{n+1} + \bar{x}^2_{n+1} \right) - \right.$$

$$\left. - \left(2n \bar{x}_n \bar{x}_{n+1} + 2x_{n+1} \bar{x}_{n+1} \right) + \left(n \bar{x}^2_n + x^2_{n+1} \right) \right] =$$

y haciendo uso de la hipótesis de inducción para n

$$= \frac{1}{n+1} \left\{ \frac{1}{n} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (x_i - x_j)^2 + \left[(n+1) \bar{x}^2_{n+1} - 2\bar{x}_{n+1} (x_{n+1} + n\bar{x}_n) \right] + \right.$$

$$\left. + \left[x^2_{n+1} + \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right)^2 \right] \right\} =$$

$$= \frac{1}{n(n+1)} \left[\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (x_i - x_j)^2 - n(n+1) \bar{x}^2_{n+1} + \right.$$

$$\left. + \left(n x^2_{n+1} + \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} \sum_{j \neq i}^{n} x_i x_j \right) \right]. \tag{4}$$

Por otro lado, hay que demostrar que (4) coincide con

$$S_{n+1}^2 = \frac{1}{(n+1)^2} \sum_{i=1}^{n} \sum_{j=i+1}^{n+1} (x_i - x_j)^2 =$$

$$(5) \qquad = \frac{1}{(n+1)^2} \left[\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (x_i - x_j)^2 + \sum_{i=1}^{n} (x_i - x_{n+1})^2 \right].$$

El resultado buscado será cierto si, y sólo si, los dos últimos términos de (4) y (5) coinciden, y esto se cumple si, y sólo si, se da la siguiente igualdad:

$$\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (x_i - x_j)^2 - n(n+1)^2 \bar{x}_{n+1}^2 + n(n+1) x_{n+1}^2 +$$

$$+ (n+1) \sum_{i=1}^{n} x_i^2 + (n+1) \sum_{i=1}^{n} \sum_{j \neq i} x_i x_j -$$

$$- \left(n \sum_{i=1}^{n} x_i^2 + n^2 x_{n+1}^2 - 2n^2 x_{n+1} \bar{x}_n \right) = 0,$$

o bien, haciendo uso de la hipótesis de inducción o de la fórmula (3) para n, la condición necesaria y suficiente es que se dé la igualdad:

$$n^2 S_n^2 - n \left(\sum_{i=1}^{n+1} x_i \right)^2 + n x_{n+1}^2 + \sum_{i=1}^{n} x_i^2 +$$

$$+ (n+1) \sum_{i=1}^{n} \sum_{j \neq i}^{n} x_i x_j + \left(2n x_{n+1} \sum_{i=1}^{n} x_i \right) = 0,$$

y simplificando quedará:

$$n^2 S_n^2 - n \left[\left(\sum_{i=1}^{n} x_i \right)^2 + x_{n+1}^2 + 2x_{n+1} \sum_{i=1}^{n} x_i \right] + n x_{n+1}^2 +$$

$$+ \sum_{i=1}^{n} x_i^2 + (n+1) \sum_{i=1}^{n} \sum_{j \neq i}^{n} x_i x_j + 2n x_{n+1} \sum_{i=1}^{n} x_i = 0,$$

o bien:

$$n^2 S_n^2 + \sum_{i=1}^{n} \sum_{j \neq i}^{n} x_i x_j - (n-1) \sum_{i=1}^{n} x_i^2 = 0,$$

es decir si, y sólo si:

$$S_n^2 = \frac{n-1}{n^2} \sum_{i=1}^{n} x_i^2 - \frac{2}{n^2} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} x_i x_j,$$

y ésta es cierta ya que es la fórmula (2) debida a De Vries (1986). De aquí concluimos que para $n \geq 2$ y n número natural, la fórmula (3) es otra expresión alternativa de la varianza $S^2 = S_n^2$ que no se refiere explícitamente a la media $\bar{x} = \bar{x}_n$. La fórmula (1) refiere la varianza a la media, pero las fórmulas (2) y (3) de la varianza no se refieren a la media explícitamente.

REFERENCIA

De Vries, P.J. (1986). *Sampling Theory for Forest Inventory.* Springer-Verlag. Berlín.

M. Ruíz Espejo
UNED

PROBLEMA N. 75

Para el estadístico media muestral $\bar{x}_n$, basado en la muestra aleatoria simple con reemplazamiento $(x_1, x_2, ..., x_n)$,

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i = t_n \text{ (digamos)},$$

de donde

$$\bar{x}_{n-1,i} = \frac{1}{n-1}\sum_{j\neq i}^{n} x_j = t_{n-1,i} \text{ (digamos)}$$

es la media muestral de la muestra aleatoria simple excluyendo la componente i-ésima x_i. También,

$$\bar{t}_{n-1} = \frac{1}{n}\sum_{i=1}^{n} t_{n-1,i} = \frac{1}{n}\sum_{i=1}^{n} \bar{x}_{n-1,i} = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} x_j =$$

$$= \frac{1}{n(n-1)}\sum_{i=1}^{n}(n\bar{x}_n - x_i) =$$

$$= \frac{1}{n(n-1)}\left(n^2\bar{x}_n - n\bar{x}_n\right) = \bar{x}_n,$$

por lo que

$$t'_n = nt_n - (n-1)\bar{t}_{n-1} = n\bar{x}_n - (n-1)\bar{x}_n = \bar{x}_n,$$

y de aquí, el estimador jackknife de la varianza de la media muestral, es

$$\widehat{V}_J(\bar{x}_n) = \frac{n-1}{n}\sum_{i=1}^{n}\left(t_{n-1,i}^2 + \bar{t}_{n-1}^2 - 2t_{n-1,i}\bar{t}_{n-1}\right) =$$

$$= \frac{n-1}{n}\left(\sum_{i=1}^{n}\bar{x}_{n-1,i}^2 + n\bar{x}_n^2 - 2\bar{x}_n\sum_{i=1}^{n}\bar{x}_{n-1,i}\right) = \frac{n-1}{n}\left(\sum_{i=1}^{n}\bar{x}_{n-1,i}^2 - n\bar{x}_n^2\right) =$$

$$= (n-1)\left(\frac{1}{n}\sum_{i=1}^{n}\bar{x}_{n-1,i}^2 - \bar{x}_n^2\right). \tag{1}$$

Y ya que si denotamos por

$$a_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$$

al momento muestral respecto al origen de orden k, tenemos que

$$\begin{aligned}
\frac{1}{n}\textstyle\sum_{i=1}^{n} \bar{x}_{n-1,i}^2 &= \frac{1}{n}\textstyle\sum_{i=1}^{n}\left(\frac{1}{n-1}\sum_{j\neq i}^{n} x_j\right)^2 = \\
&= \frac{1}{n(n-1)^2}\textstyle\sum_{i=1}^{n}\left(\sum_{j\neq i}^{n} x_j^2 + \sum_{j\neq i}^{n}\sum_{k\neq i,j}^{n} x_j x_k\right) = \\
&= \frac{1}{n(n-1)^2}\left[\textstyle\sum_{i=1}^{n}\left(na_2 - x_i^2\right) + \sum_{i=1}^{n}\sum_{j\neq i}^{n} x_j \sum_{k\neq i,j}^{n} x_k\right] = \\
&= \frac{1}{n(n-1)^2}\left[\left(n^2 a_2 - na_2\right) + \textstyle\sum_{i=1}^{n}\sum_{j\neq i}^{n} x_j\left(na_1 - x_i - x_j\right)\right] = \\
&= \frac{1}{n(n-1)^2}\left[n(n-1)a_2 + \textstyle\sum_{i=1}^{n} na_1\left(na_1 - x_i\right) - \right. \\
&\left. - \textstyle\sum_{i=1}^{n} x_i\left(na_1 - x_i\right) - \sum_{i=1}^{n}\sum_{j\neq i}^{n} x_j^2\right] = \\
&= \frac{1}{(n-1)^2}\left[n(n-2)a_1^2 + a_2\right]. \qquad (2)
\end{aligned}$$

