Qüestiió

Quaderns d'Estadística i Investigació Operativa

Any 1998, volum 22, núm. 2
Segona època

Entitats patrocinadores:

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

Entitat col·laboradora:

International Biometric Society

Generalitat de Catalunya
Institut d'Estadística de Catalunya

Any 1998, volum 22, núm. 2

SUMARI

EDITORIAL

Aquest segon número del volum 22 (1998) inclou set articles, amb la presència d'originals a totes les seccions temàtiques pròpies de *Qüestiió*. D'aquesta manera, només amb l'edició dels dos primers números del 1998 ja s'han publicat setze articles i més de 380 pàgines, xifres equivalents al que aplegava un volum complet de la revista a la seva primera època, amb l'expectativa de superar enguany els promitjos enregistrats durant els darrers cinc anys de *Qüestiió*.

En segon lloc, cal esmentar que a partir del present número *Qüestiió* inclou un darrer apartat que, sota el títol «Informació per als autors i lectors», sistematitza i actualitza diverses qüestions pràctiques com són les normes per a la presentació d'articles, els criteris per a la publicació d'anuncis institucionals o els termes de la subscripció ordinària/domiciliació bancària, que ja es publicaven anteriorment de forma dispersa i menys detallada. A continuació d'aquest apartat, es preveu l'eventual inserció d'anuncis de caràcter privat o amb finalitats comercials sobre la promoció de productes i serveis a l'entorn de l'estadística o la investigació operativa que siguin d'interès per als lectors de *Qüestiió*; en aquest sentit, l'apartat esmentat informa del catàleg d'imports i normes editorials per a la seva publicació.

Per últim, i tal com s'en feia ressò l'editorial del número anterior, la recent posada en marxa del web de l'Institut d'Estadística de Catalunya (Idescat) permet accedir a una informació àmplia i detallada de la producció científica, de l'entorn institucional i de les novetats editorials de *Qüestiió*, incloent-hi la consulta sobre l'estat de situació dels articles que es troben sotmesos al procés d'avaluació. La nova adreça de les pàgines directament referides a *Qüestiió* és *http://www.idescat.es/publicacions/questiio/questiio.stm*. Igualment, cal tenir en compte l'existència d'una adreça de correu electrònic (questiio@idescat.es) per atendre les trameses d'originals, sol·licituds de subscripcions i, en general, la correspondència amb la secretaria de la revista. A partir d'aquest número, ambdues adreces figuraran impreses a les cobertes interiors de *Qüestiió*, on s'hi reflectirà qualsevol novetat en aquest àmbit.

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

La secció «Estadística» conté quatre originals. El primer, «On maximum entropy priors and a most likely likelihood in auditing», d'A. Hernández, M.C. Martel i F.J. Vázquez, proposa un enfocament baiesià a problemes d'auditoria, escollint una distribució prior de màxima entropia i calculant la distribució posterior mitjançant les corbes posteriors més versemblants. L'article «On some strategies using auxiliary information for estimating finite population mean», de L.N. Sahoo, J. Sahoo i M. Ruíz

Espejo, és un estudi empíric de cinc estratègies destinades a estimar la mitjana d'una població finita, que fan ús de diversos paràmetres d'una variable auxiliar, tot estudiant la seva eficàcia en relació a criteris tals com el biaix, l'eficiència i l'assimetria. A continuació, l'article «Reflexiones sobre la estrategia de medida de los cambios en probabilidad en modelos de elección binarios», de M.T. Aparicio i I. Villanúa, estudia com varia la probabilitat de que una unitat prengui una decisió quan varia alguna de les variables explicatives del model; les autores proposen una forma d'avaluació de la variació, alternativa a la clàssica, i fan una simulació per provar què és millor. Finalment, l'article «Estimación del número de clusters en una población aplicando el jackknife generalizado», de J.J. Prieto, planteja una tècnica no paramètrica per estimar el nombre de clusters que no necessita un model probabilístic, ni tampoc recórrer al concepte de cobriment mostral, que es basa en el jackknife generalitzat i que s'il·lustra amb dades reals i amb simulacions.

L'únic article de la secció «Investigació Operativa», «Un método primal de optimización semi-infinita para la aproximación uniforme de funciones», de T. León, S. Sanmatías i E. Vercher, presenta un algorisme per aproximar una funció mitjançant una combinació lineal de funcions conegudes, aproximació que desenvolupen a partir de programes semi-infinits lineals i que els autors comparen amb altres algorismes a fi de comprovar-ne l'eficàcia resultant.

La secció «Biometria» presenta també només un article, «Analysis of survey data investigating the Malarial endemicity of a mixed tribal population of Bihar, India», de T.K. Basu, S.K. Sarkar i S. Ganguly; el seu treball conté un interessant exemple mèdic que compara dos estats en el sistema de predicció de la malària, cercant paràmetres i variables categòriques per tècniques Anova i d'anàlisi discriminant, a fi de determinar factors vitals predictius.

Carles M. Cuadras, editor executiu

Comentari de la secció «Estadística Oficial» i altres apartats

La secció «Estadística Oficial» únicament inclou l'article «Mètode per a l'estimació de migracions en àrees petites a partir de dades incompletes», de J.A. Sánchez Cepeda, el qual il·lustra els recents treballs que ha desenvolupat l'Institut d'Estadística de Catalunya en l'àmbit de les estimacions en petites àrees, referits en aquest cas a l'aplicació de l'anàlisi cluster a l'estadística de fluxos demogràfics. El principal interès del treball rau en la modelització del fenòmen migratori segons la corba de Rogers, oferint un ajustament satisfactori a les dades i una estimació raonable de la migració per edats, informació escassament disponible en la mateixa estadística oficial.

A continuació, la «Secció docent i problemes» prossegueix amb la presentació successiva de nous enunciats i la resolució dels problemes publicats en el número immediatament anterior al present.

Finalment, l'apartat dedicat a «Cursos i congressos d'estadística» amplia progressivament la diversitat i el nombre d'activitats incloses, per la qual cosa la seva actual denominació serà substituïda properament per un títol més genèric que faci referència a les ressenyes d'activitats institucionals en l'estadística o la investigació operativa. En aquest sentit, aquest número recull, en primer lloc, una presentació actualitzada de la Sociedad Española de Biometría, amb l'anunci dels cursos que organitzarà properament i de la convocatòria de la «VII Conferencia Española de Biometría» que se celebrarà l'any vinent (Palma de Mallorca, 10-12 de març de 1999). A continuació, es publica la presentació actualitzada del TES Institute-Training for European Statisticians que *Qüestiió* difon des del volum 21 (1997), juntament amb la ressenya dels cursos inclosos al «Training Programme 1998-99» que s'impartiran fins el desembre de 1998 en el camp de l'estadística oficial d'àmbit comunitari. Seguidament, s'anuncien les jornades internacionals sobre «Generació d'informació estadística: qualitat i limitacions» (Barcelona, 30 de novembre-1 de desembre de 1998) organitzades per la Xarxa Temàtica «Enquestes i qualitat de la informació estadística», amb la col·laboració, entre d'altres institucions, de les tres entitats patrocinadores de *Qüestiió* que participen igualment en les activitats de la Xarxa des de la seva recent constitució a finals de 1997. En darrer lloc, s'anuncia la «International Conference on Combinatorics, Statistics, Pattern Recognition and Related Areas» (Mysore, India, 28-30 de desembre de 1998), constitutiva de la «Fifth International Conference of Forum for Interdisciplinary Mathematics».

Enric Ripoll, editor executiu

Estadística

QÜESTIIÓ, vol. 22, 2, p. 231-242, 1998

ON MAXIMUM ENTROPY PRIORS AND A MOST LIKELY LIKELIHOOD IN AUDITING*

A. HERNÁNDEZ BASTIDA*
Universidad de Granada

M.C. MARTEL ESCOBAR*
F.J. VÁZQUEZ POLO*
Universidad de Las Palmas de G.C.

There are two basic questions auditors and accountants must consider when developing testing and estimation applications using Bayes' Theorem: What prior probability function should be used and what likelihood function should be used. In this paper we propose to use a maximum entropy prior probability function MEP with the most likely likelihood function MLL in the Quasi-Bayesian QB model introduced by McCray (1984). It is defined on a adequate parameter. Thus procedure only needs an expected value of θ_0 *known (in this paper the expected tainting) to obtain a MEP all an auditor or accountant need to suply are the range, as with any other prior, and the expected tainting,* θ_0*. We will see some practical applications of the methodology proposed about internal control evaluation in auditing.*

Keywords: Maximum Entropy Prior, Partial Prior Information, Auditing.

AMS Classification: 62A15, 62C10

---

* A draft version of this paper was presented by authors at 4th World Congress of the Bernoulli Society. Vienna, August, 1996. We would like to thank, without implicating, Nicanor Guerra, for computation assistance.

⋆ Authors in alphabetical order. All three authors have made major contributions to this work.
Research partially supported by «Dirección General de Universidades e Investigación del Gobierno Autónomo de Canarias», under grant 2705. *Address for correspondance*: Departamento de Métodos Cuantitativos en Economía y Gestión. Facultad de CC. Económicas. Módulo D. 35017 Las Palmas de G.C. *e-mail*: polo@empresariales.ulpgc.es

– Received September 1996.
– Accepted June de 1997.

1. INTRODUCTION

Bayes' Theorem has not been widely used in auditing and accounting because accountants find some dificulties in assessing a prior probability density function about the average tainting (or total error in an account or file). There are several reasons for these inconveniences. Firstly, it is difficult for accounting firms to define standard specific procedures to obtain the prior distribution using auditing evidence accumulated in the normal course of the evaluation of internal accounting control. Therefore it may be difficult for the auditor to justify the parameters in the prior probability density function to be used since they have not an intuitive meaning. Procedures involved may be also quite complicated and cumbersome to implement without significant staff training.

This paper suggests the usage of the class of prior probability density functions known as maximum entropy priors MEP's because they do not have any of the problems mentioned above such a class of distributions includes not only those aspects about the prior density which are unquestionable, but they are also the most noninformative priors available.

Moreover, this is a very simple procedure to use. For instance, if an auditing firm is about using MEP in internal accounting control evaluation and analytical reviews, two steps should be considered:

The former one includes the specification by the auditing firm of a matrix relating any of the possible evaluations to one or more of the statistical measures considered to be relevant in the prior distribution, i.e. the mean value, one or two quantile values. This matrix will be specific for this particular auditing firm and will be reflecting its own criteria.

In the second step, once that matrix has been established, the MEP distribution is completely defined and it can be used in bayesian analysis for DUS (Dollar Unit Sampling) or phisical unit sampling.

This paper introduces a new method to specify the posterior distribution that uses the MEP. This method is called MLPC (most likely posterior curves) and does not need to know the probability distribution for the tainting in the population.

The remainder of this article is organized as follows: Section 2 describes maximum entropy approach used later. Section 3 shows how maximum entropy priors and most likely likelihood have been used to obtain the posterior distribution of total amount of error in an accounting population. By illustrations, Section 4 describes how a firm could establish a relationship between the evaluation of internal accounting control and analytical review and the MEP characteristics. Finally, Section 5 contains a summary and some concluding remarks.

2. MAXIMUM ENTROPY PRIORS

Let the parameter space Θ be a continuous and bounded subset of the real line. In practice Θ will be a closed and bounded interval of the real line, $\Theta = [a,b]$, and represents the taint in an accounting population.

Even, if the parameter space is continuous and unbounded there is not a natural definition of entropy. We will use the definition proposed by *Jaynes (1968)*, for a probability distribution π as follows:

$$(1) \qquad Ent(\pi) = -E\left\{\log\frac{\pi(\theta)}{\pi_0(\theta)}\right\} = -\int_\Theta \pi(\theta)\log\left(\frac{\pi(\theta)}{\pi_0(\theta)}\right)\cdot d\theta$$

where π_0 is the natural «invariant» noninformative prior for the problem, usually we use the natural noninformative prior uniform.

It is well known (see *Berger (1985)*, pp. 92-93) that if partial prior information is given by:

$$(2) \qquad E^\pi[g_k(\theta)] = \int_\Theta g_k(\theta)\cdot\pi(\theta)\cdot d\theta = \mu_k,\ k = 1,...,m,$$

and we try to solve:

$$\max\ Ent(\pi)$$

subject to:

$$E^\pi[g_k(\theta)] = \int_\Theta g_k(\theta)\cdot\pi(\theta)\cdot d\theta = \mu_k,\ k = 1,...,m.$$

then, the solution is given by the expression:

$$(3) \qquad \pi(\theta) \propto \pi_0(\theta)\cdot exp\left\{\sum_k \lambda_k\cdot g_k(\theta)\right\}$$

where λ_k $(k = 1,2,...,m)$ are constants to be determined from the constraints in (2).

Notice that:

- If $g_1(\theta) = \theta$ and, $g_k(\theta) = (\theta-\mu_1)^k$, $2 \le k \le m$, restrictions and hence partial information consists of specifying m central moments in the distribution,
- If $g_k(\theta) = I_{(-\infty,z_k]}(\theta)$, restrictions are now refered to the specification of m quantiles.

There are some interesting applications in statistical auditing involving the bayesian approach.

2.1. Partial Information given by the Mean

The maximum entropy prior for a location parameter specified by the mean θ_0 is given by

$$\pi(\theta) = \frac{\lambda e^{\lambda\theta}}{e^{\lambda b} - e^{\lambda a}} \tag{4}$$

where the unique restriction is $g_1(\theta) = \theta$, $\mu_1 = \theta_0$, and a and b are specified (in the domain of θ) and λ is obtained by solving the nonlinear equation

$$\frac{\lambda(ae^{-\lambda a} - be^{-\lambda b}) + e^{-\lambda a} - e^{-\lambda b}}{\lambda(e^{\lambda b} - e^{\lambda a})} = \theta_0 \tag{5}$$

if θ_0 is less than $(a+b)/2$, then λ is positive, λ is negative otherwise. If $\theta_0 = (a+b)/2$, equation (5) cannot be solved, but it can be shown that in the limiting shape of the MEP is that of the uniform prior between a and b. Equation (5) is not restricted to positive values for a and b as constraints above might suggest. If a is negative, it is a simple matter to perform a linear transformation to obtain a MEP satisfying the boundary constraints in equation (5).

2.2. Partial Information when one quantile is given

In this case the restriction is $g_1(\theta) = I_{(-\infty,z_1]}(\theta)$, where z_1 is the known α-quantile ($\alpha \in (0,1)$). We can obtain the maximum entropy prior through by easy algebraical manipulations:

$$\pi(\theta) = \begin{cases} \frac{e^{\kappa}}{e^{\kappa}(z_1-a)+(b-z_1)} & , \text{ if } \quad a \le \theta \le z_1 \\ \frac{1}{e^{\kappa}(z_1-a)+(b-z_1)} & , \text{ if } \quad z_1 < \theta \le b \end{cases} \tag{6}$$

where κ is a constant to be determined from the constraint $\alpha = \int_a^{z_1} \pi(\theta)d\theta$. Easy computation gives us: $\kappa = \log((b-z_1)\alpha/(1-\alpha)(z_1-a))$. Finally the MEP obtained from one quantile given is the two piece uniform distribution,

$$\pi(\theta) = \begin{cases} \frac{\alpha}{z_1-a} & , \text{ if } \quad a \le \theta \le z_1 \\ \frac{1-\alpha}{b-z_1} & , \text{ if } \quad z_1 < \theta \le b \end{cases} \tag{7}$$

2.3. Partial Information when two quantiles are given

Now the restrictions are $g_1(\theta) = I_{(-\infty,z_1]}(\theta)$; $g_2(\theta) = I_{(-\infty,z_2]}(\theta)$, where z_1 is the known α_1-quantile and z_2 is the known α_2-quantile ($\alpha_1, \alpha_2 \in (0,1)$), $\alpha_1 < \alpha_2$, $z_1 < z_2$, then

$$\pi(\theta) \propto \exp\{\kappa_1 g_1(\theta) + \kappa_2 g_2(\theta)\}$$

where κ_1, κ_2 are constants to be determined from the constraints,

$$\alpha_1 = \int_a^{z_1} \pi(\theta) d\theta, \text{ and } \alpha_2 = \int_a^{z_2} \pi(\theta) d\theta.$$

Calculations involved can be easily performed obtaining the three piece uniform distribution MEP given by,

$$\pi(\theta) = \begin{cases} \frac{\alpha_1}{z_1 - a} & , \text{ if } a \leq \theta \leq z_1 \\ \frac{\alpha_2 - \alpha_1}{z_2 - z_1} & , \text{ if } z_1 < \theta \leq z_2 \\ \frac{1-\alpha_2}{b - z_2} & , \text{ if } z_2 < \theta \leq b \end{cases} \tag{8}$$

Obviously, if we consider three o more quantile the procedure yields four or more piece uniform distributions MEP. The three cases considered above are specially atractive to use in auditing and accounting settings as we will show in the next section. Furthermore, it should be used with the most likely likelihood in the Quasi-Bayesian Model (*McCray (1984,1986)*).

3. MEP AND MLPC IN AUDITING

A magnitude of prime interest in accounting auditing is the total amount of error since it has an intuitive meaning for auditors and it is also something those have prior relevant information about. Suppose a range of equally spaced possible total amount of error is defined (for example 500 or 1000).

There are various sampling techniques in auditing. Dollar unit sampling (DUS) maybe particularly appealing to auditors. Roughly speaking, this is a method to select sample items such that the probability of any given item being selected is directly proportional to its record value in the book. If the auditor is interested in including zero and/or small recorded value (wich have a small chance of being selected), then he/she could design specific audit test about them.

Suppose DUS sampling is used (see *Felix and Grimlund (1977)*, *Cox and Snell (1979)*, and *Godfrey and Neter (1982)*, among others) and we are interested in combining our

sample observations to prior information to get a posterior probability distribution over all possible states of nature, the possible total amounts of error previously specified.

In DUS, the population size is the known Recorded Book Value (*RBV*) and the sample plan consists of selecting dollar units with equal chance of being selected. The amount of error for each dollar selected is the difference between its two associated values: its book value and its audited value (presumed to be correct). The fraction (error / book value) is called **taint** of the dollar-unit randomly selected. Taintings in a dollar unit sample are recorded and used to make inference about the total error amount in the population. In a empirical situation, most of these tainting values are zero. We assume that no amount can be overestimated or underestimated by a quantity bigger than its book value, therefore the variation range of taintings goes from -100 to +100 per cent.

We have then 201 categories of taintings: $T_{-100},...,T_{-1},T_0,T_{+1},...,T_{+100}$, associated to different taintings: $-100\%,...,+100\%$. When the error tainting in the sample is zero, the sample dollar is counted in category T_0, and when it is in between the category T_{i-1} and T_i it is counted in category T_i. Let θ_i $(i=-100,...,+100)$ denote the population proportion of dollar-units with i percent error. For a random sample of dollar units of size n, let n_i $(i=-100,...,+100)$ be the observed frequencies in category T_i $(i=-100,...,+100)$. The counts in categories follow a multinomial model[1]

$$M(\theta_{-100},...,\theta_{100}) = \frac{n!}{n_{-100}!\cdot...\cdot n_{100}!}\Pi_{i=-100}^{100}\theta_i^{n_i} \tag{9}$$

The relevant magnitude for auditors is the total amount of error, λ, and the prior knowledge is assessed on λ, say $\xi(\lambda)$. This parameter is given by:

$$\lambda = \frac{\text{RBV}}{100}\cdot\sum_{i=-100}^{100} i\cdot\theta_i$$

Bayes' Theorem asserts that the logical way to modify prior beliefs about a unknown parameter is to combine prior and likelihood distributions resulting in a posterior one. Prior and likelihood function must be referred to same unknown parameter. However in our model, likelihood function refers to proportions of errors, θ_i's, and prior refers to total amount of error (a linear combination of the proportion of error). Thus, a traditional Bayesian would require that auditors supply a prior probability mass function for each possible proportion of errors θ_i's. However, there has been a

[1]Exactly, the multinomial model appears when the random sample is chosen with replacement. In other case, because in practice the total book value is very large in relation to sample size, the multinomial model is a good approximation of the likelihood function.

development which can be considered to be a modification of the usual likelihood for handling data from a QB scheme. Briefly, the approach defines the likelihood function for the unknown λ as the likelihood induced by $M(\theta_{-100},...,\theta_{100})$ in *Zehna (1966)* notation. An intuitive approach of that function is in *McCray (1984)*. For a complete development of that function see *Hernández et al. (1996)* and other references therein.

The posterior mass function, called Most Likely Posterior Curve (MLPC)[2], could be calculate by:

$$\xi(\lambda \mid \text{data}) = \frac{M^*(\lambda) \cdot \xi(\lambda)}{\sum_\lambda M^*(\lambda) \cdot \xi(\lambda)}$$

where:

$$M^*(\lambda) = \sum_\lambda \left(\sup_{\phi^{-1}(\{\lambda\})} M(\theta_{-100},...,\theta_{+100}) \right) \cdot I_{\phi^{-1}(\{\lambda\})}(\theta_{-100},...,\theta_{+100})$$

$$\phi^{-1}(\{\lambda\}) = \left\{ (\theta_{-100},...,\theta_{+100}) \mid \sum_{i=-100}^{+100} \theta_i = 1 \text{ and } \frac{RBV}{1000} \cdot \sum_{i=-100}^{+100} i \cdot \theta_i = \lambda \right\}$$

Therefore the QB formulation can be summarized via Bayes'Theorem with a maximized likelihood function. Any prior can be used.

This posterior distribution has two advantages,

1. it does no require modelling the accounting population taintig to obtain the likelihood (as in the most of the audit models, *Cox and Snell (1979)*, *Godfrey and Neter (1984)*, and *Felix and Grimlund (1977)*),
2. the user doesn't have to asses neither prior distribution on few intuitive magnitudes nor a prior distribution on a high dimensional parameter space (*Tsui et al. (1985)*).

Despite this two advantages, this model needs to elicit a complet prior distribution and it can be a very difficul task for auditors, though this is a prior distribution on one parameter, with a strongly intuitive meaning, the total amount of error. But this assignment is allways a subjective matter, and it could cause problems to justify its functional form.

It should be very useful a more objective procedure such that the prior distribution could include those issues with the most certainty, and outside of those, it could be the least informative prior distribution.

[2]The Quasi-Bayesian educational PC share ware **MLPC v.3.02** and manual are available through the Internet at the anonymous ftp site: **csg.uwaterloo.ca** in the directory: **pub/dmg/mlpc**.

This procedure even could include qualitative judgements like «excellent», «very good», ... about the evaluation of internal accounting control of the firm. It could be possible to identify each of the previous judgements with some aspects of the prior distribution, like the mean, one or more quantiles,... (of course, depending on the policy of the firm). This method can be possible using the Maximum Entropy Prior distribution, and this is the method we propose in this work, clearly more advantageous than the other usual estimation procedures of the total amount of error in auditing.

4. ILLUSTRATIONS

The following examples illustrate how above MEP's might be used in audit situations and the effect of different mean and/or quantiles on a few descriptives posterior upper bounds: $qb_{50}, qb_{80}, qb_{90}, qb_{95}, qb_{99}$. Consider the following DUS data from an inventory with a reported book value of $\$1,000,000$, a sample size of 100 items, and taints observed of 0, 10, 90 and -25 with number of cases of 94, 1, 1, 4, respectively. As an example, we give in Table 1 below a matrix of possible relationship between Internal Control and/or Analitical Review Evaluation (IC/AR) and prior information about the total amount of error in the inventory (usually, overstatement error).

Table 1. *Matrix of Possible Relationships between IC/AR Evaluation and Prior Information*

IC/AR Evaluation	Prior Information			
	Expected average tainting θ_0 (percent)			
Excellent	3			
Very Good	5			
Good	10			
Poor	15			
Very Poor	30			
	One Quantile			
	Maximum Tainting Percent (MTP)		Credibility (CR)	
Excellent	3		0.95	
Very Good	5		0.95	
Good	10		0.95	
Poor	15		0.80	
Very Poor	30		0.70	
	Two Quantiles			
	MTP	CR	MTP	CR
Excellent	3	0.95	5	0.99
Very Good	5	0.95	10	0.99
Good	10	0.95	15	0.99
Poor	15	0.80	30	0.95
Very Poor	30	0.70	50	0.80

For instance, in the one quantile case if IC/AR evaluation yields an output of Very Good, then auditors can feel comfortable accepting that the probability that total amount of error be minor than 5% is 0.95. The following table shows the behaviour of posterior probability in these situations. MEP refers to MEP with mean known, and MEP1 and MEP2 refer to MEP with one and two quantiles given, respectively.

Table 2. *Posterior Descriptive Quantities of Total Amount of Error*[3]

Distribution		IC/AR Evaluation				
		Excellent	V. Good	Good	Poor	V. Poor
Post. Mean	MEP	27872.07	26690.69	31237.37	31802.57	32488.20
	MEP1	32863.98	32972.65	32972.65	32972.65	32972.65
	MEP2	32873.64	32972.65	32972.65	32972.65	32972.65
Post. Mode	MEP	21050.04	28476.98	23682.89	24121.46	24651.69
	MEP1	25177.08	25040.71	25040.71	25040.71	25040.71
	MEP2	25164.96	25040.71	25040.71	25040.71	25040.71
Post. Median	MEP	25598.06	27286.12	28719.21	29242.20	29876.03
	MEP1	30301.68	30328.67	30328.67	30328.67	30328.67
	MEP2	30304.08	30328.67	30328.67	30328.67	30328.67
qb_{80}	MEP	35876.41	38302.43	40367.56	41122.54	42038.42
	MEP1	42618.63	42694.82	42694.82	42694.82	42694.82
	MEP2	42625.43	42694.82	42694.82	42694.82	42694.82
qb_{90}	MEP	42263.74	45155.88	47620.98	48522.93	49617.84
	MEP1	50253.08	50407.74	50407.74	50407.94	50407.74
	MEP2	50266.88	50407.74	50407.74	50407.94	50407.74
qb_{95}	MEP	48095.48	51417.08	54252.19	55290.94	56553.19
	MEP1	57152.31	57450.34	57450.34	57450.34	57450.34
	MEP2	57178.19	57450.34	57450.34	57450.34	57450.34
qb_{99}	MEP	60458.79	64718.94	68390.80	69748.42	71409.98
	MEP1	71221.28	72615.84	72615.84	72615.84	72615.84
	MEP2	71336.30	72615.84	72615.84	72615.84	72615.84

[3]These calculations were made using the educational freeware program MLPC, a Quasi-Bayesian software package. Number of data points was set at 500.

These results are shown in Table 2. For example, if there is a good evaluation of IC/AR, the qb_{80} is \$40367. This means the most likely probability the actual overstatement in the inventory balance is less than \$40367 is 0.80. Also for qb_{95}, an excellent evaluation of IC/AR results in almost a 15% reduction in the upper bound compare with a very poor evaluation of IC/AR.

5. COMMENTS AND CONCLUSIONS

In the proposed model in this paper, we must bear in mind that the magnitude λ is commonly more familiar to the auditor, the simplification is clear. Moreover, if we compare this model with other models based in the multinomial likelihood (see *Tsui et al. (1985)* or *Byekwaso (1994)*) we may take out some conclusions. All this models have the inconvenience of a high dimensional parametric space. Therefore, the choice of prior distribution, in practice, may only be carried out, if Dirichlet distribution is selected, requiring also a big effort to get the posterior.

The proposed model in this paper can be seen as a more simplified methodology compared to most of proposed models in the literature. This simplification is clear both is the conceptual level and for practical applications. Finally, since we have insisted on the advantage of this model for practical purposes, it is essential to optimize some nonlinear restricted mathematical programs. This equations can be solved by finding the minimum of an unconstrained function. These calculations are included in the MLPC software.

The combination of a maximum entropy prior and the most likely likelihood function appears to be well suited for audit and accounting applications of Bayesian analysis because it is easy to defend and support. The above example suggests that the resulting upper bounds are consistent with the prior and likelihood used.

In all situations, the fact that the mean is greater than the mode reflects that the posterior distribution has a moderate skew towards higher amount of error. Perhaps because one taint of 90% has been observed and the model is very sensible to it.

Respect to IC/AR initial classification, more conservatives upper bounds are obtained when quantiles are given than mean is given.

The use of the mean as prior information yields differents posterior distributions according to different IC/AR evaluation, note a reduction in the upper bounds from Very Poor to Excellent classification.

Using quantiles no differences are appreciated in the posterior distribution respect to IC/AR evaluation, only between Excellent and the remainder situations. There isn't difference between this situations because all priors are identical in the support of the

most likely likelihood and differ where the likelihood is almost zero. Also, this fact occurs between MEP1 and MEP2.

Thus, when the minimum amount of information is required prior mean is a good election. It maybe better than quantiles, because in auditing context usually 95 and 99 quantiles are very close. Possibly, 50 and 95 quantiles yields better results.

This work could be extended incorporating another intuitive magnitude like the mode (see *Brockett et al. (1984)*).

REFERENCES

[1] **Berger, J.O.** (1985). *Statistical Decision Theory and Bayesian Analysis* (2nd edition). Springer - Verlag, New York.

[2] **Brockett, P., Charnes, A.** and **Paick, K.** (1984). «A Method for Constructing a Unimodal Inferential or Prior Distribution». *Research Report* #473. Center for Cybernetic Studies, The University of Texas at Austin.

[3] **Byekwaso, S.** (1994). «Bayesian Sequential Inference for Error Rate and Error Amounts in Accounting Data». *Ph.D. Thesis*, The American University, Dpt. of Mathematics and Statistics. Washington.

[4] **Cox, D.R.** and **Snell, E.J.** (1979). «On Sampling and the Estimation for Rare Errors». *Biometrika*, **66, 1**, 125–132. Amendments and Corrections in vol. 69, page 491, 1982.

[5] **Felix, W.L. (Jr.)** and **Grimlund, R.A.** (1977). «A Sampling Model for Audit Tests of Composite Accounts». *Journal of Accounting Research*, Spring, 23–41.

[6] **Godfrey, J.T.** and **Neter, J.** (1984). «Bayesian Bounds for Monetary — Unit — Sampling in Accounting and Auditing». *Journal of Accounting Research*, **XX, 2**, Part 1, 304–315.

[7] **Hernández, A., McCray, J.H.** and **Vázquez-Polo, F.J.** (1996). «Bayesian Inference on a Subparameter. An Application in Auditing». Submitted to *Journal of Applied Statistics.*

[8] **Jaynes, E.T.** (1968). «Prior Probabilities». *IEEE Trans. Systems, Science and Cybernetics*, **4**, 227–291.

[9] **McCray, J.H.** (1984). «A Quasi-Bayesian Audit Risk Model for Dollar Unit Sampling». *The Accounting Review*, **LIX, 1**, 35–51.

[10] **McCray, J.H.** (1986). «A General Bayesian Risk Model for Dollar Unit Sampling and Multiple Populations».*Proceedings of VII Biennial Audit Research Symposium* University of Illinois, 181–252.

[11] **Tsui, K., Matsumura, E.M.** and **Tsui, K.L.** (1985). «Multinomial — Dirichlet Bounds for Dollar — Unit Sampling in Auditing». *The Accounting Review*, **LX, 1**, 76–96.

[12] **Zehna, P.W.** (1966) «Invariance of Maximum Likelihood Estimation». *Annals of Mathematical Statistics*, **37**, 755.

QÜESTIIÓ, vol. 22, 2, p. 243-252, 1998

ON SOME STRATEGIES USING AUXILIARY INFORMATION FOR ESTIMATING FINITE POPULATION MEAN

L.N. SAHOO
Utkal University*

J. SAHOO
Orissa University**

M. RUÍZ ESPEJO
UNED*

This paper presents an empirical investigation of the performances of five strategies for estimating the finite population mean using parameters such as mean or variance or both of an auxiliary variable. The criteria used for the choices of these strategies are bias, efficiency and approach to normality (asymmetry).

Keywords: Auxiliary information, empirical study, finite population mean, sampling.

AMS Classification: 62 D 05

* Department of Statistics, Utkal University. Bhubaneswar 751004, India.
** Department of Statistics, Orissa University of Agriculture and Technology, Bhubaneswar 751003, India.
⋆ Departamento de Economía Aplicada Cuantitativa. Universidad Nacional de Educación a Distancia (UNED), 28040 Madrid.

- Received January 1997.
- Accepted September 1997.

1. INTRODUCTION

Let y_i, x_i $(i = 1,2,...,N)$ be the values of a survey variable y and a positively correlated auxiliary variable x on the ith unit of a finite population of size N. Suppose that x_i's are known and an estimate is needed for the population mean $\bar{Y}$ of y on the basis of a sample s of fixed size n, selected by simple random sampling without replacement (SRSWOR). It is known that the sample mean

$$\bar{y} = \frac{1}{n}\sum_{i\in s} y_i$$

provides an unbiased estimate of $\bar{Y}$. But, use of the auxiliary variable x very often provides an estimator with increased accuracy. With this objective when the population mean $\bar{X}$ of x is known, traditional ratio method estimates $\bar{Y}$ by

$$t_1 = \bar{y}\frac{\bar{X}}{\bar{x}}$$

where

$$\bar{x} = \frac{1}{n}\sum_{i\in s} x_i.$$

However, this can be rendered unbiased under Sen-Midzuno (1952) [SM, say] scheme in which the sample s is selected with probability

$$p(s) = \binom{N}{n}^{-1}\frac{\bar{x}}{\bar{X}}.$$

In many practical situations σ_x^2, the population variance of x or equivalently $S_x^2 = N\sigma_x^2/(N-1)$ is known, or can be calculated from the known x-values of the population units. Then it is quite reasonable to consider the ratio-type estimator of $\bar{Y}$

$$t_2 = \bar{y}\frac{S_x^2}{s_x^2}$$

where

$$s_x^2 = \frac{1}{n-1}\sum_{i\in s}(x_i - \bar{x})^2.$$

Singh and Srivastava (1980) forwarded a sampling scheme [SS, say] in which a pair of units, (i,j) say, is selected with probability proportional to $(x_i - x_j)^2$ and then an SRSWOR sample of $(n-2)$ units is drawn from the $(N-2)$ that remain. That is, the probability of selecting a sample is

$$p(s) = \binom{N}{n}^{-1}\frac{s_x^2}{S_x^2}.$$

Cleary, for this design, t_2 provides an unbiased estimator for $\bar{Y}$.

For many populations encountered in practice, both $\bar{X}$ and S_x^2 are known in advance. Then one may think of using these parameters simultaneously and lead to consider a ratio-in-ratio estimator (which is of course biased)

$$t_3 = \bar{y}\frac{\bar{X}}{\bar{x}}\frac{S_x^2}{s_x^2}.$$

One basic question, not yet definitely answered, is how to make a choice between the parameters $\bar{X}$ or S_x^2 or both simultaneously, providing a good estimate of $\bar{y}$. The literature to date offers little guidance in this choice. The main reason perhaps being that one cannot easily evaluate the relative performance of one estimator over another. Because, the estimators considered above do not involve the same parameters in the resulting expressions of their variances or mean square errors. Also, in most of the cases the results are only in asymptotical forms which do not usually provide any meaningful conclusions preferably for smallish samples. So, the present investigation on the comparative performance is made with the help of an empirical study carried over a wide variety of natural populations.

2. STRATEGIES UNDER STUDY AND THEIR PERFORMANCE MEASURES

For the evaluation of the comparative performances, we now consider five competing strategies, viz H_1 = (SRSWOR, t_1), H_2 = (SM, t_1), H_3 = (SRSWOR, t_2), H_4 = (SS, t_2) and H_5 = (SRSWOR, t_3). To further facilitate the comparison, especially with efficiency, we consider the conventional unbiased strategy H_0 = (SRSWOR, $\bar{y}$). The strategies are compared under the following performance measures:

(a) Relative Bias $(RB) = |Bias|/\bar{Y}$:

We leave aside the strategies H_2 and H_4 in this assessment since they are completely unbiased.

(b) Relative Efficiency (RE):

The relative efficiency of a strategy is calculated in comparison to the strategy H_0. For this purpose, mean square error or variance is taken as a measure of efficiency according as a strategy is biased or unbiased.

(c) Approach to Normality (Asymmetry):

The coefficients of skewness and kurtosis, i.e. β_1 and β_2 coefficients are considered as the indices for measuring the asymmetry of the sampling distribution of a strategy. We say asymmetricity is nullified when $\beta_1 = 0$ and $\beta_2 = 3$.

3. DESCRIPTION OF THE EMPIRICAL STUDY

Our empirical study considers 20 natural populations available in some traditional sampling texts and journals articles described in table 1. We draw all $\binom{N}{n}$ possible samples for $n = 3, 4$ and 5 from a given population and calculate the average behaviour of the estimators through different performance measures. To save space, the results for $n = 4$ are not given. However, the main findings are discussed in subsections 3.1 to 3.4.

3.1. Results Based on *RB*

The results in table 2 indicate that the *RB* of H_1 is the least in all the populations. It is also true for varying size of the sample. The strategy H_3 follows H_1 in almost all the cases except in population 11 for $n = 3$, and in populations 11 and 17 for $n = 5$. On the other hand, H_5 seems to be highly biased and has very poor performance in the sense of *RB*.

Table 1. *Description of populations*

Pop. n°	Source	Size (N)	Pop. n°	Source	Size (N)
1	Cochran (1977, p. 203)	10	11	Singh *et al.* (1986, p. 279)	20
2	Cochran (1977, p. 325)	12	12	Singh *et al.* (1986, p. 286)	16
3	Konijn (1973, p. 49)	16	13	Singh *et al.* (1986, p. 287)	12
4	Murthy (1967, p. 422)	24	14***	Sukhatme *et al.* (1970, p. 185)	34
5	Singh *et al.* (1986, p. 144)	11	15′	Sukhatme *et al.* (1970, p. 185)	34
6	Singh *et al.* (1986, p. 155)	17	16″	Murthy (1967, p. 228)	32
7*	Singh *et al.* (1986, p. 166)	16	17‴	Murthy (1967, p. 228)	32
8**	Singh *et al.* (1986, p. 166)	16	18	Murthy (1967, p. 398)	32
9	Singh *et al.* (1986, p. 176)	13	19	Horvitz *et al.* (1952)	20
10	Singh *et al.* (1986, p. 176)	20	20	Sampford (1962, p. 61)	35

*x = area under wheat during 1978-79;
**x = total cultivated area during 1978-79;
***x = area under wheat in 1936;
′x = total cultivated area in 1931;
″x = n° of workers;
‴x = fixed capital.

Table 2. *Features of the RB of strategies = |Bias|/Mean*

Pop.	$n = 3$			$n = 5$		
n°	H_1	H_3	H_5	H_1	H_3	H_5
1	0.0008	5.6959	5.6298	0.0003	0.3829	0.4005
2	0.0011	2.0859	2.1596	0.0005	0.3883	0.4099
3	0.0006	2.2992	2.3647	0.0003	0.2385	0.2755
4	0.0313	8.7504	20.358	0.0174	2.3678	2.8392
5	0.0131	1.3775	1.8390	0.0054	0.1909	0.2725
6	0.0006	3.1119	3.1378	0.0004	0.9403	0.9436
7	0.0069	4.6651	6.6123	0.0043	0.6373	1.1035
8	0.0158	4.3740	6.4912	0.0080	1.1520	1.9969
9	0.0242	7.8101	9.4719	0.0100	0.8412	0.9158
10	0.0460	12.975	16.398	0.0325	1.5978	3.0081
11	0.0006	7.2642	7.1822	0.0002	0.2522	0.1891
12	0.0002	4.8377	4.8564	0.0001	0.6161	0.6260
13	0.0159	1.4597	2.2027	0.0080	0.2632	0.4099
14	0.0089	9.9166	22.294	0.0023	1.0743	1.9404
15	0.0664	3.6774	4.0725	0.0360	0.3531	0.4757
16	0.0172	4.3482	6.7701	0.0059	1.0134	1.7833
17	0.0470	0.2536	0.6299	0.0287	0.1439	0.1182
18	0.0336	3.3065	4.5126	0.0193	0.4481	0.4912
19	0.0019	3.0279	3.1656	0.0009	0.6904	0.9251
20	0.0530	15.822	44.207	0.0205	7.7060	21.983

3.2. Results Based on *RE*

Table 3 reveals that the strategies H_3, H_4 and H_5 perform very badly in comparison to H_0 in view of efficiency. Both H_1 and H_2 fare well in relation to H_0 having noticeable performances in some populations like 1, 11 and 12 etc. H_2 is more efficient than H_1 in most of the cases except a very few, as for example in populations 11, 14 and 19

for $n = 3$, and in populations 6, 11, 14 and 18 for $n = 5$. However from $n = 3$ to $n = 5$, H_1 decreases weakly in *RE* for eight populations (3, 6, 7, 10, 14, 15, 18 and 19); also H_2 decreases weakly in *RE* from $n = 3$ to $n = 5$, for other seven populations (6, 7, 10, 13, 14, 15 and 18). But, for the rest of the twelve or thirteen natural populations, H_1 and H_2 (respectively) have bigger *RE* for $n = 5$ with respect $n = 3$.

Table 3. *Features of the RE of strategies w.r.t.* H_0 *(in %)*

Pop.	$n = 3$					$n = 5$				
n°	H_1	H_2	H_3	H_4	H_5	H_1	H_2	H_3	H_4	H_5
1	1613	1639	0.0009	0.1487	0.0010	1636	1647	0.3194	1.0035	0.3050
2	155.00	157.23	0.0127	0.2755	0.0122	158.00	158.55	0.1569	0.6359	0.1435
3	653.15	653.20	0.0048	0.1607	0.0045	645.07	662.44	0.0942	0.5973	0.0795
4	779.82	1063	0.0057	4.4305	0.0008	1129	1186	0.5650	5.1541	0.4448
5	259.38	315.15	0.3843	5.9564	0.2618	302.65	330.68	7.0257	19.603	3.0228
6	131.58	132.21	0.0001	0.1969	0.0001	125.36	125.18	0.0006	0.0037	0.0006
7	1660	2050	0.0290	3.0787	0.0073	1608	1997	3.6131	11.828	1.2884
8	745.21	990.33	0.0576	3.5645	0.0310	773.64	995.88	1.0131	7.4796	0.3356
9	351.40	432.18	0.0024	1.2003	0.0019	386.78	433.88	1.5360	8.0791	1.3611
10	346.73	393.66	0.0041	1.1193	0.0030	237.43	330.28	2.2957	6.1539	0.1379
11	15324	15163	0.0002	0.1313	0.0003	19600	18195	1.1436	1.8824	1.3473
12	21366	21563	0.0017	0.2242	0.0018	22174	22284	0.1251	0.9129	0.1187
13	719.33	838.99	0.6454	13.439	0.3120	760.78	830.37	10.775	35.358	3.858
14	1618	1255	0.0012	2.8122	0.0016	986.89	920.75	1.1681	11.601	0.3320
15	571.11	590.67	0.0098	3.5286	0.0193	287.30	311.99	7.1565	22.604	4.5010
16	270.62	301.39	0.0012	0.8000	0.0005	621.97	734.83	0.0578	8.2344	0.0243
17	185.83	262.69	0.3355	2.9183	0.1267	535.12	891.82	26.591	13.875	19.117
18	132.03	152.14	0.0181	2.7558	0.0071	114.03	107.08	0.9121	6.5094	1.0867
19	446.11	426.88	0.0281	1.7332	0.0313	430.08	446.97	1.3172	4.9535	0.7936
20	252.31	330.51	0.0009	2.1083	0.0001	349.37	596.19	0.0275	4.4305	0.0033

Table 4. *Features of the coefficient of skewness*

Pop.			$n=3$					$n=5$		
n°	H_1	H_2	H_3	H_4	H_5	H_1	H_2	H_3	H_4	H_5
1	0.0001	0.0002	63.70	4790	63.95	0.0097	0.0113	12.94	12.51	10.39
2	0.0281	0.0316	25.66	190.7	24.87	0.0026	0.0035	16.11	22.59	13.76
3	0.0185	0.0157	47.59	445.6	46.67	0.0053	0.9431	72.49	73.81	74.53
4	0.2258	0.1513	1077	25914	136953	0.0311	0.0142	32.67	37.24	12.26
5	0.3009	0.2805	43.74	241.5	22.91	0.0793	0.0766	16.59	12.14	34.29
6	0.0578	0.0565	81.38	827.0	82.92	0.0166	0.1387	38.22	27.48	37.58
7	0.0001	0.0002	60.97	1926	267.6	0.0024	0.0392	3.8595	4.8007	10.56
8	0.1808	0.3463	46.85	777.7	35.75	0.1539	0.2099	8.8371	23.43	15.01
9	0.0186	0.0021	186.8	34483	119.3	0.0259	0.0325	38.92	37.01	21.22
10	0.0030	0.0041	206.6	5776	125.7	0.0417	0.8553	57.25	10.32	302.4
11	0.0579	0.0515	702.8	43523	677.9	0.0012	1.0053	1.9188	1.5136	1.9405
12	0.0137	0.0111	162.3	3409	160.9	0.0040	0.1524	45.30	62.45	44.32
13	0.1036	0.1666	37.56	209.9	32.40	0.0923	0.1243	13.30	10.96	22.59
14	1.2850	2.0191	488.4	268192	602.6	0.7797	0.8653	9.6147	36.24	11.35
15	0.0042	0.8857	1241	137580	311.4	0.0259	1.3711	3.7598	6.1541	3.9093
16	0.0341	0.2554	254.7	47038	255.7	0.0089	1.7938	645.4	7400	656.7
17	0.4526	0.4448	196.8	43.99	440.8	0.1674	3.5797	8.2159	0.7102	16.22
18	0.7658	0.0728	259.1	5513	384.5	0.5208	1.0834	46.00	77.16	41.46
19	0.5121	0.4977	136.1	1447	109.8	0.3882	0.2067	7.3445	9.1215	7.2631
20	0.0068	0.3711	572.2	109294	656.1	0.0020	1.8820	21.39	1230	23.99

3.3. Results Based on Skewness

The results in table 4 show that the strategies H_3, H_4 and H_5 have very much skewed distributions for $n = 3$. As the sample size increases to 5 through 4, there is a sudden fall in the asymmetricity of their distributions. Even for moderate size sample they have erratic behaviour and their performance under this measure is not at all

convincing excepting a very few. On the other hand, the strategies H_1 and H_2 have their distributions near the symmetry. H_2 dominates H_1 in 8 populations for $n = 3$ and in only 3 populations for $n = 5$ in the sense of approaching towards symmetry. This means that H_1 would have a decidedly better performance over H_2 when sample is of moderate size.

Table 5. *Features of the coefficient of kurtosis*

Pop.	$n = 3$					$n = 5$				
n°	H_1	H_2	H_3	H_4	H_5	H_1	H_2	H_3	H_4	H_5
1	2.511	2.526	72.81	7135	73.17	2.576	2.583	25.45	30.88	21.75
2	2.583	2.599	32.01	339.6	31.26	2.566	2.571	29.29	47.85	25.15
3	2.598	2.594	59.28	869.2	57.70	2.608	2.630	101.4	208.3	103.7
4	3.730	4.188	1277	166953	1348	2.876	2.757	68.40	93.17	34.60
5	3.569	3.722	59.97	482.1	31.30	3.032	3.045	31.65	31.06	52.59
6	2.736	2.733	107.5	1867	109.7	2.646	2.691	84.73	77.95	83.77
7	3.132	3.288	67.99	3519	318.7	2.849	3.266	9.538	12.34	21.17
8	3.502	4.072	59.17	1550	45.16	3.255	3.845	16.62	46.85	29.93
9	3.125	3.040	207.1	58235	126.2	2.585	2.547	66.88	98.71	42.87
10	3.050	2.800	295.6	17439	174.1	2.825	3.041	130.4	37.40	421.7
11	2.617	2.586	792.8	128536	770.4	2.690	2.568	5.483	5.745	5.453
12	2.579	2.572	212.5	8364	211.0	2.657	2.599	65.60	156.3	63.59
13	2.793	2.898	45.34	416.1	42.02	2.805	2.905	27.08	27.36	34.10
14	5.480	4.976	561.2	533050	696.2	4.500	4.408	14.72	65.79	15.66
15	2.688	2.576	1404	272466	367.9	2.571	2.478	7.768	14.33	8.136
16	2.659	2.771	288.9	88816	288.9	2.565	1.938	768.8	26006	779.9
17	3.578	4.950	315.4	287.9	651.0	3.023	4.277	23.30	0.8021	40.96
18	3.675	3.285	315.1	13675	446.4	3.302	1.847	68.30	169.8	62.60
19	2.811	2.727	169.9	3582	138.5	2.733	2.729	16.73	21.13	14.97
20	2.465	3.344	658.8	375810	757.2	2.344	4.350	29.33	2089	32.82

3.4. Results Based on Kurtosis

The results in table 5 show that the strategies H_3, H_4 and H_5 have extremely leptokurtic type of distributions which lead to the conclusions that the strategies provide estimates highly concentrated around the parameter they are estimating. But with steady increase in the sample size, they considerably slow down their peakedness having distributions spectacularly above the normality. On the other hand, both H_1 and H_2 having their sampling distributions near the normality. They perform equally well in view of their approach to normality and this tendency is also nearly true even for moderate size of the sample. For $n = 5$, H_1 seems to be slightly better than H_2.

4. CONCLUSIONS

Since our study is realized for 20 particular populations, the conclusions following are justified for these concrete natural data. The present empirical study leads to the overall conclusions that the performances of the strategies H_3, H_4 and H_5 are decidedly very much unsatisfactory under different measures. H_1 has the least bias among the biased strategies. The unbiased strategy H_2 is more efficient than H_1 in most of the cases. But in view of asymmetry, H_1 seems to be better than H_2 although they are almost compatible when considered under their approach to normality. The study thus suggests a straightforward rejection of H_3, H_4 and H_5. H_2 is advisable to be preferred over H_1 from efficiency viewpoint. The only demerit that lies with H_2 is that it may assume negative variance estimators for certain samples. But, a sufficient condition that H_2 will provide non-negative variance estimator is given in a recent paper by Sahoo and Sahoo (1995). One can then choose a sample just fulfilling the requirements of such a condition.