Luego, de (1) y (2) en las dos primeras igualdades respectivamente:

$$\begin{aligned}
\widehat{V}_J(\bar{x}_n) &= (n-1)\left(\frac{1}{n}\textstyle\sum_{i=1}^{n}\bar{x}_{n-1,i}^2 - \bar{x}_n^2\right) = \\
&= (n-1)\left\{\frac{1}{(n-1)^2}\left[n(n-2)a_1^2 - a_2\right] - a_1^2\right\} = \\
&= \frac{1}{n-1}\left[n(n-2)a_1^2 + a_2 - (n-1)^2 a_1^2\right] = \\
&= \frac{1}{n-1}\left(a_2 - a_1^2\right) = \frac{1}{n(n-1)}\textstyle\sum_{i=1}^{n}\left(x_i - a_1\right)^2.
\end{aligned}$$

M. Ruíz Espejo
UNED

PROBLEMES PROPOSATS

PROBLEMA N. 76

Sea un sistema de espera abierto con *una* estación de servicio, en el que los clientes llegan según una ley de Poison con media λ y en el que la duración del servicio sigue una ley exponencial con media $1/\mu$. En el supuesto de existir régimen permanente, obtener:

1. La distribución de probabilidad del número de clientes: $\Pi_n \equiv p(X = n)$.
2. En número medio de clientes en el sistema, de clientes esperando y de *unidades desocupadas.*
3. Determinar $P(X \leq n)$, la probabilidad de espera y el tiempo medio de espera.
4. ¿Qué número de usuarios se encuentra en el sistema con mayor probabilidad ?
5. ¿Para qué valor de la intensidad de tráfico se encuentra la máxima probabilidad de un número de unidades en el sistema dado? Analizar los casos $n = 1$ y $n = 2$.

R. Alonso

PROBLEMA N. 77

En los supuestos del Problema 1, se considera adicionalmente que un cliente que llega al sistema y no encuentra la unidad de servicio vacía rechaza entrar con probabilidad α.

a) Resolver 1), 2) y 3) del Problema 1.

b) Estudiar el supuesto $\alpha = 1$.

R. Alonso

PROBLEMA N. 78

En los supuestos del Problema 1, se considera adicionalmente que se produce abandono del sistema, antes de ser atendido, según una tasa β constante, independiente del número de clientes. Determinar:

1) La distribución de probabilidad del número de clientes.

2) El número medio de clientes en el sistema

R. Alonso

PROBLEMA N. 79

En los supuestos del Problema 1, se considera adicionalmente que se produce abandono del sistema, antes de ser atendido, según una tasa β y que un cliente que llega al sistema y no encuentra la unidad de servicio libre rechaza entrar con probabilidad α (supuestos adicionales de los problemas 2 y 3). Obtener:

a) La distribución de probabilidad del número de clientes

b) El número medio de clientes en el sistema.

R. Alonso

PROBLEMA N. 80

En los supuestos del Problema 1, se considera adicionalmente que un cliente que llega al sistema y no encuentra la unidad de servicio vacía rechaza entrar con probabilidad $b(n) = \frac{n}{n+1}$. Obtener:

a) La distribución de probabilidad del número de clientes (X).

b) El número medio de clientes en el sistema.

R. Alonso

PROBLEMA N. 81

En los supuestos del Problema 1, se considera adicionalmente que el sistema es de capacidad limitada (K) y la tasa de circulación (ψ) es igual a la unidad. Determinar:

a) La distribución de probabilidad del número de clientes.

b) El número medio de clientes en el sistema.

c) El número medio de unidades desocupadas.

d) El número medio de clientes en espera.

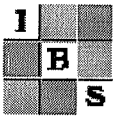
R. Alonso

Ressenyes d'activitats institucionals

Sociedad Española de Biometría

Región Española
de la Sociedad Internacional de Biometría

Spanish Region
of the International Biometric Society



http://www.iata.csic.es/ibsresp

La **Sociedad Española de Biometría/Región Española de la Sociedad Internacional de Biometría** (abreviadamente **SEB** o ***REsp***) tiene como objetivos promover, impulsar y difundir el desarrollo y la aplicación de los métodos matemáticos y estadísticos a la biología, medicina, psicología, farmacología, agricultura y otras ciencias afines (ciencias relacionadas con los seres vivos). Cualquier profesional o alumno de estas disciplinas puede ser miembro de la SEB.

Consejo Directivo

Presidenta:	Guadalupe Gómez i Melis (Biología)
Vicepresidente:	Emilio A. Carbonell Guevara (Agronomía)
Secretario y Tesorero:	Fernando López Santoveña (Agronomía)
Vocal en calidad de Miembro del Consejo de la IBS:	Carles M. Cuadras Avellana (Biología)
Vocales:	María Jesús Bayarri García (Medicina)
	Juan Luis Chorro Gascó (Psicología)
	Rosa Estarelles Rodríguez (Psicología)
	José Luis González Andújar (Agronomía)
	Martín Ríos Alcolea (Medicina)
	Alex Sánchez Pla (Biología)
Corresponsal de la REsp en el «Biometric Bulletin» de la IBS:	María Luz Calle Rosingana

La SEB promueve directamente o participa en la promoción de cursos monográficos sobre distintas técnicas de Análisis Estadístico.

Durante 1998 se celebró un curso sobre «Regresión Logística» y otro sobre «Modelos Mixtos con S-Plus», en colaboración con la Universidad Politécnica de Valencia.

En 1999 se ha celebrado en Barcelona un curso sobre Análisis de Supervivencia con S-Plus, en colaboración con la Fundació Politècnica de Catalunya, y ha tenido lugar en Palma de Mallorca la VII Conferencia Española de Biometría.

The TES Institute
Training of European Statisticians

GENERAL INTRODUCTION

Seeing the need of harmonising statistics at the European level, Eurostat decided in the early nineties to set up the TES Project. This programme was in charge of offering truly European vocational training and staff development opportunities at post-graduate level through annual training programmes for target groups ranging from young statisticians to executives of National Statistical Institutes. In November 1996, the TES Project became the TES Institute, a non-profit association created by ten Member States of the European union and the four Member States of the European Free Trade Association. At present it counts among its members the representatives of seventeen European National Statistical Institute and of the Centre Universitaire de Luxembourg.

The training programmes offered by the TES Institute provide both theoretical and practical background but the courses have all a very strong applied character.

These programmes also offer participants the opportunity to meet colleagues form all over Europe and other countries since the TES Institute has extended its activities to the Central European, Mediterranean Basin and TACIS countries.

The above characteristics represent the basic conditions to acquire sharper competence in their work environment and highlight the European dimension of their activity.

After ten years of existence, the programme became entire part of the statistical world. For the time being, around 500 participants coming from more than 30 countries are trained every academic year. Such an interest is mainly due to the large number of courses on offer. Indeed, the TES portfolio comprises more than 80 courses of short duration all at post-graduate level.