5. REFERENCES

[1] **Cochran, W.G.** (1977). *Sampling Techniques.* Third Edition, Wiley Eastern Limited, New Delhi.

[2] **Horvitz, D.G.** and **Thompson, D.J.** (1952). «A generalization of sampling without replacement from a finite universe». *J. Amer. Statist. Assoc.*, **47**, 663-685.

[3] **Konijn, H.S.** (1973). *Statistical Theory of Sample Survey Design and Analysis.* North-Holland, Amsterdam.

[4] **Midzuno, H.** (1952). «On the sampling system with probability proportionate to sum of sizes». *Ann. Inst. Statist. Math.*, **3**, 99-107.

[5] **Murthy, M.N.** (1967). *Sampling Theory and Methods.* Statistical Publishing Society, Calcutta.

[6] **Sahoo, J.** and **Sahoo, L.N.** (1995). *On three unbiased strategies in sample surveys. Unpublished manuscript.*

[7] **Sampford, M.R.** (1962). *An Introduction to Sampling Theory.* Oliver and Boyd, Edinburg.

[8] **Sen, A.R.** (1952). «Present status of probability sampling and its use in estimation in farm characteristics (Abstract)». *Econometrica,* **20**, 130.

[9] **Singh, D.** and **Chaudhary, F.S.** (1986). *Theory and Analysis of Sample Survey Designs.* Wiley Eastern Limited, New Delhi.

[10] **Singh, P.** and **Srivastava, A.K.** (1980). «Sampling schemes providing unbiased regression estimators». *Biometrika,* **67**, 205-209.

[11] **Sukhatme, P.V.** and **Sukhatme, B.V.** (1970). *Sampling Theory of Surveys with Applications.* Second Revised Edition, Iowa State University Press, Ames.

QÜESTIIÓ, vol. 22, 2, p. 253-289, 1998

REFLEXIONES SOBRE LA ESTRATEGIA DE MEDIDA DE LOS CAMBIOS EN PROBABILIDAD EN MODELOS DE ELECCIÓN BINARIOS

M.T. APARICIO*
I. VILLANÚA*
Universidad de Zaragoza*

Este trabajo se centra en la evaluación de la medida que, en el marco de los modelos de elección binarios o dicotómicos, se utiliza para reflejar el cambio en la probabilidad ante la variación de una de las variables explicativas. La opción de cuantificación más común ha consistido en utilizar el vector de valores medios de las variables explicativas, lo que podemos entender como poner el énfasis en el comportamiento de «un individuo medio». Frente a esta práctica habitual, efectuamos una propuesta de evaluación alternativa que atiende fundamentalmente al «comportamiento medio de la muestra». Se diseña, asimismo, un experimento de Monte Carlo que permite probar que la propuesta desarrollada en este trabajo es más robusta que la tradicionalmente planteada en la literatura.

Reflections on the probability change measurement strategy in binary choice models.

Palabras clave: Elección discreta, variación de probabilidad, robustez.

Clasificación AMS: 62J02, 65C05

*Los autores desean agradecer los comentarios de Antonio Aznar, Carmen García-Olaverri y de un evaluador anónimo, así como la financiación de DGICYT PB94-0602.

⋆Departamento de Análisis Económico. Facultad de Ciencias Económicas y Empresariales. Universidad de Zaragoza. Gran Vía, 2. 50005 Zaragoza.

–Recibido en julio de 1996.
–Aceptado en enero de 1998.

1. INTRODUCCIÓN

En los últimos años, los denominados modelos de elección discreta (MED) han tenido un auge considerable, debido al mayor énfasis de la investigación econométrica en modelos desagregados. Este hecho ha venido propiciado por dos razones fundamentales; en primer lugar, el reconocimiento de que el comportamiento económico relevante debe ser tratado a nivel individual, mejor que mediante la consideración de agregados de acciones individuales y, en segundo lugar, por la posibilidad cada vez mayor de disponer de datos a nivel microeconómico, es decir, de las unidades que toman las decisiones.

El término MED hace referencia a toda una clase de modelos que tratan de calcular la probabilidad de que una unidad de decisión escoja una alternativa concreta de entre un conjunto de opciones, dada la respuesta del individuo sobre la elección efectuada (y) y todo un conjunto de variables socio-económicas (x) de las cuales depende obviamente la elección y que representan características de las alternativas disponibles y características del sujeto decisor. Se pueden distinguir diferentes modelos dentro de la clase mencionada, atendiendo a la diferente forma funcional que relaciona los datos observados con la probabilidad o bien según la extensión del conjunto de elección; en este último sentido, el planteamiento más sencillo corresponde a los MED dicotómicos. Centrándonos en estos, la probabilidad de que la unidad de decisión $i-$ésima elija una de las alternativas se especifica como una función general de la forma:

$$p_i = F(x_i'\beta) \tag{1}$$

donde x_i es el vector de variables socio-económicas continuas y discretas asociadas con la unidad $i-$ésima y β es un vector de parámetros desconocidos. La justificación teórica de esta formulación se establece en base a la naturaleza de los procesos de toma de decisiones, utilizando los conceptos de la teoría de la utilidad como puede verse, entre otros, en Amemiya (1981), McFadden (1970), Maddala (1983) o Train (1986).

A partir de (1), la determinación concreta de F conduce a modelos específicos como el modelo probit o el modelo logit, según se identifique F con la función de distribución de la normal tipificada o de la logística estandarizada, respectivamente. En ambos casos, la estimación de los coeficientes se realiza por el método de máxima verosimilitud, requiriéndose el uso de algoritmos de optimización para modelos no lineales; tal estimación permite obtener, para cada unidad de decisión, la probabilidad estimada de elección.

A pesar del gran número de trabajos empíricos que utilizan este tipo de modelos, la etapa de validación de los mismos suele recibir bastante poca atención y lo normal es presentar los coeficientes estimados y sus correspondientes $t-$ratios, acompañados

de una de las múltiples medidas escalares de bondad del ajuste propuestas para planteamientos de este tipo[1] y del contraste LR relativo a la significación conjunta del modelo. La interpretación de las magnitudes anteriores conforman las conclusiones del estudio, las cuales se completan frecuentemente con el análisis del efecto que sobre la probabilidad de elección tienen las variaciones de las diferentes variables socio-económicas que intervienen en el supuesto concreto. Esta forma tan escueta de entender la etapa de evaluación del modelo se refleja claramente en la afirmación de Pagan y Vella (1989), cuando en relación con los estudios que utilizan MED establecen: «Un pequeño muestreo de la literatura revela una gran diferencia respecto de la vieja tradición de los modelos de series temporales, en el sentido de no prestar gran atención a la calidad del modelo estimado».

El objetivo de este trabajo es centrarse en una de las cuestiones que hemos mencionado como integrantes de la valoración habitual de este tipo de modelos. En concreto, nos interesa profundizar en la medida utilizada para determinar la variación de la probabilidad ante cambios de las variables explicativas. Como es bien conocido, en este tipo de modelos no lineales la magnitud de los coeficientes por sí sola no permite sintetizar tal información y el instrumento adoptado a tal fin exige calcular $x_i \hat{\beta}$, con lo cual, una vez obtenido el vector de estimadores habrá que concretar el valor que se adopta para las diferentes variables explicativas.

El análisis de las diferentes estrategias de evaluación, habitualmente propuestas en la literatura, constituye el núcleo central de este trabajo, de modo que la discusión acerca de las ventajas e inconvenientes de cada una de ellas nos sirve de base para proponer una vía alternativa de evaluación que hace hincapié en el comportamiento medio de la muestra y que, desde nuestro punto de vista, refleja un mejor uso de la evidencia muestral.

La comparación de la estrategia propuesta con la práctica habitual se realiza diseñando un experimento de simulación que permita analizar la «robustez» de la nueva medida.

Establecido el tema objeto de discusión, el trabajo se estructura de la forma siguiente: en la Sección 2 nos centramos en el instrumento que refleja la variación en la probabilidad, con la única intención de tratar de rebatir algunas interpretaciones del mismo que consideramos erróneas y que han llevado a algunos autores a cuestionar su validez; en la Sección 3 establecemos las diversas estrategias existentes para concretar la variación mencionada poniendo de manifiesto las limitaciones que, a nuestro juicio, contienen lo que nos lleva a proponer un esquema amplio de razonamiento alternativo; en la Sección 4 se lleva a cabo un ejercicio de simulación para determinar posibles pautas de comportamiento de la estrategia propuesta en relación con el

[1] Medidas de este tipo, denominadas comúnmente como pseudo-R^2 han sido propuestas por McFadden (1974), Maddala (1983), Aldrich y Nelson (1984) y McKelvey y Zavoina (1975). Una comparación de todos ellos es presentada en Windmeijer (1995).

enfoque habitual en la literatura; en la Sección 5 se ilustran las distintas estrategias de análisis y su comportamiento mediante un ejemplo con datos reales, y en la última se resumen las conclusiones principales.

2. LA VARIACIÓN DE LA PROBABILIDAD ANTE CAMBIOS DE LAS VARIABLES INDEPENDIENTES

Como es bien conocido, en un modelo econométrico lineal la magnitud de los coeficientes que acompañan a las variables explicativas, proporcionan la variación de la variable endógena ante un cambio unitario de las primeras; asimismo, el signo del coeficiente está indicando la dirección de tal cambio. En consecuencia, los coeficientes son interpretables directamente, tanto en magnitud como en signo.

En el caso de los modelos no lineales que nos ocupan, lo que se predice es la probabilidad de elección de una alternativa, de modo que el efecto de interés se centra en la variación de tal probabilidad con respecto a la variable explicativa; pero dada la forma funcional específica del modelo, la magnitud del coeficiente no nos da directamente este efecto, aunque el signo del coeficiente seguirá reflejando el sentido del cambio.

Lo que interesa, en definitiva, es obtener el ratio entre la variación de la probabilidad y la propia variación de dicha variable, es decir, $\Delta p_i/\Delta x_j$ cuya expresión analítica podemos escribir como:

$$\frac{\Delta p_i}{\Delta x_j} = \frac{F(x_i'\beta + \beta_j \Delta x_j) - F(x_i'\beta)}{\Delta x_j} \tag{2}$$

de modo que para variaciones infinitesimales de x_j resulta:

$$\lim_{\Delta x_j \to 0} \frac{\Delta p_i}{\Delta x_j} = \lim_{\Delta x_j \to 0} \frac{F(x_i'\beta + \beta_j \Delta x_j) - F(x_i'\beta)}{\Delta x_j} = \frac{\partial p_i}{\partial x_j} \tag{3}$$

concretando $\partial p_i/\partial x_j$ para los modelos expresados en (1) se tiene:

$$\frac{\partial p_i}{\partial x_j} = f(x_i'\beta)\,\beta_j \tag{4}$$

donde $f(\cdot)$ denota la correspondiente función de densidad. La expresión (4) muestra con claridad lo manifestado anteriormente, en el sentido de que β_j por sí solo no indica la magnitud de la variación, la cual, además, depende de los valores concretos adoptados por las variables explicativas, mientras que el signo del coeficiente sí que determina la dirección del cambio, puesto que $f(\cdot)$ siempre será positiva por ser una función de densidad.

La consideración de la igualdad reflejada en (3) permite una interpretación correcta de la derivada parcial en modelos de este tipo (no lineales), y evita confusiones como las reflejadas en los artículos de Petersen (1985) y LeCleer (1992). Ambos autores critican la utilización de (4) por no ser una medida acotada entre cero y uno y, en consecuencia, posibilitar la generación de probabilidades fuera del rango plausible. En este sentido, LeCleer (1992) argumenta: «Una interpretación común pero inapropiada de la derivada parcial evaluada en la media de las variables independientes, es como probabilidad marginal con respecto a un cambio unitario en cada variable independiente. El problema está en que la derivada parcial no tiene límite y puede generar probabilidades fuera del rango plausible».

Por su parte Petersen (1985) apunta: «La fórmula frecuente e incorrectamente utilizada para calcular el cambio en p_i ante un cambio en x_j es $\partial p_i/\partial x_j$. Pero esto no nos da el cambio en p_i por un cambio unitario en x_j, ni tampoco por un cambio infinitesimal en x_j». El autor justifica estas afirmaciones diciendo que la derivada en funciones no lineales tiene una interpretación diferente, en el sentido de proporcionarnos la pendiente de p_i con respecto a x_j, cuando el denominador se aproxima a cero, y concluye afirmando: «... la pendiente no tiene límite porque los coeficientes no están acotados. Así pues, la pendiente puede ser mayor que uno y generar probabilidades fuera del intervalo cero-uno».

Desde nuestro punto de vista, no existe ningún problema en la utilización de (4) si, como hemos dicho antes, entendemos lo que realmente significa, esto es, el cambio en p_i en proporción al de x_j, de modo que para variaciones pequeñas de x_j, el cambio en probabilidad se aproximará a $\dfrac{\partial p_i}{\partial x_j}\Delta x_j$. En cualquier caso, la variación de x_j se considerará grande o pequeña dependiendo del rango de valores en el que se mueve la variable explicativa.

LeCleer (1992) ilustra lo que considera el problema de la no acotación, utilizando datos reales obtenidos de un estudio realizado por Hill y Ingram (1989) sobre la elección por parte de las entidades de préstamos entre dos principios contables, mediante la aplicación de un modelo logit. En base a la información suministrada, via cuestionario, por 113 entidades financieras de Estados Unidos, correspondiente al primer semestre de 1981, sobre un conjunto de variables financieras (activo total, activo neto/depósitos, préstamos hipotecarios/activo, depósitos/activo, etc.) se formulan el conjunto de variables explicativas que aparecen en el modelo logit objeto de estimación. Utilizando los resultados de esta etapa y el conjunto de medidas descriptivas que, de cada factor, proporcionan los autores, LeClere (1992) calcula $\partial p_i/\partial x_j$ para cada variable independiente de tipo continuo, obteniendo para una de ellas un valor de 89,0082 (para el resto de variables continuas, $\partial p_i/\partial x_j$ se encuentra entre cero y uno), lo que le lleva a cuestionar la consideración de la expresión (4) como media del cambio en probabilidad.

Ahora bien, si efectuamos una interpretación correcta de la derivada considerando como unidad de cambio no la unidad en sí, sino lo que podemos denominar como «unidad de cambio coherente», esto es, la correspondiente al intervalo en que se mueve la variable, la utilización de este instrumento no plantearía ningún problema. En este caso concreto, las medidas descriptivas correspondientes a la variable en cuestión proporcionan una media de 0'0049, una desviación típica de 0'0028 y un rango de valores entre 0 y 0'0209, todo lo cual indica claramente que la «unidad de cambio coherente» no va a ser la unidad y, en consecuencia, la variación en la probabilidad ante cambios unitarios (en el sentido mencionado) de dicha variable, estará ya comprendida entre cero y uno como corresponde a tal concepto.

Podemos concluir, por tanto, que cuando estamos analizando el cambio en probabilidad ante un cambio unitario (entendiendo la unidad en el contexto correspondiente) de la variable independiente, la expresión analítica que se aproxima a dicho cambio es $\frac{\partial p_i}{\partial x_j}\Delta x_j$ y esta aproximación será mejor a medida que $\Delta x_j \to 0$; de esta manera, la no acotación de la derivada no plantea ningún problema.

Todo el análisis efectuado exige que las variables x's sean continuas, por cuanto la derivada parcial no existe cuando tales variables son de tipo discreto. Ahora bien, en la mayoría de las especificaciones que nos ocupan, alguna o varias de las variables que representan las características socio-económicas son variables discretas, normalmente variables ficticias con las que se intentan representar factores no cuantificables. Para tales variables, la evaluación del cambio en probabilidad debe hacerse teniendo en cuenta los cambios discretos de dicha variable y lo que interesa es obtener:

$$\frac{\Delta p_i}{\Delta x_j} = \frac{P(y_i = 1|x_j = 1) - P(y_i = 1|x_j = 0)}{1 - 0} \tag{5}$$

en otros términos, realizar un análisis de sensibilidad, calculando p_i para cada uno de los posibles valores adoptados por la variable discreta y obteniendo mediante su diferencia el cambio en la probabilidad de elección.

3. ESTRATEGIAS DE ANÁLISIS

En el caso de variables continuas, hemos aceptado que la interpretación de la derivada como razón entre la variación de p_i y la variación de x_j, cuando Δx_j tiende a cero, permite establecer que la variación absoluta de p_i vendrá expresada, aproximadamente, mediante $\frac{\partial p_i}{\partial x_j}\Delta x_j$ y tomando como «unidad de cambio» de x_j una unidad finita, pero coherente con el rango plausible, en los términos definidos anteriormente, de valores de la correspondiente variable explicativa, la falta de acotación de la derivada no supone ningún problema. Cuando las variables explicativas son de tipo discreto, la

variación de p_i se establece llevando a cabo lo que hemos denominado en el apartado anterior, un análisis de sensibilidad.

En cualquiera de las dos situaciones, la concreción de la variación en la probabilidad exige calcular $x_i'\beta$. El vector de parámetros se habrá estimado previamente, pero necesitamos decidir el valor que adoptamos para el conjunto de las variables explicativas y a este respecto no existe una estrategia única, pudiéndose encontrar en la literatura diferentes formas de actuación que pasamos a describir y analizar.

La opción más común consiste en efectuar la evaluación en el vector de valores medios de las diferentes variables explicativas, como puede verse, entre otros, en LeCleer (1992), Greene (1993) y Greene Knapp y Seaks (1992). En nuestra opinión, esta forma de actuación no resulta del todo adecuada debido, fundamentalmente, a dos hechos. En primer lugar, la evaluación del conjunto de variables explicativas en su valor medio, cuando como tales suele ser corriente que aparezcan variables ficticias, resulta incoherente con el sentido que se les otorga a dichas variables. La media en tal caso simplemente establece la proporción de una de las características sobre el total, pero no es un valor que la variable ficticia pueda adoptar y que tenga significado en dicho contexto. Por otra parte, utilizar la media muestral para las variables de tipo contínuo, puede plantear problemas de falta de representatividad en el caso de que los valores de una determinada variable estén muy dispersos. En cualquier caso, la cuestión relevante es el tratamiento eficiente de la información y, a este respecto, parece más adecuado utilizar el conjunto de datos relativos a cada unidad de decisión —obteniendo $\partial p_i/\partial x_j$ para cada una de las unidades de la muestra— y resumir el resultado en un solo valor, calculando entonces el valor medio. Se trata, en definitiva, de poner el énfasis en el comportamiento medio de la muestra antes que en el comportamiento de un «individuo medio», al objeto de evitar una pérdida de información sin ningún tipo de justificación. Esta postura parece ser también defendida por Train (1986), cuando plantea la inconsistencia de calcular la derivada en la media en lugar de la media de las derivadas, basándose en la no coincidencia, para modelos no lineales, entre la media de una función y el valor de la función en el punto medio.

No obstante, tal alternativa de evaluación no es posteriormente utilizada por el autor y su propuesta, al igual que la de otros autores como Jhonson (1987) y Maddala (1992), consiste en calcular $\partial p_i/\partial x_j$ para aquellos valores de las variables explicativas que sean más significativos, con el fin de obtener una idea sobre el rango de variación de los cambios resultantes en la probabilidad. Esta opción adolece de cierta ambigüedad por cuanto no se explicita lo que se consideran «valores más significativos» de las explicativas, ¿serían los valores de las x's más frecuentes en la muestra concreta?, ¿nos referimos a los valores máximo y mínimo de cada variable y sus posibles combinaciones? Train (1986) discute este aspecto teniendo en cuenta que el valor de la derivada será máximo para $p_i = 0'5$ (o equivalentemente, de acuerdo con (1), $x_i'\beta = 0$,

lo cual, si el modelo no contiene término independiente determina que todas las explicativas deben ser igual a cero y caso de existir término autónomo dará lugar a diferentes combinaciones de valores de x que satisfagan $x_i'\beta = 0$) y mínimo cuando $p_i = 0$ y $p_i = 1$ (lo que se corresponde respectivamente con $x_i'\beta = -\infty$ y $x_i'\beta = +\infty$, situaciones que sólo tienen interés como acotación, pero que son inalcanzables en la práctica) de tal modo que son estos tres puntos los que constituyen la propuesta de evaluación. El interés de esta estrategia como instrumento para informar sobre el comportamiento medio de la muestra es bastante escaso, puesto que se limita a establecer, exclusivamente, el rango de variación de la probabilidad tomando como referencia los puntos límite.

Fisher (1991) plantea una variación de esta última propuesta basada en la utilización de la media de la variable dicotómica de respuesta observada (y). En concreto, tal valor le permite, invirtiendo (1) de modo que $\widetilde{x_i'\beta} = F^{-1}(\bar{y})$, obtener la magnitud de $x_i'\beta$ a utilizar para evaluar la derivada. La idea de este autor resulta difícilmente justificable, por cuanto no considera el conjunto de información sintetizado en las variables explicativas del modelo, y en su lugar identifica a $\bar{y}$ como una estimación de la probabilidad de elección, lo que equivale a obviar por completo el marco teórico de razonamiento que subyace en la especificación de los MED.

Podemos decir, por tanto, que una característica común al conjunto de alternativas de evaluación analizadas, es la falta de un buen tratamiento de la información que permita dar cuenta del comportamiento medio de la muestra. En consecuencia, al objeto de plantear una vía adecuada para medir la variación en la probabilidad ante cambios de las variables explicativas, nuestra propuesta consiste en calcular la derivada para cada elemento de decisión y plasmar, posteriormente, el resultado en el valor medio de tales derivadas. De esta manera, se está utilizando el conjunto de información muestral del que realmente se dispone y se evitan las posibles inconsistencias asociadas con un uso parcial de la información.

Hay que tener presente, no obstante, que esta propuesta de análisis solo es válida para variables de tipo continuo, de manera que el estudio referido a las posibles variables ficticias que integren el conjunto de factores explicativos deberá hacerse utilizando la expresión (5). El cálculo de dicha expresión ha consistido, normalmente, en fijar las variables continuas en el nivel asociado con el valor medio de las mismas. En nuestra opinión, y de acuerdo con la propuesta de estudiar el «comportamiento medio de la muestra», nos parece más coherente obtener la probabilidad p_i para cada elemento de decisión y calcular la media para cada submuestra concreta distinguida por el valor asociado a la variable ficticia, de modo que la diferencia entre estas medias nos dará el cambio en probabilidad ante un cambio en el valor de la variable ficticia.

La metodología propuesta para analizar la variación de la probabilidad ante cambios en las variables explicativas se fundamenta, por tanto, en la idea de un uso eficiente

de la información muestral disponible que permita dar cuenta, de manera coherente, del comportamiento medio de la muestra.

4. EJERCICIO DE SIMULACIÓN

En el apartado anterior hemos revisado, en primer lugar, las distintas estrategias relativas a la concreción de la variación de la probabilidad y los problemas que, en nuestra opinión llevan asociados, lo cual nos ha permitido justificar, posteriormente, una opción alternativa.

Al trabajar en el contexto de funciones no lineales, se verifica que la media de una función de este tipo no va a coincidir con la función evaluada en el punto medio; en consecuencia, la alternativa de actuación propuesta en este trabajo, fundamentada en el cálculo de la media de las derivadas, no va a ser igual a la comúnmente utilizada, basada en la derivada en el punto medio. Por ello, resulta de interés efectuar un análisis de ambas medidas al objeto de poder determinar, si es posible, la superioridad de una frente a la otra de acuerdo con algún tipo de criterio.

Un aspecto previo a la cuestión central mencionada consistirá en estudiar si la diferencia que va a producirse entre ambas estrategias tiene, en la mayoría de los casos, el mismo sentido o, por el contrario, no responde a ninguna pauta fija. En otros términos, se intenta verificar si existe un posible patrón de comportamiento para las dos alternativas planteadas, lo cual, aunque no permite establecer preferencias entre ellas, sí que parece, en el caso de encontrarse, un resultado más satisfactorio.

Con este objeto se ha diseñado un ejercicio de simulación que nos permitirá, utilizando como proceso generador de los datos (PGD) bien un modelo probit o bien un modelo logit, generar diferentes muestras. En concreto, generaremos 100 muestras —es decir, se efectúan 100 replicaciones— de tamaño 100 cada una de ellas.

El procedimiento de generación está basado en los trabajos de Gourieroux, Monfort, Renault y Trognon (1987) y Griffiths, Carter y Pope (1987). En el primero de ellos se adopta como punto de partida la justificación teórica de los MED y, más concretamente, la relación lineal que se establece entre el índice subjetivo (no observable) denominado de utilidad o de satisfacción correspondiente al individuo $i-$ésimo (y_i^*) y el conjunto de características socio-económicas que influyen en el problema concreto, pudiendo escribir:

$$y_i^* = x_i'\beta + u_i$$

de modo que si la función de distribución asumida para u_i es $N(0,1)$ especificaremos, en última instancia, un modelo probit y caso de suponer una distribución logística resultará un modelo logit.

Suponiendo que el vector x_i', además del término independiente, está integrado por dos variables que suponemos cuantitativas, y que denotaremos por x_1 y x_2, el proceso seguido consiste en generar éstas como $N(0,1)$, de igual modo que el término de perturbación y fijar los parámetros igual a la unidad resultando, por tanto, un primer «índice subjetivo» dado por:

$$y_i^* = 1 + x_{1i} + x_{2i} + u_i$$

De esta manera, el conocimiento del conjunto de observaciones de y_i^*, x_{1i} y x_{2i}, permite estimar por máxima verosimilitud el modelo:

$$y_i^* = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i \tag{6}$$

A partir de esta expresión, utilizando el vector de coeficientes estimados y una muestra de tamaño 100 de x_1, x_2 y del término de perturbación (que también hemos asumido $N(0,1)$), generamos la correspondiente muestra de tamaño 100 para el índice subjetivo (y_i^*). Como lo que realmente interesa es el conocimiento de la respuesta del individuo, esto es, la obtención de la variable dicotómica representada como y_i, deberemos hacer uso de la conocida relación entre y_i^* e y_i que podemos concretar como:

$$y_i = \begin{cases} 1 & \text{si} \quad y_i^* > 0 \\ 0 & \text{si} \quad y_i^* \leqslant 0 \end{cases}$$

y, de este modo, obtenemos una muestra de tamaño 100 para y_i. Este proceso, tal como se ha mencionado anteriormente, se repite 100 veces, es decir, establecemos el número de réplicas igual a 100.

En el segundo de los trabajos citados se plantea una forma alternativa de generación de los datos según la cual, fijados los parámetros y las variables explicativas, de acuerdo con los mismos supuestos establecidos anteriormente, se obtiene $p_i = F(x_i'\beta)$. A continuación se genera una variable uniforme en el intervalo (0,1) que denominamos como v_i y la variable dicotómica se obtiene de acuerdo con:

$$y_i = \begin{cases} 1 & \text{si} \quad v_i \in [0, P_i] \\ 0 & \text{si} \quad v_i \in (P_i, 1] \end{cases}$$

De este modo, la consideración de diferentes variables v_i permite obtener las diversas muestras de la variable dicotómica.

Griffiths, Carter y Pope (1987) desarrollan este procedimiento en el marco del modelo probit, aunque su aplicación para el modelo logit resulta inmediata, por cuanto sólo requiere identificar F con la función de distribución logística en la obtención de p_i.

En la generación de los modelos probit hemos aplicado el primero de los procedimientos expuestos comprobando, adicionalmente, que las conclusiones obtenidas no variaban cuando tales modelos eran generados de acuerdo con el segundo de los métodos. Esta similitud nos lleva a utilizar el planteamiento de Griffiths, Carter y Pope (1987), que requiere menor complejidad de cálculo, para la generación de los modelos logit. En ambos casos, el programa informático utilizado ha sido Gauss.

Descrito el proceso de generación de los diferentes modelos, podemos pasar a verificar la cuestión planteada en cuanto a la existencia o no de un patrón de comportamiento estable entre las dos medidas objeto de análisis. Denominando D_K $(K = 1,2)$ a la derivada en el punto medio y DM_K $(K = 1,2)$ a la media de las derivadas, las tablas 1 y 2 del Apéndice presentan ambas medidas para el conjunto de modelos probit y logit, respectivamente. La primera columna de estas tablas recoge la réplica correspondiente, las columnas segunda y cuarta denotadas como D_1 y D_2 presentan la derivada en el punto medio, es decir, el instrumento tradicionalmente utilizado en la literatura para medir el cambio en probabilidad ante una variación unitaria de la respectiva variable explicativa (en este sentido, D_1 se asocia con x_1 y D_2 con x_2) y las otras dos columnas, DM_1 y DM_2, sintetizan la propuesta de medición efectuada basada en la media de las derivadas.

Centrándonos en la tabla 1 podemos concluir que de las 100 replicaciones sólo en cuatro de ellas (las numeradas como 7, 37, 49 y 79) la desigualdad obtenida es $D_K < DM_K$ $(K = 1,2)$; para el modelo logit (tabla 2), la anterior desigualdad no se presenta en ninguna ocasión. Por tanto, podemos establecer que de 100 casos —independientemente de que se trate de un modelo probit o logit— la mayoría de los mismos va a satisfacer que $D_K > DM_K$, lo que podemos entender como existencia de una pauta de comportamiento bastante sistemática entre ambas medidas.

Al objeto de poder aportar argumentos a favor de alguna de las medidas, parece razonable pensar en términos del criterio de «robustez». La robustez es una propiedad deseable de un buen estimador y determina, caso de cumplirse, que tal estimador sigue siendo razonablemente bueno ante pequeñas desviaciones del modelo.

En nuestro caso concreto, diremos que una medida es más robusta que la otra cuando ante pequeños cambios en la información muestral la primera altera su valor en menor medida que la segunda; en otros términos, tendremos más confianza en una medida cuyo comportamiento viene confirmado por toda la muestra que en otra que responda sólo a un subconjunto de individuos.

Para estudiar esta cuestión nos planteamos analizar, para cada muestra, la influencia que la eliminación de un individuo tiene sobre el valor de D_K o de DM_K en relación con el valor adoptado por la medida cuando hemos considerado $N = 100$. En concreto, se va eliminando cada vez un elemento distinto de la muestra, pasando a trabajar con muestras de tamaño 99, posteriormente, se compara el valor obtenido de D_K y DM_K,

para cada muestra de tamaño 99, con el que denominamos valor de referencia que es el correspondiente a la muestra completa de tamaño 100.

Una forma adecuada de resumir las variaciones sufridas por D_K o DM_K, ante las pequeñas variaciones muestrales que hemos planteado, es atender a la dispersión o variación de las mismas calculada, para cada modelo, como:

$$(7.a) \qquad DD_i = \frac{\sum_{i=1}^{100} (D_{Ki} - VR_K)^2}{100} \qquad (K = 1,2)$$

$$(7.b) \qquad DDM_i = \frac{\sum_{i=1}^{100} (DM_{Ki} - VR'_K)^2}{100}$$

donde VR y VR' denotan los valores de referencia mencionados que indican, respectivamente, el valor de la derivada en el punto medio y la media de las derivadas para la muestra completa.

La aplicación de las expresiones dadas en (7.a) y (7.b) se efectua tanto para el enfoque probit como para el logit y las concreciones de las mismas aparecen, respectivamente, en las tablas 3 y 4 del Apéndice. Tales tablas presentan para cada réplica, numerada en la primera columna, los valores de (7.a) —columnas segunda y cuarta— para las dos variables explicativas que conforman la parte sistemática de cada modelo y, análogamente, la cuantificación de (7.b) aparece en las dos columnas restantes.

Tanto en el caso probit como en el logit se comprueba cómo, para todas las réplicas efectuadas, la medida D_K presenta una mayor dispersión —reflejada en el valor de DD_K— que DM_K.

Al objeto de resumir todo este conjunto de información, podemos calcular la media de estas dispersiones, es decir, hallar el promedio de cada una de las columnas que componen las tablas 3 y 4 y, en este sentido, definimos:

$$(8) \qquad PDD_K = \frac{\sum_{i=1}^{100} DD_i}{100} \qquad PDDM_K = \frac{\sum_{i=1}^{100} DDM_i}{100} \qquad (K = 1,2)$$

como los mencionados valores promedio, lo que nos permite obtener el conjunto de resultados siguientes:

	PDD_1	$PDDM_1$	PDD_2	$PDDM_2$
Probit	0'000066838	0'000015144	0'000072187	0'000014424
Logit	0'000042917	0'000015422	0'000048772	0'000017606

que reflejan, claramente, cómo para el caso probit, la dispersión de la medida basada en la derivada en el punto medio es en torno a $4'5-5$ veces la correspondiente a la propuesta alternativa, mientras que para el caso logit, esta relación resulta ser de $2'7$ veces[2].

Por tanto, atendiendo al criterio de robustez, podemos concluir que la propuesta defendida en este trabajo presenta una mayor estabilidad ante pequeños cambios en la información muestral y, en este sentido, resulta superior a la estrategia de actuación que habitualmente se viene utilizando para analizar las variaciones en la probabilidad de elección.

5. EJEMPLO: MODELOS PROBIT Y LOGIT APLICADOS A DATOS ELECTORALES

Con el fin de contrastar el comportamiento que, del ejercicio de simulación anteriormente desarrollado, se derivaba para las medidas D_K y DM_K, consideramos el siguiente caso real extraído de Aznar, García-Ferrer y Martín (1994). El estudio tiene como objetivo predecir la elección de los votantes sobre la construcción de un colegio local en Troy (Michigan) en el año 1973. La muestra está constituida por 95 personas y el conjunto de variables explicativas consideradas incluye las siguientes:

X1: 1 si tiene uno o dos hijos en colegio público, 0 en otro caso.

X2: 1 si tiene tres o cuatro hijos en colegio público, 0 en otro caso.

X3: 1 si tiene cinco o más hijos en colegio público, 0 en otro caso.

X4: 1 si tiene uno o más hijos en colegio privado, 0 en otro caso.

X5: número de años viviendo en la comunidad de Troy.

X6: 1 si el individuo es profesor (público o privado), 0 en otro caso.

X7: logaritmo del ingreso familiar anual.

X8: logaritmo del impuesto sobre bienes inmuebles pagado anualmente.

La variable de respuesta y_i se define como 1 si el individuo vota afirmativamente y 0 en otro caso.

En base a este conjunto de información se ajusta tanto un modelo logit como un modelo probit resultado:

[2]Resultados similares son obtenidos si razonamos, alternativamente, formando en base a las tablas 3 y 4 el cociente de dispersiones para cada réplica y cada x_i y resumiendo los 100 cocientes resultantes (para cada una de las variables explicativas) en su valor medio.

Modelo logit						
Variables	Coeficientes estimados	D_K	DM_K	DD_K	DDM_K	DD_K/DDM_K
Término constante	$-5'20248$	–	–	–	–	–
X1	$0'58323$	–	–	–	–	–
X2	$1'12558$	–	–	–	–	–
X3	$0'52536$	–	–	–	–	–
X4	$-0'3415$	–	–	–	–	–
X5	$-0'026114$	$-0'0057961$	$-0'0050319$	$9'1104 \cdot 10^{-7}$	$6'31531 \cdot 10^{-7}$	$1'44259$
X6	$2'62739$	–	–	–	–	–
X7	$2'18742$	$0'48551$	$0'4215$	$0'00036434$	$0'00022363$	$1'62924$
X8	$-2'39464$	$-0'53151$	$-0'46143$	$0'00063372$	$0'00040738$	$1'55561$

Modelo probit						
Variables	Coeficientes estimados	D_K	DM_K	DD_K	DDM_K	DD_K/DDM_K
Término constante	$-2'95916$	–	–	–	–	–
X1	$0'36807$	–	–	–	–	–
X2	$0'69117$	–	–	–	–	–
X3	$0'29522$	–	–	–	–	–
X4	$-0'21115$	–	–	–	–	–
X5	$-0'015761$	$-0'0057524$	$-0'0049942$	$7'21438 \cdot 10^{-7}$	$5'09011 \cdot 10^{-7}$	$1'41733$
X6	$1'58427$	–	–	–	–	–
X7	$1'31437$	$0'47971$	$0'41648$	$0'00035812$	$0'00021799$	$1'64279$
X8	$-1'46417$	$-0'53438$	$-0'46395$	$0'00058285$	$0'00037707$	$1'5457$

Notar que el conjunto de medidas analizadas en este trabajo sólo se ha calculado para las variables continuas, ya que únicamente eran este tipo de variables las que hemos considerado en el experimento de Monte Carlo.

La consideración de ambos modelos refleja que, en todos los casos, se satisface $|DM_K| < |D_K|$ lo cual coincide con la conclusión que, a este respecto, se desprendía del ejercicio de simulación. En cuanto a la dispersión de ambas medidas (DD_K y DDM_K) se mantiene la pauta de comportamiento esperada por cuanto $DD_K > DDM_K$ y ello se puede interpretar como evidencia a favor de la mayor «robustez» de esta última. Por último, con la consideración del cociente entre DD_K y DDM_K queremos proporcionar una cierta idea sobre la cuantía en que la dispersión de la medida

habitualmente considerada supera a la propuesta efectuada; estos ratios también son coherentes con la conducta derivada en este sentido al trabajar con datos simulados.

6. CONCLUSIONES

El objetivo planteado en este trabajo ha consistido en efectuar algunas reflexiones sobre cómo analizar, en el marco de los modelos dicotómicos, la variación en la probabilidad de elección ante variaciones en una variable explicativa. Para ello, la discusión se ha centrado básicamente en dos cuestiones: en primer lugar, cuál debe ser el instrumento a utilizar y, posteriormente, cómo debe efectuarse la concreción de dicho instrumento.

En lo que respecta a la primera cuestión, se ha defendido que el cambio en probabilidad debe aproximarse a través de $\frac{\partial p_i}{\partial x_j}\Delta x_j$ para el caso de variables continuas, rechazando las críticas que algunos autores achacan a esta medida, en base a que el denominado problema de acotación no va a existir si se hace una interpretación correcta. Adicionalmente, en el supuesto de variables discretas el instrumento a utilizar es el análisis de sensibilidad.

En cuanto a la medición concreta del instrumento, en el apartado 3 se han tratado de sintetizar las diversas opciones existentes y se ha efectuado una valoración de cada una de ellas en términos del «uso eficiente de la información»; los puntos débiles detectados en las propuestas tradicionales se han tratado de subsanar planteando una concreción alternativa que atiende al valor medio de las derivadas, para el caso de variables continuas y, si se opera con variables ficticias, se defiende un análisis de sensibilidad en base a la media de cada submuestra concreta.

Establecido el marco de razonamiento alternativo, interesa profundizar en su comportamiento en relación a la opción comúnmente utilizada, basada en proceder a la cuantificación en la media de las variables. Para ello se ha planteado un ejercicio de simulación, que ha permitido verificar en primera instancia la existencia de una pauta de desenvolvimiento bastante sistemática entre ambas alternativas, por cuanto de los 100 casos analizados la casi totalidad de los mismos satisfacen $D_K > DM_K$ $(K = 1,2)$, lo cual aunque, como ya se ha mencionado, no aporta ningún tipo de evidencia sobre la superioridad de una medida frente a la otra, sí que parece resultar más satisfactorio que el hecho de no haberse podido llegar a establecer ningún tipo de regularidad.

Planteado el tema en términos de criterios que permitan avalar la decantación hacia una de las estrategias, nos ha parecido oportuno razonar en base al criterio de robustez, el cual es ampliamente defendido en la literatura. El análisis de la misma para el

marco de simulación planteado, nos permite concluir con bastante claridad que la propuesta alternativa expuesta en este trabajo, que entendemos constituye una vía coherente de determinar el comportamiento medio de la muestra, es más robusta que la habitualmente utilizada.

Las pautas de comportamiento derivadas del ejercicio de simulación se ven refrendadas en el caso real planteado como ejemplo. En definitiva, el cumplimiento de la propiedad de «robustez» por parte de la estrategia de actuación propuesta, es un resultado alentador para proseguir con el estudio de la misma y profundizar en otros aspectos de su comportamiento.

APÉNDICE

Tabla 1. *Modelo probit*

RÉPLICA	D_1	DM_1	D_2	DM_2
1	0.29938	0.20887	0.34193	0.23856
2	0.25206	0.19267	0.25817	0.19734
3	0.26402	0.21164	0.16769	0.13442
4	0.20228	0.17255	0.18783	0.16022
5	0.20141	0.15969	0.23990	0.19021
6	0.28364	0.22509	0.26436	0.20979
7	0.26788	0.26918	0.18100	0.18188
8	0.38533	0.27581	0.27001	0.19327
9	0.32265	0.21458	0.38220	0.25418
10	0.34428	0.23993	0.33918	0.23637
11	0.33830	0.22307	0.37752	0.24893
12	0.25637	0.20564	0.19293	0.15475
13	0.37863	0.24019	0.42532	0.26980
14	0.19163	0.16002	0.21051	0.17579
15	0.33915	0.24698	0.29297	0.21336
16	0.31129	0.21825	0.33449	0.23451
17	0.24867	0.21968	0.22481	0.19861
18	0.21752	0.17876	0.22163	0.18214
19	0.27778	0.23311	0.23363	0.19606
20	0.23542	0.17427	0.31659	0.23435
21	0.26785	0.18142	0.37703	0.25538
22	0.13649	0.11501	0.24746	0.20851
23	0.44524	0.26938	0.39697	0.24018
24	0.19847	0.19253	0.20090	0.19488
25	0.24605	0.22545	0.19378	0.17755
26	0.19499	0.16705	0.19152	0.16407
27	0.22320	0.17766	0.24625	0.19600
28	0.29108	0.20306	0.33456	0.23339
29	0.14784	0.13152	0.24129	0.21466
30	0.28622	0.20994	0.24379	0.17882

Tabla 1. (cont.) *Modelo probit*

RÉPLICA	D_1	DM_1	D_2	DM_2
31	0.31340	0.21262	0.33465	0.22703
32	0.29649	0.21712	0.28364	0.20771
33	0.096313	0.087231	0.22809	0.20658
34	0.27648	0.20863	0.24601	0.18563
35	0.26712	0.21922	0.24842	0.20387
36	0.28245	0.24792	0.19531	0.17143
37	0.11215	0.13527	0.18557	0.22382
38	0.39264	0.25873	0.31547	0.20788
39	0.34028	0.25137	0.26449	0.19539
40	0.30951	0.22780	0.23517	0.17308
41	0.29030	0.21551	0.25986	0.19291
42	0.35914	0.23747	0.37139	0.24557
43	0.36605	0.25654	0.25050	0.17556
44	0.28139	0.23291	0.24061	0.19916
45	0.29130	0.20350	0.38184	0.26675
46	0.28067	0.20847	0.29261	0.21734
47	0.24974	0.18072	0.31269	0.22628
48	0.27215	0.21252	0.30426	0.23760
49	0.17140	0.17941	0.15079	0.15784
50	0.22777	0.15715	0.36917	0.25471
51	0.27075	0.18425	0.36210	0.24642
52	0.25616	0.16060	0.44682	0.28013
53	0.33707	0.24286	0.32702	0.23561
54	0.30920	0.21593	0.30580	0.21356
55	0.23494	0.18705	0.23443	0.18664
56	0.26018	0.18247	0.33154	0.23251
57	0.21808	0.19425	0.20238	0.18026
58	0.29230	0.23741	0.25878	0.21019
59	0.28221	0.21273	0.24903	0.18772
60	0.28184	0.20389	0.28269	0.20450
61	0.22339	0.17059	0.29902	0.22834

Tabla 1. (cont.) *Modelo probit*

RÉPLICA	D_1	DM_1	D_2	DM_2
62	0.22710	0.18496	0.23454	0.19102
63	0.30099	0.21268	0.31603	0.22330
64	0.34976	0.24539	0.26866	0.18849
65	0.30349	0.21257	0.30550	0.21397
66	0.33462	0.21400	0.39545	0.25290
67	0.20487	0.17068	0.17282	0.14397
68	0.29225	0.21871	0.23739	0.17765
69	0.42138	0.28620	0.25751	0.17490
70	0.19093	0.18225	0.18985	0.18122
71	0.25454	0.19239	0.25316	0.19134
72	0.14537	0.14845	0.19110	0.19515
73	0.18628	0.16214	0.21552	0.18759
74	0.33545	0.23216	0.37501	0.25954
75	0.20817	0.16935	0.20457	0.16642
76	0.23933	0.15604	0.42059	0.27422
77	0.34932	0.24054	0.33695	0.23203
78	0.27095	0.20161	0.25842	0.19228
79	0.24827	0.17884	0.34808	0.25074
80	0.20303	0.15594	0.28174	0.21640
81	0.26513	0.17136	0.42101	0.27210
82	0.38170	0.26500	0.25691	0.17836
83	0.30005	0.20460	0.34534	0.23548
84	0.25534	0.21587	0.27301	0.23081
85	0.26048	0.23819	0.25182	0.23028
86	0.32725	0.23696	0.23745	0.17194
87	0.22519	0.17173	0.28059	0.21398
88	0.24001	0.18093	0.29534	0.22264
89	0.21894	0.17112	0.27758	0.21696
90	0.28828	0.19819	0.33403	0.22964
91	0.16827	0.14208	0.22265	0.18799
92	0.36031	0.22679	0.39272	0.24719

Tabla 1. (cont.) *Modelo probit*

RÉPLICA	D_1	DM_1	D_2	DM_2
93	0.21183	0.18061	0.19200	0.16370
94	0.22954	0.16713	0.31868	0.23203
95	0.24285	0.19125	0.22323	0.17581
96	0.21608	0.16085	0.29430	0.21908
97	0.36163	0.25448	0.31262	0.21999
98	0.15177	0.12168	0.26914	0.21578
99	0.24362	0.21548	0.18457	0.16325
100	0.30931	0.22273	0.31963	0.23016

Tabla 2. *Modelo logit*

RÉPLICA	D_1	DM_1	D_2	DM_2
1	0.12643	0.11286	0.13344	0.11912
2	0.29087	0.20675	0.27438	0.19503
3	0.23200	0.17188	0.27366	0.20275
4	0.23012	0.18725	0.11930	0.097080
5	0.21939	0.18114	0.23800	0.19651
6	0.34125	0.23440	0.28322	0.19453
7	0.23245	0.18256	0.20433	0.16048
8	0.25008	0.18073	0.28744	0.20773
9	0.21919	0.18021	0.15098	0.12413
10	0.21716	0.17558	0.19102	0.15445
11	0.19225	0.15137	0.23253	0.18308
12	0.20060	0.15690	0.23138	0.18098
13	0.21578	0.18116	0.13520	0.11351
14	0.22882	0.17747	0.25494	0.19774
15	0.11067	0.097809	0.16274	0.14383
16	0.24914	0.19021	0.23022	0.17577
17	0.19494	0.16411	0.18147	0.15277
18	0.10331	0.089947	0.17714	0.15422
19	0.11263	0.093074	0.23212	0.19182
20	0.21868	0.15820	0.30157	0.21816
21	0.30306	0.21066	0.28941	0.20117
22	0.16894	0.13928	0.18911	0.15590
23	0.25055	0.20153	0.13847	0.11138
24	0.16265	0.13441	0.19986	0.16515
25	0.25603	0.17697	0.33304	0.23020
26	0.27135	0.20785	0.18520	0.14186
27	0.22403	0.18509	0.14285	0.11802
28	0.16758	0.13611	0.23552	0.19129
29	0.16596	0.14228	0.18637	0.15977
30	0.23107	0.19726	0.069065	0.058961

Tabla 2. (cont.) *Modelo logit*

RÉPLICA	D_1	DM_1	D_2	DM_2
31	0.16654	0.14427	0.13193	0.11428
32	0.13024	0.10862	0.20762	0.17315
33	0.37474	0.23885	0.40485	0.25803
34	0.19035	0.15690	0.17847	0.14710
35	0.14242	0.12405	0.18896	0.16457
36	0.15076	0.11244	0.30437	0.22700
37	0.22316	0.17115	0.23151	0.17756
38	0.29836	0.21477	0.21728	0.15641
39	0.19806	0.15651	0.27506	0.21736
40	0.17457	0.15082	0.12499	0.10799
41	0.19443	0.16416	0.14038	0.11853
42	0.24458	0.18184	0.27108	0.20155
43	0.22461	0.17998	0.18147	0.14542
44	0.17199	0.13725	0.25424	0.20289
45	0.086481	0.077286	0.16440	0.14692
46	0.25485	0.20695	0.11735	0.095294
47	0.25133	0.18906	0.22534	0.16951
48	0.19168	0.15864	0.17106	0.14158
48	0.15583	0.13349	0.16681	0.14290
50	0.23465	0.18464	0.17562	0.13819
51	0.15666	0.13493	0.16323	0.14059
52	0.28036	0.20776	0.21292	0.15778
53	0.25333	0.19686	0.15819	0.12292
54	0.22715	0.18338	0.19457	0.15707
55	0.36606	0.24552	0.27974	0.18763
56	0.17666	0.14286	0.21868	0.17683
57	0.30034	0.20461	0.31595	0.21524
58	0.15569	0.12452	0.24297	0.19433
59	0.28016	0.19414	0.33138	0.22963
60	0.24730	0.17429	0.31898	0.22480
61	0.16627	0.14711	0.17176	0.15196