After a few years of co-operation with the Central European countries, the TES Institute has recently extended the co-operation to the MEDSTAT and TACIS region. Such an internationalisation is the direct result of the growing importance of training as a part of the current technological and intellectual development. Therefore, as far as statistics and economics are concerned, it is of the utmost importance to extend the best national practices to an international level.

It is obvious that the TES programmes should be considered as a complement and not a substitute to the training provided at national level. In brief, one may say that by offering training opportunities which are complementary to the ones provided at national level, the TES Institute is offering a new approach of the subsidiarity concept.

COURSES ON OFFER

The Core Programme

The annual training programmes are designed for public sector statisticians in the Member States of the EU and EFTA. However, the programmes are also open to the National Statistical Institutes of countries in Central and Eastern Europe, the Mediterranean Basin, some other selected countries and private sector statisticians.

The leading principle for the definition of the annual programmes and the detailed content of each of the courses is that choices are made on the basis of the training needs in the Member States of the EU and the EFTA. The *Core Programme* is subsidised by the Statistical Office of European Commission (Eurostat).

The TES Institute will offer 28 courses in the framework of its 1999-2000 Core Programme. This year, the accent is put on the *Stream 2: Economic Statistics* and on the *Support Courses* with respectively eight and seven courses.

The TES Institute will carry nine courses until the end of this year, other sixteen courses will be held next year.

Course Code	Course Title	Duration (days)	Course Leader	Course Site	From	To	TES*
SOC-001-E	Systems of Social Statistics	5	EVERAERS	Barcelona	04-Oct-99	08-Oct-99	ND
ECO-102-E	Business Registers: Set-up, Use and Maintenance	5	BERGSTRØM	Oslo	11-Oct-99	15-Oct-99	MC
PDU-103-E	Techniques of Electronic Data Dissemination	5	ARGÜESO JIMENEZ	Sevilla	18-Oct-99	22-Oct-99	ND
SUP-204-E	Measurement of the Quality of Statistics	3	DEPOUTOT & ELVERS	Luxembourg	25-Oct-99	27-Oct-99	MC
ECO-202-E	The Revised European System of Accounts (ESA 95) - Sector Accounts	3	NEWSON	Luxembourg	08-Nov-99	10-Nov-99	ND
ECO-001-E	National Accounts Statistics in Practice	10	VAN NUNSPEET	Voorburg	08-Nov-99	19-Nov-99	MC
SUP-103-E	Price Index Theory and Price Statistics	5	VON DER LIPPE	Düsseldorf	22-Nov-99	26-Nov-99	ND
ECO-205-E	The Revised European System of Accounts (ESA 95) - Quarterly Accounts	4	MAZZI	Luxembourg	06-Dec-99	09-Dec-99	MC
DAT-003-F	Sampling Techniques and Practice	10	DEVILLE	Bruz	06-Dec-99	17-Dec-99	ND
ECO-204-E	The Revised European System of Accounts (ESA 95) - Goods and Services	3	To be announced	Luxembourg	10-Jan-00	12-Jan-00	MC
PDU-001-E	Basic Principles of Publication and Dissemination of Statistical Products	5	SWIRES-HENNESSY	London	17-Jan-00	21-Jan-00	ND
SOC-103-E	Living Conditions, Social Indicators Social Reporting	4	EVERAERS	Lisbon	24-Jan-00	27-Jan-00	MC
SUP-202-E	Introduction to the Analysis of Multivariate Discrete Data	5	ISRAËLS	Voorburg	07-Feb-00	11-Feb-00	ND
SUP-206-E	Symbolic Data Analysis	5	DIDAY	Paris	07-Feb-00	11-Feb-00	MC
ECO-101-E	Nomenclatures, Classifications and Their Harmonisation	4	LANGKJÆR	Luxembourg	14-Feb-00	17-Feb-00	ND
PDU-109-E	Workshop on the Internet	2	FELDBÆK	Neuchâtel	17-Feb-00	18-Feb-00	MC
COM-001-M1	The European Statistical System	2.5	RODRIGUEZ PRIETO	Luxembourg	06-Mar-00	08-Mar-00	ND
COM-003-M	Confidentiality and Protection of Privacy	3	NANOPOULOS	Luxembourg	13-Mar-00	15-Mar-00	MC
SUP-303-E	Adding Value through Strategic Management	5	SCRIVENER	London	20-Mar-00	24-Mar-00	ND
DAT-102-E	Dealing with non-Response	5	LYNN	London	3-Apr-00	7-Apr-00	MC
SUP-201-E	Seasonal Adjustment Methods	5	MARAVALL	Luxembourg	10-Apr-00	14-Apr-00	ND
ECO-203-E	The Revised European System of Accounts (ESA 95) - Financial Accounts	3	COIN	Luxembourg	26-Apr-00	28-Apr-00	MC
DAT-001-E	Notions of Sampling and Survey for Managers	3	DROESBEKE	Rome	08-May-00	10-May-00	ND
SOC-106-E	Demographic Data and Their Analysis	5	ANDERSEN	Copenhagen	22-May-00	26-May-00	MC
SUP-104-E	Comparative Analysis of Statistical Packages and Data Bases for Statistics	3	COLE & CAMPBELL	Manchester	05-Jun-00	07-Jun-00	ND
PDU-105-E	Marketing and Sales of Statistical Products and Services	3	CARLSEN	Copenhagen	19-Jun-00	21-Jun-00	MC
ECO-001-F	National Accounts Statistics in Practice	10	LEQUILLER	Paris	19-Jun-00	30-Jun-00	ND
DAT-003-E	Sampling Techniques and Practice	10	SMITH	Southampton	19-Jun-00	30-Jun-00	MC

*** Programme Assistants:**

MC: Martine CORMAN - tel: 352-29858551 - e-mail: mcorman@tes-institute.lu

ND: Nora DERRAS - tel: 352-29858539 - e-mail: nderras@tes-institute.lu

Please note that all the courses offered by the TES Institute are divided between both theoretical and practical sessions. The lectures will provide the theoretical background necessary for further developments and the practical sessions will offer you the opportunity to discuss and exchange experience with colleagues from other countries and other working areas.

The Special Courses Programme

The *Special Courses Programme* is designed for public sector statisticians from countries outside the EU and the EFTA. At this moment its main clients are statisticians from the Central European countries and the countries of the Mediterranean Basin. The *Special Courses Programme* is not subsidised.

The *Special Courses Programme* consists of courses which are either repeats of existing courses in the *Core Programme* or tailor-made courses at the request of a country or a group of countries. Whenever a course from the *Core Programme* is repeated, the content is reviewed and adapted to the specific needs of the participants.

CONSULTING

Apart from the above mentioned training activities, the TES Institute becomes more and more involved in consulting activities. The main objective of setting-up of a training centre for statisticians in any specific country will be reached through the following actions:

- Consulting for curriculum development for their future training programme.
- Training of future trainers of the programme.

PUBLICATIONS

Articles

The TES Institute is regularly publishing articles on its current activities in periodic Newsletter of several Statistical Institutes and some statistical journals as Qüestiió (Quaderns d'Estadística i Investigació Operativa), edited by the Institut d'Estadística de Catalunya.

Manuals

The TES Institute has started the production of TES Manuals on subjects covered by the vocational training programme.

The first manual available at the TES Institute presents *The Role of Statistics in a Democracy.*

The second manual will be published during the summertime and presents the *Indices for bilateral and multilateral Comparison of Prices, Quantities and Values.*

Further manuals will cover topics of *Sampling Techniques* and *Social Statistics.*

OTHER ACTIVITIES

The TES institute has been associated with the Maastricht School of Management as training and advice provider in the framework of their MBA in Decision Support Systems. You can directly contact the TES Institute for any further information on this MBA.