Tabla 2. (cont.) *Modelo logit*

RÉPLICA	D_1	DM_1	D_2	DM_2
62	0.14556	0.12728	0.13708	0.11986
63	0.16860	0.13843	0.21021	0.17259
64	0.19107	0.15952	0.16905	0.14113
65	0.39247	0.26247	0.10689	0.071482
66	0.15087	0.13865	0.095620	0.087875
67	0.19403	0.15919	0.20832	0.17092
68	0.24857	0.20178	0.13518	0.10973
69	0.22022	0.16320	0.27835	0.20627
70	0.18310	0.15186	0.16338	0.13551
71	0.17588	0.14123	0.21507	0.17270
72	0.22089	0.17112	0.24111	0.18678
73	0.24333	0.18565	0.21380	0.16312
74	0.23589	0.18036	0.24787	0.18951
75	0.13953	0.12078	0.16696	0.14452
76	0.19747	0.14911	0.27977	0.21126
77	0.24395	0.19153	0.18290	0.14359
78	0.22249	0.17105	0.23159	0.17804
79	0.22160	0.17565	0.17965	0.14240
80	0.14873	0.12870	0.15971	0.13821
81	0.18939	0.15049	0.22113	0.17571
82	0.27033	0.21004	0.16652	0.12938
83	0.17161	0.14605	0.15652	0.13321
84	0.24429	0.18261	0.25272	0.18891
85	0.18912	0.13703	0.34571	0.25049
86	0.27394	0.17280	0.42165	0.26598
87	0.14242	0.12101	0.17518	0.14885
88	0.27106	0.19826	0.27452	0.20079
89	0.17779	0.15893	0.10390	0.092873
90	0.21918	0.17953	0.13911	0.11394
91	0.14962	0.13435	0.081254	0.072964
92	0.32357	0.21130	0.37521	0.24501

Tabla 2. (cont.) *Modelo logit*

RÉPLICA	D_1	DM_1	D_2	DM_2
93	0.20552	0.16392	0.20170	0.16087
94	0.21319	0.16788	0.20840	0.16411
95	0.24501	0.17748	0.29683	0.21502
96	0.28377	0.18461	0.38142	0.24813
97	0.20760	0.16430	0.21772	0.17231
98	0.14742	0.13027	0.12265	0.10838
99	0.23483	0.18213	0.19721	0.15295
100	0.15070	0.13492	0.11110	0.099465

Tabla 3. *Modelo probit. Medida de dispersión*

RÉPLICA	DD_1	DDM_1	DD_2	DDM_2
1	0.6623200E-04	0.1496600E-04	0.9964000E-04	0.1868400E-04
2	0.5145000E-04	0.1785500E-04	0.5155200E-04	0.1561800E-04
3	0.3810700E-04	0.1158600E-04	0.2999600E-04	0.1424000E-04
4	0.3540700E-04	0.1661700E-04	0.3412500E-04	0.1663700E-04
5	0.4137200E-04	0.1840100E-04	0.4152400E-04	0.1411200E-04
6	0.7812200E-04	0.1663600E-04	0.7927500E-04	0.1533200E-04
7	0.1941900E-03	0.1675100E-04	0.9422800E-04	0.1220300E-04
8	0.1483000E-03	0.1498900E-04	0.8937400E-04	0.1867900E-04
9	0.7918500E-04	0.1491800E-04	0.1076500E-03	0.1509300E-04
10	0.1015600E-03	0.1528500E-04	0.1016500E-03	0.1499000E-04
11	0.8056800E-04	0.1491600E-04	0.1063100E-03	0.1441600E-04
12	0.4438700E-04	0.2046300E-04	0.3136700E-04	0.1180000E-04
13	0.1240500E-03	0.1441400E-04	0.1590800E-03	0.1172800E-04
14	0.2436000E-04	0.9411000E-05	0.3266500E-04	0.1183900E-04
15	0.9469000E-04	0.1405700E-04	0.8387400E-04	0.1413600E-04
16	0.7340500E-04	0.1370300E-04	0.8433300E-04	0.1398800E-04
17	0.7819400E-04	0.1284500E-04	0.7175900E-04	0.1754900E-04
18	0.3657600E-04	0.1308700E-04	0.4141600E-04	0.1412600E-04
19	0.9208500E-04	0.1623300E-04	0.7832800E-04	0.1746400E-04
20	0.4801400E-04	0.1387800E-04	0.8179400E-04	0.1611400E-04
21	0.5698300E-04	0.1569600E-04	0.8038100E-04	0.1257700E-04
22	0.2142900E-04	0.1220200E-04	0.4716300E-04	0.1233100E-04
23	0.1070200E-03	0.1396500E-04	0.1089800E-03	0.1645900E-04
24	0.5854300E-04	0.1913000E-04	0.6019100E-04	0.1855400E-04
25	0.1001800E-03	0.1590100E-04	0.6919700E-04	0.1584700E-04
26	0.2548600E-04	0.1083200E-04	0.2821200E-04	0.9785700E-05
27	0.4230300E-04	0.1586100E-04	0.4408400E-04	0.1298700E-04
28	0.4563500E-04	0.9041200E-05	0.7309100E-04	0.1470800E-04
29	0.2912100E-04	0.8938400E-05	0.6485900E-04	0.9873100E-05
30	0.3483300E-04	0.9293600E-05	0.4100200E-04	0.1379400E-04

Tabla 3. (cont.) *Modelo probit. Medida de dispersión*

RÉPLICA	DD_1	DDM_1	DD_2	DDM_2
31	0.7699500E-04	0.1632300E-04	0.9149700E-04	0.1814600E-04
32	0.5513500E-04	0.1266000E-04	0.5947500E-04	0.1272700E-04
33	0.2274500E-04	0.1225900E-04	0.4637200E-04	0.1267100E-04
34	0.4111600E-04	0.1025500E-04	0.5019800E-04	0.1600300E-04
35	0.7461000E-04	0.1480100E-04	0.6726900E-04	0.1242700E-04
36	0.1045300E-03	0.2178900E-04	0.6044800E-04	0.2351200E-04
37	0.4910400E-04	0.1139200E-04	0.1356700E-03	0.1349800E-04
38	0.8339800E-04	0.1164900E-04	0.7375900E-04	0.1452700E-04
39	0.9536600E-04	0.1515800E-04	0.6193600E-04	0.1318700E-04
40	0.4688900E-04	0.1086300E-04	0.3594100E-04	0.1050800E-04
41	0.4974100E-04	0.1214900E-04	0.4647800E-04	0.1332500E-04
42	0.1041300E-03	0.1766100E-04	0.1143200E-03	0.1556300E-04
43	0.7336200E-04	0.1131700E-04	0.4870400E-04	0.1338500E-04
44	0.9079900E-04	0.1264600E-04	0.7108800E-04	0.1318600E-04
45	0.9418500E-04	0.1686800E-04	0.1587700E-03	0.1304000E-04
46	0.5438100E-04	0.1274900E-04	0.6397100E-04	0.1203700E-04
47	0.5484200E-04	0.1761700E-04	0.5627100E-04	0.1108400E-04
48	0.9550100E-04	0.1666000E-04	0.1197400E-03	0.1768500E-04
49	0.4465900E-04	0.1289300E-04	0.3840000E-04	0.1133900E-04
50	0.4802500E-04	0.1675600E-04	0.7789200E-04	0.1123500E-04
51	0.4787300E-04	0.1160900E-04	0.9292300E-04	0.1648600E-04
52	0.6761400E-04	0.1639800E-04	0.9443300E-04	0.1022000E-04
53	0.1148600E-03	0.1269500E-04	0.1110500E-03	0.1740300E-04
54	0.5111600E-04	0.1240800E-04	0.4993200E-04	0.1099100E-04
55	0.3782100E-04	0.1226600E-04	0.4430200E-04	0.1521000E-04
56	0.4474300E-04	0.1155900E-04	0.6105500E-04	0.1207500E-04
57	0.5030100E-04	0.1434000E-04	0.5402500E-04	0.1439600E-04
58	0.1003200E-03	0.1780600E-04	0.8739200E-04	0.1463700E-04
59	0.5476700E-04	0.1483600E-04	0.5350100E-04	0.1677300E-04
60	0.4867400E-04	0.1433900E-04	0.4175800E-04	0.9769500E-05
61	0.4921000E-04	0.1678600E-04	0.7251400E-04	0.1589300E-04

Tabla 3. (cont.) *Modelo probit. Medida de dispersión*

RÉPLICA	DD_1	DDM_1	DD_2	DDM_2
62	0.3909300E-04	0.1348100E-04	0.4485400E-04	0.1005300E-04
63	0.5543200E-04	0.1204400E-04	0.6086800E-04	0.1168200E-04
64	0.6414100E-04	0.1153900E-04	0.5335000E-04	0.1359700E-04
65	0.4996100E-04	0.1235500E-04	0.5755300E-04	0.1384800E-04
66	0.7529100E-04	0.1287700E-04	0.1303000E-03	0.2008500E-04
67	0.2467600E-04	0.1014500E-04	0.2444800E-04	0.1125400E-04
68	0.4545100E-04	0.1190400E-04	0.4765400E-04	0.1613000E-04
69	0.1510200E-03	0.1625600E-04	0.7534000E-04	0.1532100E-04
70	0.4562300E-04	0.1339100E-04	0.5103200E-04	0.1431500E-04
71	0.5684400E-04	0.1953500E-04	0.5523600E-04	0.1801800E-04
72	0.4474600E-04	0.1095500E-04	0.6879100E-04	0.1102900E-04
73	0.3381200E-04	0.1464100E-04	0.4533900E-04	0.1365900E-04
74	0.1223000E-03	0.1897600E-04	0.1568100E-03	0.1312300E-04
75	0.3535300E-04	0.1576600E-04	0.3626800E-04	0.1477400E-04
76	0.7060100E-04	0.1869100E-04	0.1177800E-03	0.1338600E-04
77	0.9684800E-04	0.1597900E-04	0.1009900E-03	0.1598300E-04
78	0.4084800E-04	0.1153600E-04	0.4896200E-04	0.1490900E-04
79	0.5668100E-04	0.1254700E-04	0.9856500E-04	0.1166300E-04
80	0.2917300E-04	0.9739300E-05	0.5411200E-04	0.1379400E-04
81	0.7697900E-04	0.2554500E-04	0.1042400E-03	0.1268800E-04
82	0.8923200E-04	0.1252800E-04	0.5220600E-04	0.1082700E-04
83	0.5253100E-04	0.1185900E-04	0.7515800E-04	0.1452000E-04
84	0.1869000E-03	0.5635800E-04	0.2458700E-03	0.3370500E-04
85	0.1514100E-03	0.1495200E-04	0.1394200E-03	0.1250600E-04
86	0.5912600E-04	0.1295400E-04	0.4838000E-04	0.1475800E-04
87	0.3423700E-04	0.1064600E-04	0.5029500E-04	0.1224500E-04
88	0.4477500E-04	0.1564300E-04	0.5608300E-04	0.1266800E-04
89	0.4787800E-04	0.1804200E-04	0.5366300E-04	0.1014000E-04
90	0.7247300E-04	0.1881600E-04	0.6756500E-04	0.1292900E-04
91	0.2296300E-04	0.9895300E-05	0.3559500E-04	0.1216000E-04
92	0.1032600E-03	0.1887900E-04	0.9917200E-04	0.1425800E-04

Tabla 3. (cont.) *Modelo probit. Medida de dispersión*

RÉPLICA	DD_1	DDM_1	DD_2	DDM_2
93	0.3953400E-04	0.1861400E-04	0.3478300E-04	0.1583000E-04
94	0.1365400E-03	0.5495300E-04	0.1186100E-03	0.3486100E-04
95	0.3152000E-04	0.9275400E-05	0.3727200E-04	0.1333100E-04
96	0.4714500E-04	0.1692500E-04	0.5345900E-04	0.1306600E-04
97	0.1101200E-03	0.1425700E-04	0.8603700E-04	0.1398300E-04
98	0.2135200E-04	0.8661000E-05	0.4578500E-04	0.1141000E-04
99	0.6002100E-04	0.1655200E-04	0.4724300E-04	0.1924500E-04
100	0.7932300E-04	0.1534300E-04	0.8141000E-04	0.1208500E-04

Tabla 4. *Modelo logit. Medida de dispersión*

RÉPLICA	DD_1	DDM_1	DD_2	DDM_2
1	0.3317700E-04	0.2070100E-04	0.3496000E-04	0.2150500E-04
2	0.8078600E-04	0.2143000E-04	0.6342400E-04	0.1654700E-04
3	0.5892300E-04	0.1933900E-04	0.5916000E-04	0.1555200E-04
4	0.3763900E-04	0.1281600E-04	0.4071800E-04	0.2259200E-04
5	0.3856500E-04	0.1240400E-04	0.5192800E-04	0.1447600E-04
6	0.7488700E-04	0.1292500E-04	0.7963400E-04	0.2006000E-04
7	0.3740600E-04	0.1374600E-04	0.4471500E-04	0.1886100E-04
8	0.4728300E-04	0.1480200E-04	0.4589500E-04	0.1099400E-04
9	0.3794700E-04	0.1522500E-04	0.3455500E-04	0.1777500E-04
10	0.4995000E-04	0.2064500E-04	0.4358800E-04	0.1855100E-04
11	0.4148700E-04	0.1792000E-04	0.4302000E-04	0.1542300E-04
12	0.3215500E-04	0.1094600E-04	0.6320900E-04	0.2095200E-04
13	0.3148100E-04	0.1315600E-04	0.2298500E-04	0.1264800E-04
14	0.3792400E-04	0.1377200E-04	0.4708900E-04	0.1320400E-04
15	0.2209500E-04	0.1381000E-04	0.4808500E-04	0.2569900E-04
16	0.8289800E-04	0.2734500E-04	0.8393600E-04	0.2888100E-04
17	0.3967600E-04	0.2008500E-04	0.4451200E-04	0.2269400E-04
18	0.2349800E-04	0.1539800E-04	0.4638400E-04	0.2405900E-04
19	0.2370300E-04	0.1424000E-04	0.4043900E-04	0.1526900E-04
20	0.5901500E-04	0.2118500E-04	0.5437300E-04	0.1284200E-04
21	0.5345100E-04	0.1216700E-04	0.5618400E-04	0.1370100E-04
22	0.3060500E-04	0.1411900E-04	0.4331800E-04	0.1912600E-04
23	0.3588100E-04	0.1150700E-04	0.2377200E-04	0.1204500E-04
24	0.2910900E-04	0.1357200E-04	0.5079100E-04	0.2209600E-04
25	0.5472000E-04	0.1483000E-04	0.5808100E-04	0.1031300E-04
26	0.6337600E-04	0.1765400E-04	0.7972900E-04	0.3306900E-04
27	0.5039100E-04	0.2112000E-04	0.2938200E-04	0.1584600E-04
28	0.3070400E-04	0.1451300E-04	0.3763700E-04	0.1233600E-04
29	0.2505200E-04	0.1332000E-04	0.2870600E-04	0.1287600E-04
30	0.3330200E-04	0.1244600E-04	0.2488700E-04	0.1723300E-04

Tabla 4. (cont.) *Modelo logit. Medida de dispersión*

RÉPLICA	DD_1	DDM_1	DD_2	DDM_2
31	0.2682100E-04	0.1445200E-04	0.2755100E-04	0.1693700E-04
32	0.3327300E-04	0.1870700E-04	0.4120200E-04	0.1732400E-04
33	0.1200500E-03	0.1598000E-04	0.1381200E-03	0.1409000E-04
34	0.2721500E-04	0.1186700E-04	0.3334800E-04	0.1574400E-04
35	0.2193700E-04	0.1280400E-04	0.2753100E-04	0.1217600E-04
36	0.3496800E-04	0.1591700E-04	0.6245300E-04	0.1479100E-04
37	0.5297400E-04	0.1942700E-04	0.5065600E-04	0.1782900E-04
38	0.4945400E-04	0.1126300E-04	0.4975500E-04	0.1813800E-04
39	0.3336200E-04	0.1021000E-04	0.5704000E-04	0.1232400E-04
40	0.2464000E-04	0.1268600E-04	0.2270600E-04	0.1417500E-04
41	0.3241400E-04	0.1477700E-04	0.2884300E-04	0.1633300E-04
42	0.5256100E-04	0.1670100E-04	0.6070300E-04	0.1620900E-04
43	0.2704900E-04	0.9646000E-05	0.3194100E-04	0.1493100E-04
44	0.2948800E-04	0.1149900E-04	0.5012900E-04	0.1396700E-04
45	0.2155500E-04	0.1493800E-04	0.4202600E-04	0.2375900E-04
46	0.4048600E-04	0.1276600E-04	0.3079200E-04	0.1683300E-04
47	0.4051700E-04	0.1290600E-04	0.4083400E-04	0.1518800E-04
48	0.3838300E-04	0.1789600E-04	0.3162700E-04	0.1577400E-04
49	0.3226700E-04	0.1826800E-04	0.3582000E-04	0.1918000E-04
50	0.5106800E-04	0.1608700E-04	0.4784800E-04	0.2027000E-04
51	0.3697300E-04	0.2103400E-04	0.2799800E-04	0.1457800E-04
52	0.4910500E-04	0.1299800E-04	0.4871700E-04	0.1748100E-04
53	0.4162400E-04	0.1270200E-04	0.4134400E-04	0.2016600E-04
54	0.2884300E-04	0.9620900E-05	0.3092600E-04	0.1313900E-04
55	0.8042600E-04	0.1485300E-04	0.6569100E-04	0.1702700E-04
56	0.3122800E-04	0.1331400E-04	0.6717200E-04	0.2711100E-04
57	0.7690300E-04	0.1625000E-04	0.8501300E-04	0.1781100E-04
58	0.2917400E-04	0.1411900E-04	0.4472800E-04	0.1489100E-04
59	0.5177800E-04	0.1325800E-04	0.6892200E-04	0.1283100E-04
60	0.6409000E-04	0.1914100E-04	0.6601600E-04	0.1329400E-04
61	0.2115400E-04	0.1102400E-04	0.2234300E-04	0.1074900E-04

Tabla 4. (cont.) *Modelo logit. Medida de dispersión*

RÉPLICA	DD_1	DDM_1	DD_2	DDM_2
62	0.3012200E-04	0.1712200E-04	0.3738400E-04	0.2253200E-04
63	0.3010400E-04	0.1474600E-04	0.3601700E-04	0.1466100E-04
64	0.4268700E-04	0.2026800E-04	0.4480300E-04	0.2282100E-04
65	0.8219200E-04	0.1024400E-04	0.9083400E-04	0.3465700E-04
66	0.1723700E-04	0.1049200E-04	0.2503100E-04	0.1876000E-04
67	0.3603700E-04	0.1772200E-04	0.3667400E-04	0.1466200E-04
68	0.4421800E-04	0.1490600E-04	0.2891000E-04	0.1618800E-04
69	0.4355600E-04	0.1518900E-04	0.5501200E-04	0.1507800E-04
70	0.3499600E-04	0.1623700E-04	0.3002300E-04	0.1545200E-04
71	0.3874400E-04	0.1746600E-04	0.5240200E-04	0.2061300E-04
72	0.3752800E-04	0.1429600E-04	0.4295200E-04	0.1385200E-04
73	0.8357200E-04	0.2558700E-04	0.8374500E-04	0.2913900E-04
74	0.5480200E-04	0.1732100E-04	0.7026700E-04	0.2210900E-04
75	0.2663500E-04	0.1612800E-04	0.3014600E-04	0.1637100E-04
76	0.4005300E-04	0.1480400E-04	0.6144900E-04	0.1644600E-04
77	0.5245000E-04	0.1888800E-04	0.3535500E-04	0.1605100E-04
78	0.7161300E-04	0.2566300E-04	0.5355700E-04	0.1711400E-04
79	0.3599200E-04	0.1378400E-04	0.4627600E-04	0.2283400E-04
80	0.2331400E-04	0.1314800E-04	0.4176900E-04	0.2351400E-04
81	0.3516900E-04	0.1488000E-04	0.4019900E-04	0.1483500E-04
82	0.5167600E-04	0.1540100E-04	0.3793500E-04	0.1700000E-04
83	0.2970700E-04	0.1497600E-04	0.3182500E-04	0.1686700E-04
84	0.3396600E-04	0.1031800E-04	0.4965400E-04	0.1658800E-04
85	0.5126700E-04	0.1706500E-04	0.8714900E-04	0.1265400E-04
86	0.7312300E-04	0.1619000E-04	0.1056100E-03	0.1268900E-04
87	0.2685200E-04	0.1430400E-04	0.3877500E-04	0.1893200E-04
88	0.4812800E-04	0.1438600E-04	0.5483700E-04	0.1454900E-04
89	0.3042100E-04	0.1772700E-04	0.2293900E-04	0.1574500E-04
90	0.3534500E-04	0.1361100E-04	0.2732500E-04	0.1479600E-04
91	0.2465000E-04	0.1428800E-04	0.3180100E-04	0.2361900E-04
92	0.8940500E-04	0.1815200E-04	0.1176100E-03	0.1803600E-04

Tabla 4. (cont.) *Modelo logit. Medida de dispersión*

RÉPLICA	DD_1	DDM_1	DD_2	DDM_2
93	0.5491700E-04	0.1997900E-04	0.6680000E-04	0.2605700E-04
94	0.4583000E-04	0.1722700E-04	0.5007900E-04	0.1974600E-04
95	0.4535900E-04	0.1546300E-04	0.5517400E-04	0.1319300E-04
96	0.6988900E-04	0.1438800E-04	0.1031800E-03	0.1550300E-04
97	0.2639500E-04	0.9943600E-05	0.4484800E-04	0.1762500E-04
98	0.2805400E-04	0.1601300E-04	0.2770300E-04	0.1759700E-04
99	0.3763400E-04	0.1187700E-04	0.4539300E-04	0.1835800E-04
100	0.2920700E-04	0.1775800E-04	0.3084500E-04	0.2103600E-04

REFERENCIAS

[1] **Aldrich, J.N.** and **F.D. Nelson** (1984). *Linear Probability, Logit and Probit Models*. (Sage, California).

[2] **Amemiya, T.** (1981). «Qualitative response models: A survey», *Journal of Economic Literature*, **19**, 1483–1536.

[3] **Aznar, A., A. García-Ferrer** y **A. Martín** (1994). *Ejercicios de Econometría*. Volumen II (Pirámide, Madrid).

[4] **Fisher, T.C.G.** (1991). «An empirical study of the adverse selection model of strikes», *Canadian Journal of Economics*, **XXIV**, 499–516.

[5] **Gourieroux, C., A. Monfort, E. Renault** and **A. Trognon** (1987). «Simulated residuals», *Journal of Econometrics*, **34**, 201–252.

[6] **Greene, W.H.** (1993). *Econometric Analysis*. Second Edition (MacMillan, New York).

[7] **Greene, L.** and **T. Seaks** (1992). «An analysis of the probability of default on federally guaranteed student loans», *Review of Economics and Statistics*, **74**, 404–411.

[8] **Griffiths, W.E., R. Carter-Hill** and **P.J. Pope** (1987). «Small sample properties of probit model estimators», *Journal of the American Statistical Association*, **82**, 929–937.

[9] **Hill, J.W.** y **R.W. Ingram** (1989). «Selection of GAAP of RAP in the Savings and Loan Industry», *The Accounting Review*, **LXIV, 4**, 667–679.

[10] **Jhonson, T.** (1987). *The analysis of qualitative and limited responses*, in: W.E. Becker and W.B. Walstad, eds., Econometric modeling in economic education research (Kluwer-Nijhoff, Boston).

[11] **LeClere, M.J.** (1992). «The interpretation of coefficients in models with qualitative dependent variables», *Decision Science Journal*, **23**, 770–776.

[12] **Maddala, G.S.** (1983). *Limited dependent and qualitative variables in Econometrics*. (Cambridge University, New York).

[13] **Maddala, G.S.** (1992). *Introduction to Econometrics*. Second Edition, (Mac Millan, New York).

[14] **McFadden, D.** (1976). «Quantal Choice analysis: A survey», *Annals of Economic and Social Measurement*, **5**, 363–390.

[15] **Pagan, A.** and **F. Vella** (1989). «Diagnostic tests for model based on individual data: A survey», *Journal of Applied Econometric*, **4**, 19–59.

[16] **Petersen, T.** (1985). «A comment of presenting results from logit and probit models», *American Sociological Review*, **50**, 130–131.

[17] **Train, K.** (1986). *Qualitative choice analysis.* (MIT Press, Massachusetts).

[18] **Windmeijer, F.A.G.** (1995). «Goodness-of-fit measures in binary choice models», *Economic Reviews*, **14**, 101–116.

ENGLISH SUMMARY

REFLECTIONS ON THE PROBABILITY CHANGE MEASUREMENT STRATEGY IN BINARY CHOICE MODELS

M.T. APARICIO*
I. VILLANÚA*
Universidad de Zaragoza*

This paper focuses on the evaluation of the measure which, in the context of Binary Choice Models, is required in order to represent the probability variation in the case of changes in the independent variables. The most common option to quantify this measure consists in evaluating it in the vector of mean values of the different explanatory variables. Our alternative strategy considers a more efficient use of sample information, based on emphasizing the average behaviour of the sample, rather than the behaviour of an «average individual». Finally, a simulation exercise is presented to illustrate that the proposed approach is more robust than the traditional option.

Keywords: Discrete choice, probability variation, robustness.

AMS Classification: 62J02, 65C05

* Los autores desean agradecer los comentarios de Antonio Aznar, Carmen García-Olaverri y de un evaluador anónimo, así como la financiación de DGICYT PB94-0602.

⋆ Departamento de Análisis Económico. Facultad de Ciencias Económicas y Empresariales. Universidad de Zaragoza. Gran Vía, 2. 50005 Zaragoza.

– Received July 1996.
– Accepted January 1998.

In recent years the so-called discrete choice models (DCM) have enjoyed considerable success due to the greater emphasis that has been placed on disaggregated models in econometric research. This type of model attempts to obtain the probability that a decision-making unit choose a given alternative from among a set of options, when the response regarding the choice made (y) and a set of variables (x) which represent characteristics of the available choices and of the decision-making unit are know.

The simplest models in this category are the dichotomous or binary DCMs, which can be represented generally as follows:

$$p_i = F(x_i'\beta) \tag{1}$$

and upon which this paper will focus.

In spite of the large number of empirical studies which use these models, their validation stage usually receives little attention. It is normal to present the estimated coefficients and their corresponding $t-$ratios, as well as one of the multiple scale measures of goodness of fit proposed for this kind of approach, the LR test relative to the joint significance of the model and the analysis of the effect the variations of the different socio-economic (explanatory) variables have on the probability of choice. The aim of this paper is to study only this last mentioned aspect.

In our context (non-linear models), this effect is determined, for continuous variables, by the expression:

$$(\partial p_i/\partial x_j)\,\Delta x_j \tag{2}$$

and, for discrete variables, by:

$$\frac{\Delta p_i}{\Delta x_j} = \frac{P(y_i = 1|x_j = 1) - P(y_i = 1|x_j = 0)}{1 - 0} \tag{3}$$

In either of these two situations, the specification of the probability variation requires the calculation of $x_i'\beta$. The parameter vector will have been estimated previously, but we need to decide on the value to be adopted for the set of independent variables. In this respect, we cannot speak of a unique strategy; rather, the literature contains various approaches. The most frequent option is to evaluate in the vector of the average mean values of the x's. In our opinion, this practice has some defects. First, in the case of dummy variables, the mean value is not very coherent with the meaning given to these variables. Secondly, the use of the average of the sample can lead into problems of lack of representativeness if the values of a given variable are very disperse.

Another option is to calculate $(\partial p_i/\partial x_j)$, in what are generically referred to as the "most significant"values of the explanatory variables. This practice is somewhat ambiguous, since the meaning of the term «most significant» is not explicitly defined.

In our view, the relevant question is the efficient treatment of the information and, consequently, it appears more appropiate to use the set of data relating to the decision-making unit, i.e. to obtain $(\partial \bar{p}_i/\partial x_j)$ for each element of the sample and summarize the result in a single value by calculating the average of these derivatives. In short, we propose to emphasizing on the average behaviour of the sample, rather than the behaviour of «an average decision-making unit», in order to avoid a loss of information which does not appear to be justified. When the x_i vector includes dummy variables, then, according to our strategy, the quantification of expression (3) can be summarized by calculating p_i for each element and obtaining the average for each specific sub-sample distinguished by the value associated with the dummy variable (x_j); the difference between these average probabilities gives the above mentioned change in probability associated with a change in the variable x_j.

Our next objective is to carry out an analysis of both measurements, with the purpose of determining, if possible, the superiority of one over the other according to some kind of criterion. In order to establish this superiority, the criterion of «robustness» would appear to be appropriate. In our particular approach, we will consider that one measure is more robust than another if, in the face of small variations in the sample information, the value of the former varies less than the value of the latter. In order to examine this question, we have designed a simulation exercise which, using a logit and a probit model, generates different samples; specifically, 100 models are generated with a sample size of $N = 100$.

Denoting the derivative at the mean point and the mean of the derivatives as D_k and DM_k respectively, we set out to analyze the influence that the elimination of an element has on such values. With this aim, our robustness measure can be expressed as:

$$(4) \qquad \frac{\sum_{i=1}^{100} (D_{ki} - VR_k)^2}{N} \quad \text{and} \quad \frac{\sum_{i=1}^{100} (DM_{ki} - VR'_k)^2}{N}$$

where k represents the explanatory variable for which the effect is calculated, VR and VR' denote, respectively, the value of D_k and DM_k for the whole sample, and D_{ki} and DM_{ki} are indicating the value of D_k and DM_k when the element i is eliminated.

The experiment concludes that, in all replications and in both models, D_k presents greater dispersion than DM_k. Furthermore, when calculating the mean of these dispersions, it can be observed in probit model that the dispersion of D_k is four times greater than that corresponding to DM_k. Therefore, the measure based on the average of the derivatives is more stable in the face of small changes in the sample information and, consequently, is more robust than the traditional method. This same general conclusion is obtained when we use an empirical application with real data.

QÜESTIIÓ, vol. 22, 2, p. 291-309, 1998

ESTIMACIÓN DEL NÚMERO DE CLUSTERS EN UNA POBLACIÓN APLICANDO EL JACKKNIFE GENERALIZADO

J.J. PRIETO MARTÍNEZ
Universidad Carlos III de Madrid*

Sea una población constituida por un número desconocido de clusters. Este trabajo desarrolla una secuencia finita de estimadores no paramétricos para el número de clusters, basándose en el método jackknife generalizado. Estos estimadores resultan ser una combinación lineal de las frecuencias de observación de cada cluster. Se propone entonces un procedimiento de selección para elegir el más apropiado. La técnica es aplicada a un conjunto de datos reales procedentes de un estudio de captura de especies de una población. Además, se lleva a cabo un estudio de simulación para investigar su comportamiento.

Estimation of the number of cluster in a population through the generalized jackknife.

Palabras clave: Número de clusters, jacknife generalizado.

Clasificación AMS: 162G05

* Universidad Carlos III de Madrid. Dpto. de Estadística y Econometría. C/ Madrid, 126. 28903 Madrid.
- Recibido en enero de 1997.
- Aceptado en diciembre de 1997.

1. INTRODUCCIÓN

Existe una gran cantidad de trabajos en la literatura estadística sobre los métodos de estimación del número de clusters en una población, pero la mayoría de ellos han sido desarrollados entorno a la idea de que las probabilidades de observación de los distintos clusters son iguales. Ver, por ejemplo, Lewontin y Prout (1956), Darroch (1958), Harris (1968), Johnson y Kotz (1977), Darroch y Ratcliff (1982) Marchand y Schrowck (1982), Holst (1981) y Esty (1985).

Existe un concepto que está muy ligado con el de número de clusters de una población, que es el cubrimiento muestral. Se define como la suma de las probabilidades de los clusters observados en una muestra. En el caso de clusters igualmente probables, el cubrimiento viene dado por el número de clusters observados en una muestra, dividido por el número de clusters que constituyen la población. Darroch y Ratclif (1980) utilizaron exactamente la idea del cubrimiento muestral para estimar K.

Ahora bien, considerar la hipótesis de que las probabilidades de los distintos clusters son iguales es, en principio, un caso muy particular y poco frecuente. Por ejemplo, no existe una misma cantidad de animales para cada especie en un ecosistema; no se repite con la misma frecuencia cada una de las diferentes palabras que constituyen un texto; no se acuña la misma cantidad de las distintas monedas utilizadas en un país durante un centenario, etc. La mayoría de los trabajos realizados para poblaciones heterogéneas (es decir, constituidas por clusters no equiprobables) adoptan un enfoque paramétrico. Por ejemplo, Fisher, Corbet y Williams (1943) asumen que para cada cluster, el número de observaciones en la muestra se distribuye según una distribución de Poisson, y el parámetro de dicha distribución se asume que sigue una distribución Gamma. Muchos otros artículos sobre modelos de abundancia de especies en un ecosistema también hacen consideraciones paramétricas. Ver, por ejemplo, Mc Neill (1973), Engen (1978), Efron y Thisted (1976). Fue Esty (1985), el primero en estimar el número de clusters en una población heterogénea mediante el concepto de cubrimiento muestral, aunque bajo un modelo paramétrico. Chao (1992) propone una técnica de estimación no paramétrica, pero utilizando también la idea del cubrimiento muestral.

La propuesta de este trabajo es justamente plantear una técnica de estimación no paramétrica alternativa a los estimadores mencionados anteriormente, sin necesidad de plantear un modelo de probabilidad ni de recurrir al concepto de cubrimiento muestral. La técnica empleada es el método jackknife generalizado.

Por tanto, considérese una población cerrada en la cual las observaciones están agrupadas en K clusters. El significado de cerrada hace alusión a que durante el estudio no se producen entradas o salidas de clusters existentes. Se propone inicialmente un estimador sesgado, el cual es corregido y ajustado mediante el jackknife generalizado.

Este trabajo está dividido, básicamente, en tres partes. La primera presenta, fundamentalmente, la idea general del método jackknife generalizado (apartado 2). En la segunda se aplica dicho método, obteniéndose una secuencia finita de estimadores para el número de clusters en una población. Todos ellos son una combinación lineal de las frecuencias con que los clusters son observados. Se calculan sus esperanzas matematicas, sus varianzas, y además se dan intervalos de confianza (apartado 3). Para elegir el estimador más apropiado se propone un procedimiento de selección basado en contraste de hipótesis (apartado 4). En la tercera y última parte se presenta un estudio numérico realizado por simulación, para comprobar la eficacia de las principales fórmulas presentadas aquí, así como una aplicación a un conjunto de datos reales procedentes de la captura y recaptura de especies en un determinado ecosistema (apartado 5).

2. EL MÉTODO JACKKNIFE GENERALIZADO

El jackknife (Quenouille (1949)) constituye una técnica básica de remuestreo o de reutilización muestral. Quenouille introdujo el jackknife como estimador del sesgo de un estimador, y Tukey (1958) sugirió su utilidad también en la estimación de la varianza. Miller (1974) hace una revisión sobre este método y Efron (1982) presenta de una forma concisa las ideas esenciales.

Una panorámica algo diferente del jackknife es justamente el jackknife generalizado, aunque con la misma idea central de estimación del sesgo de un estimador. Ver Gray y Shucany (1972).

A continuación se presentan los pasos generales que sigue dicho método.

Sea $y_1,\ldots,y_n$ una muestra aleatoria de una función de distribución $F(\theta)$. Sea $\hat{\theta}_{(n)} = \hat{\theta}(y_1,\ldots,y_n)$ un estimador de θ que verifica:

$$E\left(\hat{\theta}_{(n)}\right) = \theta + (a_1/n) + \left(a_2/n^2\right) + \cdots$$

donde $a_1, a_2, \ldots$ son constantes.

Sea $j_1,\ldots,j_i$ una combinación de i enteros del conjunto $\{1,\ldots,n\}$.

Se define $\hat{\theta}_{n-i,j_1,\ldots,j_i}$ como un estimador de θ basado en $(n-i)$ observaciones, después de haber eliminado al azar $y_{j_1},\ldots,j_{j_i}$ de la muestra. Entonces, el estimador

$$\hat{\theta}_{(n-i)} = \frac{\sum_{j_1<\cdots<j_i} \hat{\theta}_{n-i,j_1,\ldots,j_i}}{\binom{n}{i}},$$

llamado por Fraser (1957, p. 142) el estadístico $-U$, es la base del método jackknife generalizado para la reducción del sesgo del estimador $\hat{\theta}_{(n)}$. Esto se debe porque el estimador generalizado de orden h viene dado por:

$$\hat{\theta}_{J,h} = \frac{1}{h!}\sum_{i=0}^{h}(-1)^i \binom{h}{i}(n-i)^h \hat{\theta}_{(n-i)},$$

donde el subíndice J hace alusión al nombre de jackknife generalizado. Obsérvese que es una combinación lineal de los estimadores $\hat{\theta}_{(n-i)}$.

A continuación este método será aplicado eliminando grupos de observaciones (muestras) en vez de observaciones individuales.

3. UN ESTIMADOR PARA EL NÚMERO DE CLUSTERS EN UNA POBLACIÓN APLICANDO EL JACKKNIFE GENERALIZADO

Sea una población infinita de individuos, cerrada y formada por un número desconocido K de clusters. Sean t muestras aleatorias simples extraídas de la población con reemplazamiento, y p_j (con $j = 1,\ldots,K$) la probabilidad de observar el cluster j en cualquiera de las t muestras, $0 < p_j < 1$, $\sum_{j=1}^{K} p_j = 1$.

Sea f_r el número de clusters observados exactamente r veces de las t posibles muestras tal que,

$$S = \sum_{r=1}^{t} f_r$$

es el número de clusters observados al menos en una muestra.

Se define $\hat{K}_t = S$ un estimador natural de K obtenido a partir de la información de las t muestras, que es sesgado. Ahora el objetivo es corregir y ajustar $\hat{K}_t$ mediante su sesgo, el cual es estimado mediante el método jackknife generalizado.

Sea $\hat{K}_{t-1;r}$ un estimador de K al eliminar la muestra r–ésima (con $r = 1,\ldots,t$), el cual se define como:

$$\hat{K}_{t-1;r} = \hat{K}_t - \begin{pmatrix} \text{número de clusters observados} \\ \text{una vez en la muestra } r \end{pmatrix}, \quad \text{con } r = 1,\ldots,t;$$

y sea,

$$\hat{K}_{(t-1)} = \frac{1}{t}\sum_{r=1}^{t} \hat{K}_{(t-1);r}, \tag{1}$$

el estimador jackknife de primer orden, que es el promedio de los valores $\hat{K}_{t-1;r}$.

Desarrollando (1),

$$\hat{K}_{(t-1)} = \frac{1}{t}\sum_{r=1}^{t}\left\{\hat{K}_t - \left(\begin{array}{l}\text{número de clusters observados}\\ \text{una vez en la muestra } r.\end{array}\right)\right\} =$$

$$= \hat{K}_t - \frac{1}{t}\sum_{r=1}^{t}\left(\begin{array}{l}\text{número de clusters observados}\\ \text{una vez en la muestra } r.\end{array}\right) = \hat{K}_t - \frac{1}{t}f_1.$$

De la misma forma, si $\hat{K}_{t-2;r,l}$ es un estimador de K al eliminar las muestras r y l (con $r,l = 1,\ldots,t$) tal que,

$$\hat{K}_{t-2;r,l} = \hat{K}_t - \left\{\left(\begin{array}{l}\text{número de clusters observados}\\ \text{una vez en la muestra } r \text{ o en la } l.\end{array}\right) + \left(\begin{array}{l}\text{número de clusters observados dos}\\ \text{veces en las muestras } r \text{ y } l.\end{array}\right)\right\},$$

entonces el estimador jackknife de segundo orden es:

$$\hat{K}_{(t-2)} = \frac{1}{\binom{t}{2}}\sum_{\substack{r=1\\ l=1\\ r\neq 1}}^{t}\hat{K}_t - \left\{\left(\begin{array}{l}\text{número de clusters observados una}\\ \text{vez en la muestra } r \text{ o en la } l.\end{array}\right) -\right.$$

$$\left. - \left(\begin{array}{l}\text{número de clusters observados dos}\\ \text{veces en la muestra } r \text{ y } l.\end{array}\right)\right\} = \hat{K}_t - \frac{2}{t(t-1)}(t-1)f_1 - \frac{2}{t(t-1)}f_2 =$$

$$= \hat{K}_t - \frac{2}{t}f_1 - \frac{2}{t(t-1)}f_2.$$

$w_{m;i_1,\ldots,i_m}$ = el número de clusters observados exactamente m veces de las t posibles muestras, donde $i_1,\ldots,i_m$ es una combinación de los enteros $1,\ldots,t$.

Así, el número de clusters observados al menos una vez en las $(t-r)$ muestras es:

$$\hat{K}_{(t-r);i_1,\ldots,i_r} = \hat{K}_t - \sum_{m=1}^{r}\left\{\sum_{\{n_1,\ldots,n_m\}\subseteq\{i_1,\ldots,i_r\}} w_{m;n_1,\ldots,n_m}\right\}.$$

El sumatorio que está dentro del paréntesis es sobre las $\binom{r}{m}$ posibles combinaciones de enteros $\{n_1,\ldots,n_m\} \subseteq \{i_1,\ldots,i_r\}$. Se sigue que:

$$\hat{K}_{(t-r)} = \hat{K}_t - \frac{1}{\binom{t}{r}}\sum_{\{i_1,\ldots,i_r\}\subseteq\{1,\ldots,t\}}\sum_{m=1}^{r}\left\{\sum_{\{n_1,\ldots,n_m\}\subseteq\{i_1,\ldots,i_r\}} w_{m;n_1,\ldots,n_m}\right\} =$$

$$= \hat{K}_t - \frac{1}{\binom{t}{r}} \sum_{m=1}^{r} \binom{t-m}{r-m} \sum_{\{n_1,\ldots,n_m\} \subseteq \{1,\ldots,t\}} w_{m;n_1,\ldots,n_m} =$$

$$= \hat{K}_t - \frac{1}{\binom{t}{r}} \sum_{m=1}^{r} \binom{t-m}{r-m} \sum_{\{n_1,\ldots,n_m\} \subseteq \{1,\ldots,t\}} f_m.$$

A partir de los estimadores $\hat{K}_{(t-m)}$, el estimador jackknife generalizado para K de orden h es:

$$\hat{K}_{J,h} = \frac{1}{h!} \sum_{m=0}^{h} (-1)^m \binom{h}{m} (t-m)^h \hat{K}_{(t-m)}. \tag{2}$$

Nótese que (2) es una fórmula un poco engorrosa de manejar. A continuación se han dado valores a h desde 1 hasta 4, obteniéndose:

$$h = 1: \qquad \hat{K}_{J,1} = \hat{K}_t - \left(\frac{t-1}{t}\right) f_1,$$

$$h = 2: \qquad \hat{K}_{J,2} = \hat{K}_t + \left(\frac{2t-3}{t}\right) f_1 - \left(\frac{(t-2)^2}{t(t-1)}\right) f_2,$$

$$h = 3: \quad \hat{K}_{J,3} = \hat{K}_t + \left(\frac{3t-6}{t}\right) f_1 - \left(\frac{3t^2-15t+19}{t(t-1)}\right) f_2 + \left(\frac{(t-3)^3}{t(t-1)(t-2)}\right) f_3,$$

$$h = 4: \qquad \hat{K}_{J,4} = \hat{K}_t + \left(\frac{4t-10}{t}\right) f_1 - \left(\frac{6t^2-36t+55}{t(t-1)}\right) f_2 +$$

$$+ \left(\frac{4t^3-42t^2+148t-175}{t(t-1)(t-2)}\right) f_3 - \left(\frac{(t-4)^4}{t(t-1)(t-2)(t-3)}\right) f_4.$$

Obsérvese que $\hat{K}_{J,h}$ es una combinación lineal de las frecuencias con que los clusters son observados, de manera que dicho estimador se puede escribir de una forma general como:

$$\hat{K}_{J,h} = \sum_{r=1}^{t} a_{rh} f_r, \quad \text{con} \quad h \leqslant t,$$

donde a_{rh} son justamente los coeficientes que acompañan a las f_r los cuales están en función de t. Burnham y Overton (1978) obtuvieron este mismo resultado para la estimación del número de individuos en una población aplicando la técnica jackknife.

Justamente este artículo presenta otros resultados para los momentos del estimador $\hat{K}_{J,h}$.

Ahora, definiendo la variable aleatoria indicatriz:

$$Z_{j,r} = \begin{cases} 1 \text{ si el clusters } j \text{ es observado } r \text{ veces de las } t \text{ posibles muestras} \\ 0 \text{ en otro caso.} \end{cases}$$

$$E(\hat{K}_{J,h}) = \sum_{r=1}^{t} a_{rh} E(f_r) = \sum_{r=1}^{t} a_{rh} E\left(\sum_{j=1}^{K} Z_{j,r}\right) =$$

$$= \sum_{r=1}^{t} a_{rh} \sum_{j=1}^{K} P(Z_{j,r} = 1) = \sum_{r=1}^{t} a_{rh} \sum_{j=1}^{K} \binom{t}{r} p_j^r (1-p_j)^{t-r}.$$

Haciendo $\pi_r = E(f_r) = \sum_{j=1}^{K} \binom{t}{r} p_j^r (1-p_j)^{t-r}$, tal que

$$\text{Var}(f_r) = K\pi_r(1-\pi_r) \quad \text{y} \quad \text{Cov}(f_r r_l) = -K\pi_r\pi_l,$$

se tiene que:

(3)

$$\text{Var}(\hat{K}_{J,h}) = \sum_{r=1}^{t} a_{rh}^2 \text{Var}(f_r) + 2 \sum_{\substack{r=1 \\ r>1}}^{t} \sum^{t} a_{rh} a_{lh} \text{Cov}(f_r r_l) =$$

$$= K\sum_{r=1}^{t} a_{rh}^2 \pi_r(1-\pi_r) - 2K \sum_{\substack{r=1 \\ r>1}}^{t} \sum^{t} a_{rh}\, a_{lh}\, \pi_r\, \pi_l =$$

$$= K\sum_{r=1}^{t} a_{rh}^2 \pi_r - K\sum_{r=1}^{t} a_{rh}^2 \pi_r^2 - 2K \sum_{\substack{r=1 \\ r>1}}^{t} \sum^{t} a_{rh}\, a_{lh}\, \pi_r\, \pi_l =$$

$$= K\sum_{r=1}^{t} a_{rh}^2 \pi_r - K\sum_{r=1}^{t}\sum_{l=1}^{k} a_{rh}\, \pi_r\, a_{lh}\, \pi_l = K\sum_{r=1}^{t} a_{rh}^2 \pi_r - K\sum_{r=1}^{t} a_{rh}\, \pi_r \sum_{l=1}^{t} a_{lh}\, \pi_l =$$

$$= K\sum_{r=1}^{t} a_{rh}^2 \pi_r - \frac{K\sum_{r=1}^{t} a_{rh}\, \pi_r\, K\sum_{l=1}^{t} a_{lh}\, \pi_l}{K} = K\sum_{r=1}^{t} a_{rh}^2 \pi_r - \frac{E(\hat{K}_{J,h}\, E(\hat{K}_{J,h})}{K} =$$

$$= K\sum_{r=1}^{t} a_{rh}^2 \pi_r - \frac{E^2(\hat{K}_{J,h})}{K}.$$

Un estimador insesgado de mínima varianza para $\text{Var}(\hat{K}_{J,h})$ es:

$$\widehat{\text{Var}}(\hat{K}_{J,h}) = \frac{\hat{K}_t}{\hat{K}_t - 1}\left\{\hat{K}_t \sum_{r=1}^{t} a_{rh}^2 \hat{f}_r - \frac{(\hat{K}_{J,h})^2}{\hat{K}_t}\right\}, \tag{4}$$

donde $\hat{f}_r$ son los valores observados de π_r.

Por otra parte, $(f_1, f_2, \ldots, f_t)$ es una variable aleatoria multinomial tal, que cualquier combinación lineal de éstas es bien sabido que se distribuye asintóticamente como una normal. Así $\hat{K}_{J,h}$ se distribuye aproximadamente como:

$$N\left(E(\hat{K}_{J,h}),\ \widehat{\text{Var}}(\hat{K}_{J,h})\right).$$

Sea $z_{\alpha/2}$ el percentil $1-(\alpha/2)$ de la distribución $N(0,1)$. Un intervalo de confianza para K con un nivel de confianza $1-\alpha$ es:

$$\left(\hat{K}_{j,h} - z_{\alpha/2}\left(\widehat{\text{Var}}(\hat{K}_{J,h})\right)^{1/2},\ \hat{K}_{J,h} + z_{\alpha/2}\left(\widehat{\text{Var}}(\hat{K}_{J,h})\right)^{1/2}\right).$$

4. UN PROCEDIMIENTO DE ESTIMACIÓN

Una vez calculados los estimadores $\hat{K}_{J,h}$ para distintos valores de h, cabe preguntarse cuál es el que mejor estima o menor error comete. Para ello se propone el siguiente procedimiento basado en contrastes de hipótesis.

Contrastar la hipótesis nula:

$$H_{0,h} : E(\hat{K}_{J,h+1} - \hat{K}_{J,h}) = 0$$

frente a la hipótesis alternativa:

$$H_{1,h} : E(K_{J,h+1} - K_{J,h}) \neq 0,$$

y elegir como estimador de K, $\hat{K} = \hat{K}_{J,h}$, tal que $H_{0,h}$ es la primera hipótesis nula no rechazada. El mecanismo es el siguiente. Si $h = 1$ y H_{01} no es rechazada es que hay evidencia de que se logre un descenso en el sesgo por utilizar $\hat{K}_{J,2}$, mucho más que utilizando $\hat{K}_{J,1}$ y, por consiguiente, se concluye que no hay razón para utilizar $\hat{K}_{J,2}$. De hecho debe utilizarse $\hat{K}_{J,1}$ como estimador de K debido a que posteriores estimadores con $h > 2$ tendrán un sesgo mucho mayor. Si H_{01} es rechazada, entonces cabe esperar una reducción del sesgo utilizando $\hat{K}_{J,2}$. Ante la aceptación de $\hat{K}_{J,2}$ como estimador de k se contrastaría la elección de $\hat{K}_{J,2}$ frente a la de $\hat{K}_{J,3}$. En caso de rechazo de H_{02}

se volvería a realizar el contraste correspondiente. El proceso continua de manera progresiva hasta conseguir el estimador secuencial, $\hat{K}_{J,h}$, cuya H_{0h} sea la primera no rechazada.

Nótese que bajo la condición de que $H_{0,h}$ sea verdadera, el estadístico del contraste es:

$$T_h = \frac{\hat{K}_{J,h+1} - \hat{K}_{J,h}}{\left(\widehat{\mathrm{Var}}(\hat{K}_{J,h+1} - \hat{K}_{J,h}/\hat{K}_t)\right)^{1/2}},$$

el cual se distribuye aproximadamente como una $N(0,1)$. Para un nivel de significación α, la región crítica es:

$$C = \{z : |z| > z_{\alpha/2}\},$$

siendo $z_{\alpha/2}$ el valor de una normal (0,1) que deja a la derecha un área de probabilidad igual a $\alpha/2$. Por consiguiente se rechaza $H_{0,h}$ si $T_\alpha > z_{\alpha/2}$.

Es necesario indicar que como $\hat{K}_{J,h} = \sum_{r=1}^{t} a_{rh} f_r$, entonces

$$\hat{K}_{J,h+1} - \hat{K}_{J,h} = \sum_{r=1}^{t} a_{rh} f_r,$$

es decir, es también una combinación lineal de las frecuencias f_r. Por tanto, se sigue de la expresión (4):

$$\widehat{\mathrm{Var}}(\hat{K}_{J,h+1} - \hat{K}_{J,h}) = \frac{\hat{K}_t}{\hat{K}_t - 1}\left\{\hat{K}_t \sum_{r=1}^{t} a_{rh}^2 \hat{f}_r - \frac{(\hat{K}_{J,h+1} - \hat{K}_{J,h})^2}{\hat{K}_t}\right\}. \tag{5}$$

5. ESTUDIOS NUMÉRICOS

Ejemplo 1

La evaluación de los estimadores $\hat{K}_{J,h}$ ha sido llevada a cabo simulando t muestras (en particular 5, 10 y 15) de tamaño 100 de una población de K clusters (se han considerado desde una población con pocos clusters, $K = 5$, hasta una población constituida por numerosos clusters, $K = 150$). Cada caso se ha simulado 100 veces, de manera que los datos presentados son promedio de los resultados. También se ha calculado la varianza y desviación típica de los $\hat{K}_{J,h}$ para los casos de mayor interés, $K = 150$ (con $t = 5$ y $t = 15$). Nótese que son justamente los casos donde la estima

de K puede tener más sesgo debido a la heterogeneidad de la población. Obsérvese que de la expresión (5) se obtiene:

$$\widehat{\mathrm{Var}}(\hat{K}_{J,1}) = \hat{K}_t \left(1 + \frac{t-1}{t}\right)^2 f_1 + (\hat{K}_t - f_1) - \hat{K}_{J,1},$$

$$\widehat{\mathrm{Var}}(\hat{K}_{J,2}) = \hat{K}_t \left(1 + \frac{2t-3}{t}\right)^2 f_1 + \hat{K}_t \left(1 - \frac{(t-2)^2}{t(t-1)}\right)^2 f_2 + \left(\hat{K}_t - (f_1 + f_2)\right) - \hat{K}_{J,2},$$

$$\widehat{\mathrm{Var}}(\hat{K}_{J,3}) = \hat{K}_t \left(1 + \frac{3t-6}{t}\right)^2 f_1 + \hat{K}_t \left(1 - \frac{3t^2 - 15t + 19}{t(t-1)}\right)^2 f_2 +$$

$$+ \hat{K}_t \left(1 + \frac{(t-3)^2}{t(t-1)(t-2)}\right)^2 f_3 + \left(\hat{K}_t - (f_1 + f_2 + f_3)\right) - \hat{K}_{J,3}$$

y

$$\widehat{\mathrm{Var}}(\hat{K}_{J,4}) = \hat{K}_t \left(1 + \frac{4t-10}{t}\right)^2 f_1 + \hat{K}_t \left(1 - \frac{6t^2 - 36t + 55}{t(t-1)}\right)^2 f_2 +$$

$$+ \hat{K}_t \left(1 + \frac{4t^3 - 42t + 148t - 175}{t(t-1)(t-2)}\right)^2 f_3 + \hat{K}_t \left(1 - \frac{(t-4)^4}{t(t-1)(t-2)(t-3)}\right)^2 f_4 +$$

$$+ \left(\hat{K}_t - (f_1 + f_2 + f_3 + f_4)\right) - \hat{K}_{J,4}.$$

De los resultados obtenidos y reflejados en las tablas 1 y 2 se deduce que:

- $\hat{K}_j$ es sesgado, infraestimando siempre K, es decir, estima por debajo del valor de K.
- Para poblaciones constituidas por pocos clusters ($K < 50$) las estimas de $\hat{K}_{J,h}$ (con $h = 1,2,3,4$) son excelentes. Obsérvese que justamente para $K = 50$, los valores de $\hat{K}_{J,h}$ (con $j = 1,2,3$) empiezan a ser estimas sesgadas.
- En poblaciones constituidas por un gran número de clusters ($K = 100$ o $K = 150$), los estimadores $\hat{K}_{J,h}$ con $j = 1,\ldots,4$, trabajan muy bien, no produciéndose un sesgo muy grande. Nótese que a medida que t crece, la diferencia entre K y $\hat{K}_{J,h}$ es cada vez menor.
- En cualquiera de los casos, el mayor sesgo se produce para el estimador $\hat{K}_{J,1}$. Obsérvese siempre la mejora de $\hat{K}_{J,4}$ con respecto a $\hat{K}_{J,1}$.
- A medida que el número de muestras extraídas es mayor, la estimación en general de K con esta técnica es mejor.

- Si K es pequeño, el número de muestras en que cada cluster es observado, de las t posibles, es muy elevado. Si K es grande, raramente existen clusters que son observados en la mayoría de las t muestras.
- En la tabla 2 se observa que la desviación típica es menor a medida que el número de muestras crece. Además, en cualquiera de los tres casos de interés, cuando h crece, la desviación típica decrece.

Tabla 1. *Cálculo de los estimadores* $\hat{K}_{J,h}$

Caso	k	t	f_r	$\hat{K}_{J,h}$		
1	150	5	44, 42, 23, 5, 0	$\hat{K}_J = 114$	$\hat{K}_{J,1} = 159.20$	$\hat{K}_{J,2} = 156.36$
				$\hat{K}_{J,3} = 153.96$	$\hat{K}_{J,4} = 152.86$	
2		10	20, 26, 36, 33, 9	$\hat{K}_J = 134$	$\hat{K}_{J,1} = 152.0$	$\hat{K}_{J,2} = 149.51$
			9, 1, 0, 0, 0	$\hat{K}_{J,3} = 152.22$	$\hat{K}_{J,4} = 151.11$	
3		15	4, 6, 35, 27, 34	$\hat{K}_J = 150$	$\hat{K}_{J,1} = 153.37$	$\hat{K}_{J,2} = 152.37$
			22, 7, 8, 4, 2	$\hat{K}_{J,3} = 152.05$	$\hat{K}_{J,4} = 151.75$	
			1, 0, 0, 0, 0			
4	100	5	34, 45, 9, 1, 0, 0	$\hat{K}_J = 89$	$\hat{K}_{J,1} = 116.20$	$\hat{K}_{J,2} = 116.35$
				$\hat{K}_{J,3} = 97.98$	$\hat{K}_{J,4} = 98.97$	
5		10	9, 6, 32, 22, 12	$\hat{K}_J = 88$	$\hat{K}_{J,1} = 106.10$	$\hat{K}_{J,2} = 101.92$
			5, 2, 0, 0, 0	$\hat{K}_{J,3} = 102.65$	$\hat{K}_{J,4} = 98.83$	
6		15	1, 5, 8, 20, 26	$\hat{K}_J = 90$	$\hat{K}_{J,1} = 90.93$	$\hat{K}_{J,2} = 87.77$
			12, 5, 9, 2, 0	$\hat{K}_{J,3} = 98.82$	$\hat{K}_{J,4} = 100.96$	
			2, 0, 0, 0, 0			
7	50	5	12, 6, 5, 2, 0	$\hat{K}_J = 25$	$\hat{K}_{J,1} = 43.62$	$\hat{K}_{J,2} = 48.60$
				$\hat{K}_{J,3} = 49.95$	$\hat{K}_{J,4} = 49.93$	
8		10	2, 4, 11, 5, 7	$\hat{K}_J = 33$	$\hat{K}_{J,1} = 46$	$\hat{K}_{J,2} = 49.87$
			2, 2, 0, 0, 0	$\hat{K}_{J,3} = 50.73$	$\hat{K}_{J,4} = 50.87$	
9		15	0, 1, 4, 8, 8	$\hat{K}_J = 42$	$\hat{K}_{J,1} = 48$	$\hat{K}_{J,2} = 49.89$
			7, 9, 3, 0, 2	$\hat{K}_{J,3} = 50.29$	$\hat{K}_{J,4} = 50.42$	
			0, 0, 0, 0, 0			
10	20	5	1, 1, 4, 9, 2	$\hat{K}_J = 17$	$\hat{K}_{J,1} = 21.80$	$\hat{K}_{J,2} = 19.95$
				$\hat{K}_{J,3} = 20.98$	$\hat{K}_{J,4} = 20.35$	

Continúa

Continúa

Caso	k	t	f_r	$\hat{K}_{J,h}$		
11	20	10	0, 1, 0, 1, 1 3, 2, 6, 4, 0	$\hat{K}_J = 18.00$ $\hat{K}_{J,3} = 20.86$	$\hat{K}_{J,1} = 19.35$ $\hat{K}_{J,4} = 20.36$	$\hat{K}_{J,2} = 21.08$
12		15	0, 0, 1, 0, 0 0, 0, 2, 3, 2 3, 5, 3, 0, 0	$\hat{K}_J = 19.00$ $\hat{K}_{J,3} = 20.63$	$\hat{K}_{J,1} = 20.63$ $\hat{K}_{J,4} = 20.23$	$\hat{K}_{J,2} = 19.45$
13	10	5	0, 0, 1, 2, 7	$\hat{K}_J = 10.00$ $\hat{K}_{J,3} = 10.65$	$\hat{K}_{J,1} = 10.75$ $\hat{K}_{J,4} = 10.45$	$\hat{K}_{J,2} = 10.29$
14		10	0, 0, 0, 0, 0 0, 1, 1, 3, 3	$\hat{K}_J = 8.00$ $\hat{K}_{J,3} = 10.24$	$\hat{K}_{J,1} = 9.27$ $\hat{K}_{J,4} = 10.19$	$\hat{K}_{J,2} = 9.61$
15		15	0, 0, 0, 0, 0 0, 0, 0, 0, 0 1, 1, 3, 3, 1	$\hat{K}_J = 9.00$ $\hat{K}_{J,3} = 9.94$	$\hat{K}_{J,1} = 9.73$ $\hat{K}_{J,4} = 10.03$	$\hat{K}_{J,2} = 10.47$
16	5	5	0, 0, 0, 0, 5	$\hat{K}_J = 4.86$ $\hat{K}_{J,3} = 5.00$	$\hat{K}_{J,1} = 4.90$ $\hat{K}_{J,4} = 5.00$	$\hat{K}_{J,2} = 5.00$
17		10	0, 0, 0, 0, 0 0, 0, 0, 0, 5	$\hat{K}_J = 4.97$ $\hat{K}_{J,3} = 5.00$	$\hat{K}_{J,1} = 5.09$ $\hat{K}_{J,4} = 5.00$	$\hat{K}_{J,2} = 5.00$
18		15	0, 0, 0, 0, 0 0, 0, 0, 0, 0 0, 0, 0, 0, 5	$\hat{K}_J = 4.94$ $\hat{K}_{J,3} = 5.00$	$\hat{K}_{J,1} = 5.12$ $\hat{K}_{J,4} = 5.00$	$\hat{K}_{J,2} = 5.00$

Tabla 2. *Cálculo de la varianza de los $\hat{K}_{J,h}$ para los casos 1 y 2*

t	h	$\widehat{\text{Var}}(\hat{K}_{J,h})$	$\left(\widehat{\text{Var}}(\hat{K}_{J,h})\right)^{1/2}$
5	1	80.82	8.99
	2	52.78	7.26
	3	9.92	3.15
	4	8.76	2.96
10	1	8.00	2.83
	2	2.40	1.55
	3	0.16	0.41
	4	0.11	0.34
15	1	16.81	4.10
	2	11.22	3.35
	3	1.44	1.20
	4	1.06	1.03

Nótese que los valores de las f_r y K_J son valores promedio de las 100 simulaciones realizadas.

A continuación se procede a la elección del estimador secuencial para los casos 2 y 3. Para ello se aplicará el siguiente algoritmo de contraste de hipótesis explicado anteriormente.

Algoritmo

Paso 1: Hacer $h = 1$. Contrastar la hipótesis nula

$$H_{0,h} : E(\hat{K}_{J,h+1} - \hat{K}_{J,h}) = 0$$

frente a la hipótesis alternativa

$$H_{1,h} : E(\hat{K}_{J,h+1} - \hat{K}_{J,h}) \neq 0.$$

Paso 2: Rechazar $H_{0,h}$ si $|T_\alpha| > z_{\alpha/2}$, con $z_{\alpha/2}$ el valor de una $N(0,1)$ que deja a ambos lados un área de probabilidad $\alpha/2$.
Ir al paso 3.
En caso contrario, $h = h + 1$ e ir al paso 1.

Paso 3: Un estimador para K es $\hat{K}_{J,h}$.

Entonces, para el caso 3 ($K = 150, t = 15$), en el primer contraste se tiene que:

$$T_1 = \frac{152.37 - 153.73}{\left(\frac{150}{149}\left(6.89 - \frac{1.84}{150}\right)\right)^{1/2}} = 0.51.$$

Como $|T_1| < z_{\alpha/2}$ (con $\alpha = 0.05$, $z_{0.025} = 1.96$), se acepta la hipótesis nula y, por consiguiente, hay que realizar el siguiente contraste.

El estadístico del segundo contraste es $T_2 = 3.22$. Como $|T_2| > z_{\alpha/2}$ (con $\alpha = 0.05$), se rechaza la hipótesis nula, siendo el estimador apropiado para K, $\hat{K} = \hat{K}_{J,2} = 152.37$.

Los valores de T_h (para $h = 1,2,3,4$) del caso $K = 150$ ($t = 10$) están dispuestos en la tabla 3. Dependiendo del nivel de significación que se fije se obtendrá un estimador apropiado para K. Por ejemplo, para $\alpha = 0.05$, $z_{0.025} = 1.96$, $|T_2| = 0.11 < z_{0.025}$, por tanto se acepta la hipótesis nula. En cambio, $|T_3| = 3.51 > z_{0.025}$ y, por consiguiente, el estimador apropiado es $\hat{K}_{J,3} = 150.32$.

Tabla 3. *Cálculo de los T_h para el caso 2 de la tabla 1*

h	$\hat{K}_{J,h}$	T_h
1	152.00	−0.48
2	149.00	0.11
3	150.32	3.51
4	155.34	4.55

Un intervalo de confianza para los casos 2 y 3 son respectivamente:

$$\textit{Caso 2:}\ (150.32 \pm 1.96 \times 0.41) = (149.51,\ 151.12)$$

$$\textit{Caso 3:}\ (152.37 \pm 1.96 \times 3.35) = (145.81,\ 158.93)$$

Ejemplo 2

Como aplicación del procedimiento desarrollado en el apartado 3, se ha considerado un conjunto de datos reales pertenecientes a 10 muestras extraídas de un pantano junto al río Pettaquamscutt al sur de la Isla de Rhode. Son datos recogidos en abril de 1978 por Jeffrey Hyland del Colegio de Oceanografía de la Universidad de la Isla de Rhode. En Heltshe y Forrester (1983) se utilizan estos datos para realizar un estudio de muestreo, indicando que el número de clases de especies existentes en el pantano es 21. En las 10 muestras se observaron 14 especies cuyos datos fueron:

Tabla 4

	Muestras									
Especies observadas	1	2	3	4	5	6	7	8	9	10
S. benedicti.		13	21	14	5	22	13	4	4	27
N. succines.	2	2	4	1	1	1	1		1	6
P. lignis.		1						1		
S. Robustus.	1		1	2		6	1		1	2
E. heteropoda.			1	2						1
H. filiformis.	1	1	2	1		1			1	5
C. capitata.	1									
S. viridis.	2									
H. grayi.		1								
B. clavata.			1							
M. balthica.			3							2
A. abdita.			5	1		2				3
N. texana.								1		
Tubifocodies sp.	8	36	14	19	3	22	6	8	5	41

La tabla 3 indica por filas el número de observaciones realizadas por cada especie en cada una de las muestras. Se deduce fácilmente que:

$$f_1 = 5; \quad f_2 = 2; \quad f_3 = 0; \quad f_4 = 2; \quad f_5 = 0,$$
$$f_6 = 1; \quad f_7 = 1; \quad f_8 = 0; \quad f_9 = 2; \quad f_{10} = 1.$$

Entonces los valores de los estimadores son:

$$\hat{K}_{J,1} = 19.9; \quad \hat{K}_{J,2} = 19.9; \quad \hat{K}_{J,3} = 20.01; \quad \hat{K}_{J,4} = 20.01$$

los cuales se aproximan al verdadero valor de K.

6. AGRADECIMIENTOS

Quiero dar las gracias al profesor Kenneth P. Burnham, profesor de Estadística del Departamento de Biología de la Universidad de Colorado, y al profesor Kenneth H. Pollock del Departamento de Estadística de la Universidad de Carolina, por la colaboración prestada para la realización de este artículo.

7. BIBLIOGRAFÍA

[1] **Burnham, K.P.** y **Overton, W.S.** (1978). «Estimation of the size of a closed population when capture probabilities vary among animals». *Biometrika*, **63, 3**, 625–633.

[2] **Chao, A.** (1992). «Estimating the number of classes via sample coverage». *Journal of the American Association*, **87, 417**, 210–217.

[3] **Darroch, J.N.** (1958). «The multiple recapture census I: Estimation of a closed population». *Biometrika*, **40**, 343–359.

[4] **Darroch, J.N.** y **Ratclif, D.** (1980). «A note on capture-recapture estimation». *Biometrika*, **45**, 343–359.

[5] **Efron, B.** (1979). «Bootstrap methods, Another look at the jackknife». *The Annals of Statistic*, **7**, 1–26.

[6] **Efron, B.** y **Thisted, R.** (1976). «Estimating the number of unseen species: How many words did Shakespeare Know?». *Biometrics*, **63**, 435–447.

[7] **Efron, B.** (1982). «The jackknife, the bootstrap and other resampling plans». *Regional conference series and applied mathematics*. CBMS-NSF.

[8] **Engen, S.** (1978). *Stochastic Abundance Models*. London: Chapman-Hall.

[9] **Esty, W.W.** (1985). «The estimation of the number os classes in a population and the coverage of a sample». *Mathematical Scientist*, **10**, 41–50.

[10] **Fraser, D.A.S.** (1957). *Nonparametric methods in Statistic*. New York: Wiley.

[11] **Fiher, R.A., Corbet, A.S.** y **Williams, C.B.** (1943). «The relation between the number of species and the number of individuals in a random sample of an animal population». *Journal of Animal Ecology*, **12**, 42–58.

[12] **Gray, H.L.** y **Shucany, W.R.** (1972). *The Generalized Jackknife Statistic*, New York: Marcel Dekker.

[13] **Harris, B.** (1968). «Statistical inference in the classical occupancy problem unbiased estimation of the number of classes». *Journal of the American Statistical Association*, **63**, 837–847.

[14] **Heltshe, J.F.** y **Forrester, N.E.** (1983). «Estimating Species Richness using the Jackknife Procedure». *Biometrics*, **39**, 1–11.

[15] **Holst, L.** (1981). «Some asymptotic result for incomplete multinomial or Poisson samples». *Scandinavian Journal of Statistic*, **8**, 243–246.

[16] **Johnson, N.L.** y **Kotz, S.** (1977). *Urn models and their applications: An approach to modern discrete probability theory*. New York: John Wiley.

[17] **Lewontin, R.C.** y **Prout, T.** (1956). «Estimation of the number of different classes in a population». *Biometrics*, **12**, 211–223.

[18] **Marchand, J.P.** y **Schroeck, P.E.** (1982). «On the estimation of the number of equally likely classes in a population». *Communications in Statistics, Part A-Theory and Methods*, **11**, 1139–1146.

[19] **McNeil, D.** (1973). «Estimating an Author's vocabulary». *Journal of the American Statistical Association*, **68, 341**, 92–97.

[20] **Miller, R.G.** (1974). «The Jackknife-a rewiew». *Biometrika*, **61**, 1–17.

[21] **Quenouille, M.H.** (1949). «Approximate tests of correlation in time series». *Journal of the Royal Statistical Society, Series B*, **11**, 68–84.

[22] **Tukey, J.** (1958). «Bias and confidence in not quite large samples abstract». *Annals of Mathematical Statistics*, **29**, 614.

ENGLISH SUMMARY

ESTIMATION OF THE NUMBER OF CLUSTERS IN A POPULATION THROUGH THE GENERALIZED JACKKNIFE

J.J. PRIETO MARTÍNEZ
Universidad Carlos III de Madrid*

Let a population be compared by an unknown number of clusters. This work develops a finite sequence of estimators for the number of clusters using the generalized jackknife method. It is found that there estimators are a linear combination of the observation frequencies. A procedure is then proposed for selecting just one of them which will be finaly used. The technique is applied to an real set of capture histories from a study of a population. Also the paper includes a simulation study toinvestigate its behavior.

Keywords: Number of clusters, generalized jackknife.

AMS Classification: 162G05

* Universidad Carlos III de Madrid. Dpto. de Estadística y Econometría. C/ Madrid, 126. 28903 Madrid.
- Received January 1997.
- Accepted December 1997.

Assume that t random samples of size n are drawn from a population with unknown number of clusters, K, and unequal cluster probabilities.

There is a large statistical literature on estimation methods of the number of clusters in a population, In most practical applications, the equally likely assumption is not valid. For practical examples, see Lewontin and Prout (1956), Darroch (1958), Harris (1968), Johnson and Kotz (1977), Darroch and Ratclif (1982), Marchand and Schrowch (1982), Holst (1981) and Esty (1985).

The sample coverage, C, of a random sample from a multinomial population is defined to be the sum of the probabilities of the observed clusters. For an equiprobable population, C reduces to D/K where D is the number of distinct clusters observed in the sample. Darroch and Ratclif (1980) exactly used the idea of sample coverage to estimate K.

Most authors adopted a parametric approach to handle heterogenenous populations (i.e., unequal clusters probabilities). For example, Fisher, Corbet and Williams (1943) assumed that for each cluster the number of elements observed in the sample follows a Poisson distribution and the Poisson parameter is assumed to have a gamma-type distribution. Many other papers on stochastic abundance models also make parametric assumption; see, for example, McNeill (1973), Efron and Thisted (1976) and Engen (1978). For heterogeneous populations, Esty (1985) was the first to apply the concept of sample coverage to estimate the number of clusters in a parametric setup. Chao (1992) proposes a nonparametric estimation method but using also the idea the coverage sample.

In this paper, a nonparametric estimation technique is developed based on the generalized jackknife. Note that this method does't assume a parametric model and it doesn't use the sample coverage.

Then, let t random samples are taken from a population of elements belonging to K different clusters. Let p_j be the probability that any observation belongs to the $j-$th cluster in some of the t samples, with $0 < p < 1$, $\sum p_j = 1$.

Let f_r the number of clusters observed exactly r times of the t possibles samples, and

$$S = \sum_{r=1}^{t} f_r$$

the number of clusters observed.

The $\hat{K}_t = S$ is a natural estimator to K. The goal is correcting and adjusting for the bias $\hat{K}_t$; the approach to estimating the bias is under the method generalized jackknife.

The estimator obtained is given by

$$\hat{K}_{J,h} = \sum_{r=1}^{t} a_{rh} f_r,$$

which is a linear combination of the frequencies whose clusters are observed.