GENERAL INFORMATION

For further details on any of the above mentioned topics, please contact Ms. Valérie Vandewalle, co-ordinator for *Evaluation and Information* matters:

Phone: +352 29 85 85 34
Fax: +352 29 85 29/30
E-mail: vvandewalle@tes-institute.lu

In order to get any information on courses, latest activities, publications, etc., you are always welcome to visit our webside at **www.tes-institute.lu.**

tes

Institute

COURSE «SYSTEMS OF SOCIAL STATISTICS»

COURSE LEADER

Pieter EVERAERS

OBJECTIVE

To get an insight knowledge in the systems and models on social statistics in use in Eurostat and in the Member States of the EU, in the possible data sources, based on the interdependency of the themes in social statistics. Special attention is given on the methodology of harmonization and integration of data sources and results of this processes for the relevant social themes in the national and international context. New developments in harmonisation (e.g. accounting, integration of surveys) and integration techniques and the outcomes are illustrated.

TRAINING METHODS

Lectures, discussion and working groups, practical examples, case studies, exercises, role play.

TARGET GROUP

Statisticians working (or starting to work) in one of the domains in social statistics or in middle management with a task in preparing or implementing harmonisation and integration, working on the crossroads of statistical domains or preparing output/publications covering several data sources and/or statistical domains.

ENTRY QUALIFICATIONS

- Academic or comparable level.
- At least three years of experience in statistical work.
- Sound knowledge of the teaching language (see «Practical Information»).

EXPECTED OUTPUT

Direct usable knowledge of the interdependency between themes in social statistics, methods and levels of harmonisation and integration of data from different sources, insight and experience in working with systems/models of social statistics and the availability and characteristics of European 1/2comparable+ data sources for social statistics.

CONTENTS

Overview of several systems in social statistics, in the process of development or in use, their contents and characteristics, methodological and technical aspects, possible use, available data sources. Relations between the different social themes in statistics, European data sources, policy use, and experiences in NSI's with the systems/models of social statistics.

The course is structured along three main lines:

SYSTEMS OF SOCIAL STATISTICS

- Social accounting and social demographic accounting.
- Supply and demand of services (health, education).
- Situations relating to systems of living/life cycle approaches.
- Time accounting/time use systems.
- The indication approach.

LEVELS AND METHODS FOR HARMONIZATION AND INTEGRATION

- Accounting, macro linkage.
- Imputation, weighting correcting on the table level.
- Combination of information from tables, reconciliation of data.
- Integration on the micro level, micro linkage.

GENERAL BACKGROUNDS/RELEVANCE

- Principles of harmonisation, co-ordination and integration.
- Theoretical bases, methods and techniques, data sources for social statistics.
- The European system of social statistics.
- National experiences.

REQUIRED READING

C.A. van Bochove and P.C.J. Everaers. Micro-macro and micro-micro linkage in social statistics (in Netherlands Official Statistics, 1996).

SUGGESTED READING

Reading of about 50 pages of separate articles required as well as preparing presentation on one or two specific articles. Extra reading of about 100 pages of separate articles suggested.

Description of main characteristics of own national statistical systems. To be brought to the course.

PRACTICAL INFORMATION

Code: SOC-001-E

Duration: 5 days

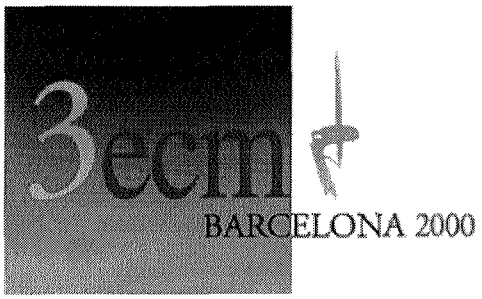
Timing: 4-8 October 1999

Training site: Institut d'Estadística de Catalunya (Idescat)
Via Laietana, 58. 08003 Barcelona. SPAIN

Language: English

Training Fee: 1.240 Euro

Deadline for registration: 31 August 1999



Tercer Congrés Europeu de Matemàtiques

Barcelona, del 10 al 14 de juliol del 2000

Primer Anunci

El Comitè Organitzador es complau a anunciar que el **Tercer Congrés Europeu de Matemàtiques (3ecm)** tindrà lloc a Barcelona del 10 al 14 de juliol de l'any 2000. L'organitza la Societat Catalana de Matemàtiques (SCM), sota els auspicis de la Societat Matemàtica Europea (EMS).

Conferenciants plenaris

- **Robbert Dijkgraaf** (Universitat d'Amsterdam, Holanda)
- **Hans Föllmer** (Universitat Humboldt de Berlín, Alemanya)
- **Hendrik W. Lenstra, Jr.** (Universitat de Califòrnia a Berkeley, Estats Units, i Universitat de Leiden, Holanda)
- **Yuri I. Manin** (Institut Max Planck de Matemàtiques, Bonn, Alemanya)
- **Yves Meyer** (Escola Normal Superior de Cachan, França)
- Carles Simó (Universitat de Barcelona)
- **Marie-France Vignéras** (Universitat de París 7, França)
- **Oleg Viro** (Universitat d'Uppsala, Suècia, i POMI de Sant Petersburg, Rússia)
- **Andrew J. Wiles** (Universitat de Princeton, Estats Units)

Programa científic

El programa del congrés inclourà nou conferències plenàries, trenta conferències invitades en sessions paral·leles, conferències impartides pels guardonats amb els premis de l'EMS, minisimposis, taules rodones i sessions de pòsters. Igual com es va fer en els congressos europeus anteriors, s'atorgarà un cert nombre de premis a investigadors/res joves en matemàtiques, de menys de trenta-dos anys d'edat. Els minisimposis són una de les novetats del **3ecm**; el Comitè Científic escollirà una llista de temes actuals i interdisciplinaris. El programa complet de conferències, minisimposis i taules rodones s'especificarà en el segon anunci. Tots els participants podran presentar comunicacions en forma de pòsters. També està previst que s'organitzin demostracions de programari matemàtic, vídeo i material multimèdia.

Amb finalitats organitzatives, s'ha establert la següent llista numerada de temes científics:

1. Lògica i fonaments
2. Àlgebra. Teoria de nombres
3. Geometria algebraica i analítica
4. Geometria diferencial
5. Topologia
6. Matemàtica discreta i informàtica
7. Modelització i simulació
8. Equacions diferencials ordinàries i sistemes dinàmics
9. Equacions en derivades parcials
10. Anàlisi funcional
11. Anàlisi complexa
12. Probabilitat i estadística
13. Anàlisi real
14. Física matemàtica

Comitès

- El Comitè Científic és presidit per **Sir Michael Atiyah** (Universitat d'Edimburg).
- El Comitè de Premis és presidit per **Jacques-Louis Lions** (Col·legi de França).
- El Comitè de Taules Rodones és presidit per **Miguel de Guzmán** (Universitat Complutense de Madrid).
- El Comitè Organitzador és presidit per **Sebastià Xambó Descamps** (Universitat Politècnica de Catalunya).

Presentació de pòsters

Tots els participants inscrits al **3ecm** podran presentar treballs en forma de pòsters. El Comitè Organitzador decidirà quins pòsters s'accepten a partir de resums que haurà d'haver rebut abans de l'1 d'abril del 2000. Us demanem que envieu el vostre resum preferiblement fent servir el programa que hi haurà a la pàgina web **http://www.iec.es/3ecm/posters.htm**. També es pot enviar per correu electrònic a **posters.3ecm@upc.es**, posant com a *subject* només el número de la secció escaient (vegeu la llista de temes indicada més amunt). Si no el podeu enviar electrònicament, utilitzeu l'adreça següent: *Pòsters 3ecm (Prof. Josep M. Font), Facultat de Matemàtiques, Universitat de Barcelona, Gran Via 585, 08007 Barcelona.*

Presentacions de programari matemàtic

Durant el congrés tindrà lloc una sessió de programari matemàtic, en la qual es podran presentar programes relacionats amb tots els camps de les matemàtiques i aplicables a objectius diversos. El Comitè Organitzador avaluarà les propostes i en seleccionarà un cert nombre, aplicant criteris d'originalitat matemàtica, novetat i possibilitats d'aplicació, i tenint en compte l'equilibri temàtic de la sessió.

Les propostes han d'arribar als organitzadors abans de l'1 de febrer del 2000. Es poden enviar electrònicament, fent servir el full que hi ha a la web **http://www.iec.es/3ecm/mathsoft.htm**, o bé a l'adreça **mathsoft.3ecm@upc.es**, posant com a *subject* només la paraula *mathsoft*. L'adreça següent també es pot utilitzar per enviar material complementari: *Mathsoft 3ecm (Prof. Santiago Zarzuela), Facultat de Matemàtiques, Universitat de Barcelona, Gran Via 585, 08007 Barcelona.*

Presentacions de vídeo i multimèdia

Durant el **3ecm** hi haurà un seguit d'activitats complementàries i actes culturals. Una d'aquestes activitats serà la producció d'un DVD amb vídeos i material multimèdia amb contingut matemàtic. Aquest DVD s'exhibirà en sessions públiques i també serà accessible en diversos llocs de la seu del **3ecm**. Es podran enviar contribucions per a aquest DVD des de totes les àrees de les matemàtiques.

Els treballs hauran d'arribar als organitzadors abans de l'1 de febrer del 2000. S'han d'enviar per correu ordinari a: *Video 3ecm (Prof. Santiago Zarzuela), Facultat de Matemàtiques, Universitat de Barcelona, Gran Via 585, 08007 Barcelona*. Trobareu més informació a la web **http://www.iec.es/3ecm/video.htm**. Podeu utilitzar l'adreça **video.3ecm@upc.es** per contactar amb els organitzadors d'aquesta activitat.

Activitats satèl·lit

Els congressos i les altres activitats de la llista següent han estat acceptats com a satèl·lits del **3ecm** pel Comitè Executiu abans del mes de febrer de 1999. Us encoratgem que feu la llista més llarga. Les propostes s'han de fer arribar al president del Comitè Organitzador abans de l'1 de febrer del 2000, per correu electrònic a **3ecm@iec.es** o bé per carta a la SCM.

- **Summer School on Interactions between Algebraic Topology and Invariant Theory. Ioannina, Grècia, del 26 de juny a l'1 de juliol del 2000.** *Contacteu amb:* Nondas Kechagias (Universitat de Ioannina), **nkechag@cc.uoi.gr.**
- **Functional Analysis Valencia 2000, an International Functional Analysis Meeting on the Occasion of the 70th Birthday of Professor Manuel Valdivia. València, del 3 al 7 de juliol del 2000.** *Contacteu amb:* José Bonet (Universitat de València), **vlc2000@mat.upv.es**, o bé Klaus D. Bierstedt (Universitat de Paderborn), **vlc2000@uni-paderborn.de.**
- **6th International Conference on Harmonic Analysis and Partial Differential Equations. El Escorial, Madrid, del 3 al 7 de juliol del 2000.** *Contacteu amb:* Eugenio Hernández (Universitat Autònoma de Madrid), **eugenio.hernandez@uam.es.**
- **Alhambra 2000, a Joint Mathematical European-Arabic Conference. Granada, del 3 al 7 de juliol del 2000.** *Contacteu amb:* Ceferino Ruiz (Universitat de Granada), **alhambra2000@ugr.es.**
- **First Euro-Mediterranean Topology Meeting. Bellaterra, del 4 al 7 de juliol del 2000.** *Contacteu amb:* Carlos Broto (Universitat Autònoma de Barcelona), **broto@mat.uab.es.**
- **cem 2000, Congrés d'Educació Matemàtica, I Jornades d'Educació Matemàtica a Catalunya. Mataró, del 3 al 5 o del 5 al 7 de juliol del 2000.** *Contacteu amb:* Xavier Vilella (FEEMCAT), **xvilella@pie.xtec.es.**
- **Distributions with Given Marginals and Statistical Modelling. Barcelona, del 17 al 19 de juliol del 2000.** *Contacteu amb:* Carles M. Cuadras (Universitat de Barcelona), **carlesm@porthos.bio.ub.es.**

Preinscripcions

Si desitgeu rebre el segon anunci i més informació per correu electrònic sobre el **3ecm**, us podeu preinscriure a través de la web **http://www.iec.es/3ecm** (si encara no ho heu fet). La preinscripció no costa diners ni us obliga a res. Per tal d'esdevenir participants del **3ecm**, caldrà que formalitzeu la inscripció quan s'obri el termini per fer-ho i pagueu la quota corresponent. També us podeu preinscriure per correu electrònic a l'adreça **3ecm@iec.es**, o bé enviant una carta a la SCM. Cal que indiqueu el vostre nom, la vostra institució, l'adreça postal completa, l'adreça de correu electrònic (si en teniu) i els camps científics que us interessin (en podeu escollir un o més d'un de la llista indicada més amunt).

Patrocinadors (relació actualitzada a febrer de 1999)

Generalitat de Catalunya, Comissionat per a Universitats i Recerca
Generalitat de Catalunya, Departament d'Ensenyament
Ministerio de Educación y Cultura, S.E.U.I.D.
Fundació Catalana per a la Recerca
Ajuntament de Barcelona
Institut d'Estudis Catalans
Universitat de Barcelona
Universitat Autònoma de Barcelona
Universitat Politècnica de Catalunya
Institut d'Estadística de Catalunya
International Mathematical Union
Real Sociedad Matemática Española
Sociedad Española de Matemática Aplicada
Fundación Retevisión
Fundació «la Caixa», Museu de la Ciència
Borsa de Barcelona
Port de Barcelona
Fundació Caixa Catalunya
Fundació Banc Sabadell
Fundació Caixa de Sabadell
Logic Control
Springer-Verlag

Adreces de contacte

Correu electrònic: **3ecm@iec.es**

Web: **http://www.iec.es/3ecm/** o també **http://www.si.upc.es/3ecm/**

Correu ordinari: Societat Catalana de Matemàtiques
Institut d'Estudis Catalans
Carrer del Carme, 47
08001 Barcelona

Telèfon: +34 93 270 16 20

Fax: +34 93 270 11 80

INTERNATIONAL WORKSHOP ON STATISTICAL MODELLING

Bilbao, Spain: Monday 17 to Friday 21 July, 2000

15th IWSM

New Trends in Statistical Modelling

First Announcement and Call for Papers

The International Workshop on Statistical Modelling concentrates on the various aspects of statistical modelling, including theoretical developments, applications and computational methods. Papers motivated by real practical problems are desirable, but theoretical contributions addressing problems of practical importance or related to software developments are also welcome.

The scientific programme is characterized by having invited lectures & tutorials, contributed papers, posters and software demonstrations. Contributed papers should be suitable for a 20 to 30 minutes oral presentation (including discussion) and focus on motivation, statement of key results and conclusions, and emphasize examples, wherever possible.