The expected value is calculated:

$$E(\hat{K}_{J,h}) = \sum_{r=1}^{t} a_{rh} \sum_{j=1}^{K} \binom{t}{r} p_j^r (1 - p_j)^{t-r};$$

and an estimator to the variance is:

$$\widehat{\text{Var}}(\hat{K}_{J,h}) = \frac{\hat{K}_t}{\hat{K}_t - 1} \left\{ \hat{K}_t \sum_{r=1}^{t} a_{rh}^2 \hat{f}_r - \frac{(\hat{K}_{J,h})^2}{\hat{K}_t} \right\}.$$

Investigació Operativa

QÜESTIIÓ, vol. 22, 2, p. 313-335, 1998

UN MÉTODO PRIMAL DE OPTIMIZACIÓN SEMI-INFINITA PARA LA APROXIMACIÓN UNIFORME DE FUNCIONES*

T. LEÓN
S. SANMATÍAS
E. VERCHER
Universitat de València*

En este trabajo presentamos un algoritmo que resuelve problemas clásicos de aproximación que pueden ser formulados como programas semi-infinitos lineales. Hemos estudiado la caracterización algebraica de los puntos extremos y demostrado algunas de sus propiedades. Hemos diseñado un procedimiento que genera direcciones factibles a partir de la solución de ciertos programas lineales finitos, que también caracteriza la solución óptima del problema. El método incorpora una etapa interna de purificación para alcanzar un punto extremo desde cualquier solución factible, mejorando el valor de la función objetivo. Finalmente, comparamos el comportamiento de las diferentes estrategias sobre varios problemas de aproximación, comprobando la eficacia del algoritmo propuesto.

A primal semi-infinite programming method for the uniform approximation problem.

Palabras clave: Programación semi-infinita, aproximación uniforme, puntos extremos, métodos de direcciones factibles.

Clasificación AMS: 90C34, 41A50, 49M35

*Este trabajo ha sido subvencionado por el Ministerio de Educación y Ciencia, España, DGICYT, PB93-0703.

* Departament d'Estadística i Investigació Operativa. Facultat de Matemàtiques. Universitat de València. C/ Dr. Moliner, 50. 46100 Burjassot (Valencia).

- Recibido en junio de 1997.
- Aceptado en marzo de 1998.

1. INTRODUCCIÓN

El problema de aproximación uniforme consiste en determinar los coeficientes x_1, x_2, $\ldots$, x_n de aquella combinación lineal de un conjunto dado de funciones $\{a_1, a_2, \ldots, a_n\}$ continuas y definidas en un intervalo $[\alpha, \beta]$, que mejor aproxima a una cierta función $b(\cdot)$, también continua y definida en $[\alpha, \beta]$. Es decir, se trata de minimizar $x_0 := \max |b(s) - P(x,s)|$, donde la variable x_0 es el error máximo y $P(x,s) := x_1 a_1(s) + x_2 a_2(s) + \cdots + x_n a_n(s)$, $x_1, \ldots, x_n \in \mathbb{R}$ denota una aproximación lineal.

Es bien conocido que cuando el conjunto de funciones aproximantes es el de los polinomios de grado menor que n, el error máximo de la mejor aproximación se alcanza al menos en $n+1$ puntos de $[\alpha, \beta]$, con signos alternos, y que esta mejor aproximación es única. Esta propiedad de alternancia no solo se verifica en el caso de aproximación mediante polinomios, ocurre siempre que el conjunto de funciones $\{a_1, a_2, \ldots, a_n\}$ cumplan determinadas condiciones (vease Cheney (1966)). El procedimento utilizado tradicionalmente para resolver esta clase de problemas es el algoritmo de Remez, en sus diferentes versiones. En el libro de Powell (1981) se puede encontrar la descripción de estos algoritmos.

El problema de aproximación uniforme puede formularse como un problema semi-infinito lineal de la siguiente manera:

$$
\begin{array}{lll}
\text{(PA1)} \quad \text{Min} & x_0 & \\
\text{s.a.} & x_0 + a(s)^T x \geq b(s) & \\
& x_0 - a(s)^T x \geq -b(s) & \\
& x_0 \in \mathbb{R}, \quad x \in \mathbb{R}^n, & s \in [\alpha, \beta]
\end{array}
$$

Los problemas de aproximación de funciones constituyen, de hecho, una de las aplicaciones más importantes de la programación semi-infinita. En el libro de Anderson y Nash (1987) se puede encontrar una discusión detallada del problema de aproximación uniforme, que incluye la formulación del dual standard de (PA1):

$$
\begin{array}{lll}
\text{(DA)} \quad \text{Max} & \left\{ \displaystyle\sum_{s \in S} w_1(s)\, b(s) - \sum_{s \in S} w_2(s)\, b(s) \right\} & \\
\text{s.a.} & \displaystyle\sum_{s \in S} w_1(s) + \sum_{s \in S} w_2(s) = 1 & \\
& \displaystyle\sum_{s \in S} w_1(s)\, a(s) - \sum_{s \in S} w_2(s)\, a(s) = 0_n & \\
& w_1(s), w_2(s) \in \mathbb{R}_+^{(S)} & s \in S
\end{array}
$$

donde $\mathbb{R}_+^{(S)}$ es el cono de las sucesiones finitas generalizadas no negativas. Si denotamos el soporte de w_i como $\text{supp}(w_i) = \{s \in S : w_i(s) > 0\}$, para $i = 1, 2$, se tiene que $\text{supp}(w_i)$ es finito.

Anderson y Nash (1987) han demostrado que uno de los métodos de cambio de Remez puede considerarse como una extensión del algoritmo dual del simplex para programas semi-infinitos lineales aplicado a (DA), pues pasa de una solución factible básica del problema dual a otra, aumentando el valor de la función objetivo. Siendo las soluciones factibles básicas del dual (w_1, w_2) aquellas cuyo soporte, $\text{supp}(w_1) \cup \text{supp}(w_2)$, contiene exactamente $n+1$ elementos que estan intercalados, de forma que si $s_1 < s_2$ están en $\text{supp}(w_1)$, existe $t \in \text{supp}(w_2)$ tal que $s_1 < t < s_2$.

En el libro de Glashoff y Gustafson (1983) se hace un estudio exhaustivo de este par de problemas duales. Bajo la exigencia de que el conjunto de funciones aproximantes $\{a_1, a_2, \ldots, a_n\}$ es un sistema de Chebyshev de orden dos, se demuestra que no hay fallo de dualidad. La caracterización de la solución óptima del primal se establece a partir de las propiedades de las soluciones del problema dual y se propone un algoritmo de tres fases para resolver el problema (PA1).

El motivo por el que nos hemos planteado desarrollar un método primal para resolver el problema (PA1), es que la aproximación uniforme de funciones y sus aplicaciones dan lugar a una amplia clase de problemas de programación semi-infinita (PSI), que se utilizan a menudo como ejemplos test para ilustrar el comportamiento de procedimientos más generales de PSI (vease, p.e. Hettich y Kortanek (1993)). Los algoritmos que presentamos en este trabajo son especializaciones de un método de direcciones factibles pensado para resolver problemas con función objetivo lineal y con restricciones indexadas en subconjuntos compactos de $\mathbb{R}$.

En la Sección 2 estudiamos la caracterización algebraica de los puntos extremos de las regiones factibles de (PA1) y de su dual como soluciones factibles básicas primales y duales, demostramos que cuando la solución óptima del problema primal cumple la propiedad de la alternancia de signos, el óptimo se alcanza en un punto extremo degenerado. En la Sección 3 introducimos un esquema para determinar direcciones factibles de descenso, basado en una condición de optimalidad del tipo Karush-Kuhn-Tucker que da lugar a los tres algoritmos que detallamos en la Sección 4. Y, finalmente, en la Sección 5 mostramos y comparamos los resultados obtenidos al resolver algunos problemas utilizando los diferentes métodos implementados.

2. PUNTOS EXTREMOS

En esta sección se comprueba que la solución óptima del problema (PA1) se encuentra en un punto extremo de la región factible y caracterizaremos la forma que tienen estos puntos. Introduciendo dos funciones de holgura $z_1(y,s) = x_0 + a(s)^T x - b(s)$ y $z_2(y,s) = x_0 - a(s)^T x + b(s)$, para $y = (x_0, x)$ podemos escribir (PA1) en forma standard:

$$
\begin{array}{lll}
\text{(PA2)} \quad \text{Min} & y_0 & \\
\text{s.a.} & v_1(s)^T y - z_1(y,s) = b(s) & s \in S \\
& v_2(s)^T y - z_2(y,s) = -b(s) & s \in S \\
& y \in \mathbb{R}^{n+1}, \quad z(y,s) \geq 0 & s \in S
\end{array}
$$

siendo $S = [\alpha,\beta]$, $v_1(s) = (1,a_1(s),\ldots,a_n(s))^T$, $v_2(s) = (1,-a_1(s),\ldots,-a_n(s))^T$ y $z(\cdot,\cdot) = (z_1(\cdot,\cdot),z_2(\cdot,\cdot))$.

En un trabajo publicado en 1985 Nash define, en un contexto puramente algebraico, los objetos fundamentales y las operaciones de la programación lineal en espacios infinito dimensionales. De esta manera, puede extender las definiciones de soluciones factibles básicas degeneradas y no degeneradas, costes reducidos y pivote de la programación lineal al caso infinito dimensional. En particular, buscando una definición algebraica de solución factible básica, de manera que hubiera una correspondencia con la de punto extremo, Nash demostró que lo que debía caracterizar a una solución factible básica es que define un subespacio sobre el cual se resuelven con unicidad la restricciones, este subespacio es precisamente lo que llamó solidificación. Siguiendo su notación, para el (PA2), tendríamos que $X = \mathbb{R}^{n+1}\, xC(S)^2$, donde $C(S)^2 = \{(u_1,u_2) : u_1 \in C(S), u_2 \in C(S)\}$ y la aplicación lineal:

$$
\begin{array}{lccc}
A: & X & \longrightarrow & C(S)^2 \\
& \left((x_0,x)^T,(u_1(\cdot),u_2(\cdot)\right) & \longrightarrow & \left(x_0 + a(\cdot)^T x - u_1(\cdot), x_0 - a(\cdot)^T x - u_2(\cdot)\right)
\end{array}
$$

siendo $C(S)$ el espacio de las funciones continuas en S. Denotamos por

$$
\begin{array}{rcl}
N(A) & = & \left\{\left((x_0,x)^T,(u_1,u_2)\right) \in \mathbb{R}^{n+1}\, XC(S)^2 : u_1(s) = \right. \\
& = & \left. x_0 + a(s)^T x, u_2(s) = x_0 - a(s)^T x\right\}
\end{array}
$$

el anulador de A y por

$$
\begin{array}{rcl}
B\left((x_0,x)^T,(u_1,u_2)\right) & = & \left\{\left((\chi_0,\chi)^T,(\nu_1,\nu_2)\right) : \exists\, \lambda > 0 : ((x_0,x)^T,(u_1,u_2)) \right. \\
& & \left. \pm\lambda((\chi_0,\chi)^T,(\nu_1,\nu_2)) \geq 0\right\}
\end{array}
$$

la solidificación de $\left((x_0,x)^T,(u_1,u_2)\right)$. En ese mismo trabajo se demuestra que si existe una solución óptima $(y^*;z^*) \in X$, tal que la dimensión de $B(y^*;z^*) \cap N(A)$ es finita, entonces el óptimo también se alcanza en un punto extremo. En nuestro caso, se comprueba fácilmente que $N(A)$ tiene dimensión finita. Por tanto, la única solución óptima del problema (PA2) es un punto extremo.

Hemos comprobado que, al particularizar las caracterizaciones de punto extremo, solución factible básica degenerada y no degenerada para el problema (PA2), obtenemos

resultados análogos a los que Anderson y Lewis (1989) han establecido para una única función de holgura. Por este motivo no incluimos las pruebas de los Teoremas 2.1 y 2.2.

Definición 2.1. *Sea* $(y;z)$ *una solución factible del problema (PA2), se define* $constr_i(y) = \{s \in S : z_i(y,s) = 0\}$ *para* $i = 1,2$, *y* $constr(y) = constr_1(y) \cup constr_2(y)$.

Para toda solución factible $(y;z)$, los elementos s de $\text{constr}_i(y)$, $i = 1,2$, son aquellos en los que se alcanza la diferencia máxima entre $b(s)$ y $P(x,s)$. En los puntos de $\text{constr}_1(y)$ ($\text{constr}_2(y)$) esta diferencia es positiva (negativa). Por tanto, $\text{constr}_1(y) \cap \text{constr}_2(y) = \emptyset$, ya que si existiera $s^* \in S$, tal que $z_1(y,s^*) = z_2(y,s^*) = 0$, entonces $x_0 + a(s^*)^T x - b(s^*) = 0 = x_0 - a(s^*)^T x + b(s^*)$, con lo que $x_0 = 0$ y la aproximación óptima sería exacta.

Asumimos las siguientes hipótesis:

Hipótesis 1. *Las funciones* $b(\cdot)$ *y* $a_i(\cdot)$, $i = 1,2,\ldots,n$, *son n veces continuamente diferenciables.*

Hipótesis 2. *El conjunto* $constr(y)$ *es finito, para todo* $y \in \mathbb{R}^{n+1}$.

Nota. Si las funciones $a_i(\cdot)$ y $b(\cdot)$ son analíticas, entonces z_1 y z_2 lo son y tienen un número finito de ceros en S o son idénticamente nulas. En particular, esta condición se cumple cuando las funciones aproximantes son polinomios y $b(\cdot)$ es analítica.

Definición 2.2. *Sea* $(y;z)$ *una solución factible de (PA2), tal que* $constr(y) = \{s_1, s_2, \ldots, s_m\}$. *Denotamos por* $d(i)$ *el grado de* $s_i \in \text{constr}(y)$, i.e. $d(i)$ es el menor entero tal que:

$$z^{(k)}(y,s_i) = 0 \qquad k = 0,1,\ldots,d(i)$$
$$z^{(k)}(y,s_i) \neq 0 \qquad k = d(i)+1 \quad i = 1,2,\ldots,m$$
donde $z \equiv z_1$ o $z \equiv z_2$.

Teorema 2.1. *Sea* $(y;z)$ *una solución factible de (PA2), entonces* $(y;z)$ *es un punto extremo si y solo si* $V(y) = \{x \in \mathbb{R}^{n+1} : v_k^{(j)}(s_i)^T x = 0,\ j = 0,\ldots,d(i),\ s_i \in constr_k(y),\ k = 1,2\} = \{0_{n+1}\}$, *donde* $v_k^{(j)}(s_i) = \left(0, a_1^{(j)}(s_i), a_2^{(j)}(s_i), \ldots, a_n^{(j)}(s_i)\right)^T$, *para* $j > 0$.

Nota. La condición $V(y) = \{0_{n+1}\}$ es equivalente a la independencia lineal de las filas de la matriz $B(y)$, cuyas columnas son:

$$v_1(s_1),\ldots,v_1(s_p),v_2(s_{p+1}),\ldots,v_2(s_m),v_1'(s_1),v_1''(s_1),\ldots,v_1^{d(1)}(s_1),\ldots,v_1'(s_p),$$
$$v_1''(s_p),\ldots,v_1^{d(p)}(s_p),v_2'(s_{p+1}),v_2''(s_{p+1}),\ldots,v_2^{d(p+1)}(s_{p+1}),\ldots,v_2^{d(m)}(s_m)$$

Esta condición es, en realidad, la extensión de la caracterización de punto extremo en Programación Lineal. La diferencia consiste en que, en el caso infinito dimensional, hay dos clases de vectores involucrados en la definición de $V(y)$: los gradientes de las restricciones activas y lo que podríamos llamar «sus derivadas». Visto de esta forma, tendríamos que una restricción activa proporciona más información en un programa semi-infinito que en uno finito. Un comentario análogo puede hacerse acerca de las soluciones factibles básicas no degeneradas que vienen caracterizadas por el siguiente resultado:

Teorema 2.2. *Sea $(y;z)$ una solución factible básica de (PA2), $(y;z)$ es un punto extremo no degenerado si y sólo si la matriz $B(y)$ es invertible.*

Como ya se ha dicho, la solución óptima de (PA2) es un punto extremo, y según se comprueba a continuación, si además verifica la propiedad de que el error máximo de la mejor aproximación se alcanza al menos en $n+1$ puntos de $[\alpha,\beta]$, con signos alternos (que denotaremos por AS), entonces es degenerado. Para nuestro enfoque, esta característica del problema condiciona el tipo de test de optimalidad y los generadores de direcciones factibles de descenso (vease Anderson y Lewis (1989) y León y Vercher (1992a)).

Teorema 2.3. *Supongamos que se cumplen las Hipótesis 1 y 2, que $n \geq 2$ y que la solución óptima cumple la propiedad AS, entonces el óptimo se alcanza en un punto extremo degenerado.*

Demostración. Sea $P^*(x,s)$ la solución óptima del problema de aproximación uniforme. Por verificarse AS, sabemos que existen al menos $n+1$ puntos en S, $s_0 < s_1 < \cdots < s_n$, tales que $h = |P^*(x,s_i) - b(s_i)|$ para $i = 0,1,\ldots,n$, donde h es el error máximo. Si $P^*(x,s) = x_1 a_1(s) + x_2 a_2(s) + \cdots + x_n a_n(s)$, entonces $y^* = (h,x_1,\ldots,x_n)^T$ es una solución óptima del problema (PA2).

Por el Teorema 4 de Nash (1985) y dado que los coeficientes asociados a las funciones aproximantes son únicos, se tiene que $(y^*;z^*)$ es un punto extremo. La matriz $B(y^*)$ tiene exactamente $n+1$ filas y $|\text{constr}(y^*)| \geq n+1$.

a) Si $|\text{constr}(y^*)| \geq n+2$, entonces $B(y^*)$ tiene al menos $n+2$ columnas.

b) Si $|\text{constr}(y^*)| = n+1$ y $n \geq 2$, entonces existe $s_i \in \text{int}(S)$, tal que s_i es un mínimo local de $z_1^*(\cdot)$ o de $z_2^*(\cdot)$, luego $d(s_i) \geq 1$, y hay en $B(y^*)$ al menos dos columnas ligadas a s_i. Así, pues, $B(y^*)$ tiene al menos $n+2$ columnas.

En cualquier caso, la matriz $B(y^*)$ no es invertible, i.e. $(y^*;z^*)$ es un punto extremo degenerado.

■

No ocurre lo mismo cuando $n=1$ o cuando la solución óptima no verifica la propiedad AS ya que podemos encontrar ejemplos, como el 2.1 y el 2.2, en los que la solución óptima es un punto extremo no degenerado.

Ejemplo 2.1. Consideremos el problema de aproximar $b(s)=s$, mediante $a_1(s)=1$, para $s \in [0,1]$. El mejor polinomio constante que aproxima $b(\cdot)$ es $P^*(x,s)=0.5$ y la solución óptima del problema (PA2) es $y^*=(0.5,0.5)^T$ con $z_1(y^*,s)=1-s$, $z_2(y^*,s)=s$, $\text{constr}_1(y^*)=\{1\}$, $\text{constr}_2(y^*)=\{0\}$, $d(1)=0=d(0)$. Por supuesto, la matriz asociada $B(y^*)$ es invertible e y^* es una solución básica no degenerada.

Ejemplo 2.2. Consideremos la aproximación de la función $b(s)=s^2$ en $[0,2]$, mediante una combinación lineal de las funciones $\{s, \exp(s)\}$. Este es un problema resuelto por Andreasson y Watson (1976) y cuya solución coincide (salvo precisión) con la que hemos obtenido al aplicar nuestro algoritmo:

$$y^* = (0.53824531859082, 0.18423413878873, 0.41863079193405)^T.$$

Se tiene que

$$\text{constr}_1(y^*) = \{2\},\ \text{constr}_2(y^*) = 0.40637642, d(2) = 0, d(0.40637642) = 1$$

y puede comprobarse fácilmente que no se cumple la propiedad AS y que y^* es un punto extremo no degenerado.

Cuando la aproximación uniforme se lleva a cabo mediante el conjunto de polinomios $\{1, s, s^2, \ldots, s^{n-1}\}$ la estructura del problema nos proporciona algunas ventajas. El Lema 2.4 recoge una forma muy simple de descartar que una solución factible sea punto extremo.

Lema 2.4. *Sea $(y;z)$ una solución factible de (PA2) y sean $a_i(s)=s^{i-1}, i=1,\ldots,n$. Si al menos uno de los conjuntos $\text{constr}_1(y)$ o $\text{constr}_2(y)$ es vacío, entonces $(y;z)$ no es un punto extremo.*

Demostración. Si ambos conjuntos son vacios, la solución factible no tiene restricciones activas y no puede ser un punto extremo. Supongamos que $\text{constr}_1(y) = \{s_1, s_2, \ldots, s_k\}$ para $k \geq 1$, y $\text{constr}_2(y) = \emptyset$, por definición $v_1(s) = (1, 1, \ldots, s^{n-1})^T$, y las dos primeras filas de $B(y)$ coinciden. En el otro caso, si $\text{constr}_1(y) = \emptyset$, las dos primeras filas de $B(y)$ son linealmente dependientes.

■

También puede comprobarse fácilmente que la dirección de descenso $(-1, 1, 0_{n-1})$ es factible para soluciones factibles con $\text{constr}_2(y) = \emptyset$, y que el movimiento a lo largo de ella transforma los mínimos globales positivos de la función de holgura, $z_2(y, s)$, en índices activos para el nuevo punto, sin cambiar el otro conjunto de índices activos. La dirección $(-1, -1, 0_{n-1})$ hace el mismo papel para la función z_1.

Ejemplo 2.3. Encontrar la mejor aproximación de $s^2 - s^4$ en $[-1, 1]$, mediante una recta:

$$\begin{array}{lll} \text{(P)} \quad \text{Min} & x_0 & \\ \text{s.a.} & x_0 + x_1 + s x_2 \geq s^2 - s^4 & s \in S \\ & x_0 - x_1 - s x_2 \geq -s^2 + s^4 & s \in S \end{array}$$

$y^0 = (1/2, 1/2, 0)^T$ es una solución factible tal que $\text{constr}_1(y^0) = \emptyset$ y $\text{constr}_2(y^0) = \{-1, 0, 1\}$. La dirección $d = (-1, -1, 0)^T$ conduce a $y^* = (1/8, 1/8, 0)^T$, con $\text{constr}_1(y^*) = \{\pm\sqrt{2}/2\}$ y $\text{constr}_2(y^*) = \{-1, 0, 1\}$, que resulta ser la solución óptima del problema.

Con respecto a la solución del problema dual, se tiene que hay un número infinito de soluciones óptimas. Sólo cinco de ellas son básicas y vienen definidas por sus soportes:

(1) $\text{supp}(w_1^1) = \{-\sqrt{2}/2\}$ y $\text{supp}(w_2^1) = \{-1, 0\}$;

(2) $\text{supp}(w_1^2) = \{-\sqrt{2}/2\}$ y $\text{supp}(w_2^2) = \{\pm 1\}$;

(3) $\text{supp}(w_1^3) = \{\sqrt{2}/2\}$ y $\text{supp}(w_2^3) = \{\pm 1\}$;

(4) $\text{supp}(w_1^4) = \{\pm\sqrt{2}/2\}$ y $\text{supp}(w_2^4) = \{0\}$, y

(5) $\text{supp}(w_1^5) = \{\sqrt{2}/2\}$ y $\text{supp}(w_2^5) = \{0, 1\}$.

Nota. Sea (y^*, z^*) la solución óptima de (PA2), cualquier solución óptima del dual (w_1^*, w_2^*) verifica que $\text{constr}_i(y^*) \supseteq \text{supp}(w_i^*)$ para $i = 1, 2$, por el Teorema de holgura complementaria. Luego, para aquellos problemas en los que $|\text{constr}(y^*)| = n + 1$ hay una única solución óptima del dual. Si $|\text{constr}(y^*)| > n + 1$, entonces el problema dual tendrá al menos dos soluciones factibles básicas óptimas.

Resumiendo, el problema (PA2) tiene una única solución óptima, que es un punto extremo (bajo ciertas condiciones, degenerado), pero el problema dual asociado puede tener más de una solución factible básica óptima, dependiendo del número de restricciones que son activas en la solución óptima del primal.

3. GENERACIÓN DE DIRECCIONES DE BÚSQUEDA

La estrategia que planteamos para resolver el problema (PA2) corresponde a un esquema de descenso con direcciones factibles. Estos métodos construyen una secuencia de soluciones factibles y calculan las direcciones de búsqueda resolviendo diferentes problemas de optimización, habitualmente lineales o cuadráticos. En el contexto de la programación semi-infinita han sido propuestos algoritmos de direcciones factibles que trabajan con ciertos conjuntos de mínimos locales de las funciones de holgura sobre el conjunto de parámetros S, previamente discretizado (vease, por ejemplo, Panier y Tits (1989) y Polak y He (1991)). Esto permite obtener reglas eficientes para el esquema de búsqueda, sin incrementar demasiado el coste computacional.

En nuestro caso, para establecer el criterio de optimalidad y para determinar las direcciones de búsqueda intervienen, no sólo las restricciones activas, sino también el conjunto de índices casi-activos que son mínimos locales sobre el conjunto de parámetros S, que no hemos discretizado:

Definición 3.2. *Sea $(y;z)$ una solución factible para el problema (PA2). Denotamos por $