Invited speakers to date:

Christopher Bishop (Cambridge, UK), Joel L. Horowitz (Iowa City, Iowa, USA), Johannes Ledolter (Vienna, Austria), Winfried Stute (Giessen, Germany), Mark Steel (Edinburgh, UK), James Zidek (Vancouver, British Columbia, Canada), Dale L. Zimmerman (Iowa City, Iowa, USA).

A Tutorial on Goodness of Fit Tests for Regression Models will be given by W. González-Manteiga (Santiago de Compostela, Spain).

Students:

Professors should encourage students to attend the workshop. The programme is designed to allow for discussions and interchange between junior and senior scientists. A special session will be devoted for students contributions, an award for the best presentation will be given.

Scientific programme committee:

Ludwig Fahrmeir (Munich, Germany), Eva Ferreira (Bilbao, Spain, Co-chair), John Hinde (Exeter, U.K., Secretary), Michel Mouchart (Louvain-La-Neuve, Belgium), Vicente Núñez-Antón (Bilbao, Spain, Chair), Jean D. Opsomer (Ames, Iowa, U.S.A.), Juan Romo (Madrid, Spain), Esther Ruíz-Ortega (Madrid, Spain), Bill Venables (Australia).

Local organizing committee:

Ma. Victoria Esteban-González, Eva Ferreira, Petr Mariel, Vicente Núñez-Antón, Jesús Orbe-Lizundia, Susan Orbe-Mandaluniz, Marta Regúlez-Castillo, Juan M. Rodríguez-Póo, Gonzalo Rubio-Irigoyen, Fernando Tusell-Palmer.

Further information:

Details about registration for the workshop, instructions for authors and further information is available from the workshop homepage

http://iwsm.bs.ehu.es

Deadlines:

Jan 31: Submission of abstracts
Mar 13: Notification of acceptance
Apr 17: Submission of final manuscripts.

For additional information please contact:

Vicente Núñez-Antón
Departamento de Econometría y Estadística
Facultad de Ciencias Económicas y Empresariales
Universidad del País Vasco
Avda. Lehendakari Aguirre, 83
48015 Bilbao, Spain
Phone: +34 94 601 37 49
Fax: +34 94 601 37 54
E-mail: vn@alcib.bs.ehu.es

Informació per als autors i lectors

NORMES PER A LA PRESENTACIÓ D'ARTICLES A *QÜESTIIÓ*

La revista accepta, per a la seva publicació, articles originals no sotmesos a consideració en cap altra revista dins els àmbits de l'Estadística, la Investigació Operativa, l'Estadística Oficial i la Biometria. Els articles poden ser teòrics o aplicats, incloent aspectes computacionals i/o de caire docent, i poden presentar-se en anglès, francès, català o qualsevol altra llengua oficial a l'Estat espanyol.

Tots els originals destinats a les esmentades seccions temàtiques de *Qüestiió*, incloent-hi els articles per a la «Secció docent i problemes», seran sotmesos sistemàticament a una avaluació prèvia a càrrec d'especialistes independents i/o membres del Consell Editorial, llevat dels articles convidats per la revista i les reimpressions d'articles. El resultat de l'avaluació serà comunicat a l'autor principal als efectes d'eventuals correccions formals o dels seus continguts.

Per a totes les trameses d'originals, la revista emetrà un acusament de recepció la data del qual figurarà com a «data de rebuda» en la publicació de l'article. Per la seva banda, la «data d'acceptació» de l'article serà la data de recepció de la versió definitiva.

Per a la presentació d'articles, l'autor trametrà a la Secretaria de *Qüestiió* (Institut d'Estadística de Catalunya) dues còpies del treball mecanografiat en DIN A4, a una sola cara, a doble espai i amb marges amplis. Cada article ha d'incloure el títol, el nom de l'autor o autors, la seva afiliació i l'adreça completa, així com un resum de 75-100 paraules al principi de l'article, seguit de les principals paraules clau (en l'idioma original) i la seva adscripció a la classificació AMS. Abans de sotmetre els articles a la revista, s'aconsella els autors que revisin la correcció lingüística de textos d'acord amb l'idioma original i les eventuals traduccions a l'anglès.

Les referències bibliogràfiques es faran indicant el cognom de l'autor seguit de l'any de la publicació entre parèntesi [i.e.: Mahalanobis (1936), Rao (1982b)] i seran llistades alfabèticament al final de l'article; les referències múltiples d'un mateix autor s'ordenaran cronològicament. Les notes explicatives es numeraran correlativament i han d'aparèixer al peu de la pàgina corresponent. Les taules i figures també es numeraran correlativament en el text i seran reproduïdes directament dels originals tramesos en cas que no sigui possible la seva autoedició.

Una vegada avaluat satisfactòriament l'article cal que, a més de la versió impresa, l'autor el trameti en disquet de 3.5 polsades i en format MS-DOS, on han de constar de forma clara els noms dels autors i el títol de l'article. Aquesta versió final s'ha de trametre preferiblement en el processador de textos LaTeX2ε [subsidiàriament, es poden trametre els textos i les taules en Word Perfect —versió 6.0A o anterior— o ASCII]; en el cas de figures, diagrames o gràfics es recomanen els formats adients per als programes-editors PS, EPS o PCX. Els autors han de garantir la correspondència exacta entre la versió impresa i la còpia electrònica. D'altra banda, si l'article no està escrit en llengua anglesa s'haurà d'adjuntar la traducció del títol original, de l'abstract i de les paraules clau, així com un ampli resum en anglès (amb una extensió d'entre 2 i 5 pàgines i amb la mateixa estructura de l'article original).

La Secretaria de *Qüestiió* posa a disposició dels autors que ho sol·licitin plantilles en format LaTeX2ε de *Qüestiió* per a la seva edició i les referències adients de la classificació AMS.

QÜESTIIÓ
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona
Tel: +34-93 412 15 36
Fax: +34-93 412 31 45
E-mail: questiio@idescat.es

GUIDELINES FOR THE SUBMISSION OF ARTICLES FOR *QÜESTIIÓ*

The journal well comes submission of articles and contributions that are not being considered for publication in any other journal in the fields of Statistics, Operational Research, Official Statistics or Biometrics. Articles may be theoretical or applied, including teaching aspects and applications, and will be accepted in English, French, Catalan or any of the other official languages in Spain.

All originals assigned to the thematic sections of *Qüestiió*, including articles for the «Teaching section and problems» will be systematically reviewed by independent referees and/or members of the Editorial Board, who will send a report to the main author of the article in order to correct, if necessary, any formal or content aspects. The articles invited by the journal and articles reprinted will be excluded from this evaluation process.

For all submissions, the journal will issue a receipt corresponding to the submission date, which will appear as «date received» in the final publication of the article. The «acceptance date» of the article, which will appear in its final publication, will be the date of sending the final version to the journal.

For the presentation of original articles, the author should send, to the Secretary of *Qüestiió* (Institut d'Estadística de Catalunya), two copies of the paper typed on A4 sheets, one side of the paper only, double spaced and with wide margins. Each article should include the title, the name of the author or authors, their affiliation, full address and also an abstract of the paper (75-100 words) at the beginning of the article, followed by the main keywords (in the original language) and its assignation in the AMS classification. Before submitting their papers, authors are advised to seek assistance in the writing of their articles for the correct use of English and/or of original language.

Bibliographical references should state the author's name followed by the year of publication in brackets [e.g.: Mahalanobis (1936), Rao (1982b)] and they should be listed at the end of the article in alphabetical order; multiple references to the same author should be given in chronological order. Footnotes should be numbered in the article and appear at the foot of the corresponding page. Figures and tables are to be numbered in consecutive order in the text using Arabic numerals and will be directly reproduced from the originals submitted if it is not impossible to print them electronically.

Once the evaluation has been passed, the author is required to provide the article on a diskette (a 3.5-inch disk in MS-DOS format) together with its paper copy; it must be a new diskette and must bear very clearly the names of the authors and the title of the article. This final version should be processed by LaTeX 2_ε, preferably, or, failing that, by Word Perfect (6.0A or earlier) or ASCII for text and tables; for figures, diagrams or graphs, the appropriate formats of PS, EPS or PCX software tools are strongly recommended. Authors must ensure that the version of the electronic copy is exactly the same as the paper copy which accompanies it. Furthermore, if the article is not written in English, the translation of its original title, short abstract and keywords should be enclosed, as well as a full summary of the article in English (that is, 2-5 pages with the same structure as the original).

The Secretary of *Qüestiió* can send, by request of the authors, the LaTeX 2_ε style of *Qüestiió* for manuscript preparation and the appropriate AMS classification references.

QÜESTIIÓ
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona

Tel: +34-93 412 15 36
Fax: +34-93 412 31 45
E-mail: questiio@idescat.es

NORMES PER A LA PUBLICACIÓ D'ANUNCIS INSTITUCIONALS A *QÜESTIIÓ*

Qüestiió convida les entitats patrocinadores, les institucions col·laboradores, els organismes públics i privats, i tota la comunitat científica vinculada a l'estadística o la investigació operativa, a la publicació d'anuncis institucionals sobre cursos, seminaris, congressos i activitats similars que, preferentment, tinguin lloc en el nostre país. Els textos poden presentar-se en anglès, francès, català o en qualsevol altra llengua oficial a l'Estat espanyol. Les iniciatives per a una possible publicació sempre són a instància de les entitats interessades, de manera que *Qüestiió* no fa una cerca sistemàtica d'esdeveniments d'aquesta naturalesa, ni té cap ànim d'exhaustivitat en les ressenyes d'activitats finalment publicades.

Una vegada aprovada la inclusió dels anuncis sol·licitats es procedirà a la seva publicació, i es reproduirà directament dels originals tramesos amb les mides adequades i la màxima qualitat tipogràfica possible; en aquest cas, *Qüestiió* no procedeix a cap mena de procés d'autoedició de la versió impresa que l'anunciant hagi tramès. Si els originals es trameten en els mateixos termes electrònics exigits per als articles (vegeu «Normes per a la presentació d'articles a *Qüestiió*»), la revista procedirà a la seva autoedició. Si es desitja una qualitat superior a la reproducció simple o l'autoedició, o bé la seva publicació en color, els sol·licitants hauran de posar-se en contacte amb la Secretaria de *Qüestiió* per tal de trametre els fotolits dels textos originals corresponents.

La disposició dels textos i les figures adjuntes dels anuncis han de procurar la màxima intel·ligibilitat i claredat expositiva, sense atapeir la informació ni utilitzar formats o fonts de lletres excessivament petites. D'altra banda, la publicitat ha de ser fidedigna, exempta d'enganys i respectuosa amb les persones i institucions. En qualsevol cas, la direcció de *Qüestiió* es reserva la decisió final pel que fa a la seva publicació.

L'anunciant es compromet a lliurar els textos/materials amb l'antelació que se li indiqui per a la inserció en els números/volums de *Qüestiió* que prèviament s'hagi establert. La revista no es fa responsable dels retards, per part de l'anunciant, que impedeixin la publicació de l'anunci en els termes previstos.

Mònica M. Jaime
Secretaria de *Qüestiió*
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona
Tel: +34-93 412 15 36
Fax: +34-93 412 31 45
E-mail: questiio@idescat.es

GUIDELINES FOR INSTITUTIONAL ADVERTISEMENTS IN *QÜESTIIÓ*

Qüestiió invites all sponsor entities, collaborating institutions, other public and private bodies and the entire scientific community related to Statistics or Operations Research to submit institutional advertisements on courses, seminars, congress and similar activities that will be held, preferably in our country. These will be accepted in English, French, Catalan or any of the other official languages in Spain. The initiative should always come from the entities interested in advertising them so that *Qüestiió*'s aim is not to do a systematic search of these events and therefore does not publish a comprehensive list of such activities.

Once their insertion is approved the advertisements will be reproduced from the most accurate photocopy of the originals sent by the advertiser to *Qüestiió* in paper copy, with the appropriate size and at the best possible typographic quality. Therefore, in this case the journal does not elaborate any further editing process to the printed version that the advertiser has sent. If the original advertisements are sent in the same electronic format requested by the articles (please see «Guidelines for the submission of articles for *Qüestiió*») the journal will print it directly from the file. If a better quality than the simple reproduction or automatic printing or a colour version of the adverts is desired, the authors should contact the Secretary of *Qüestiió* in order to negotiate this.

The typesetting of texts and figures in the advertisement should have maximum intelligibility and clearness, neither compressing the information too much nor using formats or letter fonts that are too small. Furthermore, the information has to be reliable, without errors and respectful of the people and institutions. The management of *Qüestiió* has the right to a final decision concerning the insertion of the advertisement.

Advertisers commit themselves to give the text/materials on request in order to insert them in the issues of *Qüestiió* that have been previously agreed. The journal is not responsible for any delay from the announcer that could prevent the advertisement from been published on the agreed terms.

Mònica M. Jaime
Secretaria de *Qüestiió*
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona
Tel: +34-93 412 15 36
Fax: +34-93 412 31 45
E-mail: questiio@idescat.es

NORMES PER A LA PUBLICACIÓ D'ANUNCIS PRIVATS O AMB FINALITAT COMERCIAL A *QÜESTIIÓ*

Qüestiió accepta la publicació d'anuncis privats o amb finalitat comercial sobre productes, serveis o altres eines promocionals a l'entorn de l'estadística o la investigació operativa. Els textos poden presentar-se en anglès, francès, català o en qualsevol altra llengua oficial a l'Estat espanyol. Les iniciatives per a una possible publicació sempre són a instància de les organitzacions que hi estiguin interessades, de manera que *Qüestiió* no fa una cerca sistemàtica de novetats o productes d'aquesta naturalesa ni té cap ànim d'exhaustivitat en els anuncis finalment publicats.

Els anuncis en **blanc i negre** s'elaboren a partir de la fotocòpia més acurada possible dels originals que trameti l'anunciant en versió impresa, amb les mides adequades i la màxima qualitat tipogràfica. Per tant, en aquest cas la revista no efectua cap procés d'edició ulterior respecte de la versió impresa que l'anunciant hagi tramès. Alternativament, si els anuncis originals es trameten en els mateixos termes formals exigits per als articles (vegeu «Normes per a la presentació d'articles a *Qüestiió*»), la revista procedirà a la seva autoedició. Igualment, si es desitja una qualitat superior a la reproducció simple, els sol·licitants hauran de trametre els fotolits dels originals corresponents o encarregar-los a *Qüestiió,* que els facturarà separadament.

Els anuncis **en color** requereixen els fotolits dels textos originals, que poden ser subministrats directament per l'anunciant o bé encarregats per la revista a compte de l'anunciant; en el segon cas, l'anunciant ha de trametre a la revista els originals impresos en color amb la màxima qualitat, per tal de filmar-los amb les millors garanties i condicions. El cost dels fotolits realitzats per *Qüestiió* serà sempre a càrrec de l'anunciant, a qui se li repercutirà l'import i l'IVA d'aquests, juntament amb les tarifes que corresponen a la modalitat d'anunci per la qual hagi optat.

La disposició dels textos i figures adjuntes dels anuncis ha de procurar la màxima intel·ligibilitat i claredat expositiva, sense atapeir la informació ni utilitzar formats o fonts de lletres excessivament petites. D'altra banda, la publicitat ha de ser fidedigna, exempta d'enganys i respectuosa amb les persones i institucions. En qualsevol cas, la direcció de *Qüestiió* es reserva la decisió final de la seva inclusió.

L'anunciant es compromet a lliurar els textos/materials amb l'antelació que se li indiqui per a la seva inserció en el(s) número(s)/volum(s) de *Qüestiió* que prèviament s'hagi establert. La revista no es fa responsable dels retards per part de l'anunciant que impedeixin la publicació de l'anunci en els termes previstos.

Imports:

1 pàgina en color (un número aïllat):	125.000 PTA + IVA
1 pàgina en color (tres números consecutius):	200.000 PTA + IVA
1 pàgina en blanc i negre (un número aïllat):	30.000 PTA + IVA
1 pàgina en blanc i negre (tres números consecutius):	50.000 PTA + IVA
1/2 pàgina en blanc i negre (un número aïllat):	20.000 PTA + IVA
1/2 pàgina en blanc i negre (tres números consecutius):	35.000 PTA + IVA

Mides opcionals dels anuncis:

1 pàgina sencera (espai intern):	19.0 cm. × 12.3 cm.
1 pàgina sencera (espai extern):	23.8 cm. × 17.0 cm.
1/2 pàgina (espai intern):	9.5 cm. × 12.3 cm.
1/2 pàgina (espai extern):	11.9 cm. × 17.0 cm.

Forma de Pagament:

- Transferència bancària al compte: 2013-0100-53-0200698577
- Xec bancari nominatiu a l'Institut d'Estadística de Catalunya
- Pagament amb targeta de crèdit

El pagament serà per l'import total de la factura corresponent, on hi figurarà el cost dels fotolits en el cas que l'edició de l'anunci hagi estat a càrrec de l'Institut. En el cas que l'anunciant necessiti una factura proforma, només cal que ho faci saber amb l'antelació suficient.

Correspondència:

Mònica M. Jaime
Secretaria de *Qüestiió*
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona

Tel: +34-93 412 15 36
Fax: +34-93 412 31 45
E-mail: questiio@idescat.es

GUIDELINES FOR THE PRIVATE OR COMMERCIAL ADVERTISEMENTS IN *QÜESTIIÓ*

Qüestiió accepts for their publication both private and commercial advertisements on products, services or other promotional tools related to statistics or operational research and will be accepted in English, French, Catalan or any of the official languages in Spain. The initiatives should always come from entities interested in advertising them so that *Qüestiió*'s aim is not to do a systematic search of news and therefore does not publish a comprehensive list of such private or profit activities.

The **black and white** advertisements are made out from the most accurate photocopy of the originals sent by the advertiser to *Qüestiió* in paper copy with the appropriate size and at the best possible typographic quality. Therefore, in this case the journal does not elaborate any further editorial process to the printed version that the advertiser has sent. Alternatively, if the original advertisements are sent in the same formal terms required by the articles (please see «Guidelines for the submission of articles for *Qüestiió*»), the journal will proceed to its autoedition. In the same way, if a better quality than the simple reproduction is wanted, the authors should send the photolits of the corresponding original texts or, on the other hand, order to *Qüestiió* their fulfilment, which will be invoiced separately from the rates charged as advertisements.

The advertisements **in colour** need the photolits of the original texts, which can be provided directly by the advertiser or requested by *Qüestiió* to the advertiser charge; in the second case, the advertiser must sent to the journal the originals printed in colour with the best possible quality, so that they can be filmed at the best conditions and guarantees. The cost of the photolits made by *Qüestiió* will always be charged to the advertiser together with the VAT derived from it, plus the prices corresponding to the type of the advertisement that has been chosen.

The set up of texts and figures of the advertisement should provide the maximum intelligibility and clearness, neither squeezing together the information nor using set ups or letter types that are too small. On the other hand the publicity has to be reliable, without fraud and respectful to the persons and institutions. The direction of *Qüestiió* has the right of the last decision concerning the insertion of the advertisement.

The advertiser committs himself to give the texts/materials on request, in order to insert them in the issue(s) of *Qüestiió* that had been previously agreed. The journal is not responsible for any delay from the announcer that could prevent the advertisement from been published in the agreed terms.

Rates:

1 colour page (only one issue):	125.000 PTA + VAT
1 colour page (three consecutive issues):	200.000 PTA + VAT
1 black and white page (only one issue):	30.000 PTA + VAT
1 black and white page (three consecutive issues):	50.000 PTA + VAT
1/2 black and white page (only one issue):	20.000 PTA + VAT
1/2 black and white page (three consecutive issues):	35.000 PTA + VAT

Advertisement sizes (optional):

1 full page (internal space):	19.0 cm. × 12.3 cm.
1 full page (external space):	23.8 cm. × 17.0 cm.
1/2 page (internal space):	9.5 cm. × 12.3 cm.
1/2 page (external space):	11.9 cm. × 17.0 cm.

Payment:

- A bank transfer to account number: 2013-0100-53-0200698577
- A bank cheque to Institut d'Estadística de Catalunya
- Charge on a credit card

The payment should be for the amount shown at the invoice, where it will be shouwn the total cost of the photolits, in case that *Qüestiió* would be in charge of the filmation of the advertisement. If advertiser need a pro-forma invoice, he should let us know some time in advance so that *Qüestiió* could send it to the proper address.

Mail address:

Mònica M. Jaime
Secretaria de *Qüestiió*
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona

Tel: +34-93 412 15 36
Fax: +34-93 412 31 45
E-mail: questiio@idescat.es

Butlleta de subscripció a la revista ***Qüestiió***

Nom i cognoms _____

_____

Empresa/Institució _____

_____

Adreça _____

_____

Codi postal _____ Ciutat _____

Tel. _____ Fax _____ NIF _____

Data _____

Signatura

Desitjo subscriure'm a ***Qüestiió*** per a l'any 1999.
El preu de la subscripció és de 3.000 PTA (18,03 euros) (IVA inclòs).

Forma de pagament

☐ Transferència al compte 2013-0100-53-0200698577

☐ Domiciliació bancària al compte número

☐☐☐☐ ☐☐☐☐ ☐☐ ☐☐☐☐☐☐☐☐☐☐

☐ Xec nominatiu a l'Institut d'Estadística de Catalunya

☐ Gir postal

☐ En efectiu

Retorneu aquesta butlleta (o una fotocòpia) a:

Qüestiió:
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona

Exemplar per a l'entitat bancària

Autorització de domiciliació bancària per al pagament de les subscripcions anuals de la revista ***Qüestiió***

El sotasignat _____

autoritza el Banc/Caixa _____

Adreça _____

Codi postal _____ Ciutat _____

a abonar les subscripcions a la revista ***Qüestiió*** amb càrrec al seu compte

número ☐☐☐☐ ☐☐☐☐ ☐☐ ☐☐☐☐☐☐☐☐☐☐

Data _____

Signatura

Qüestiió:
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona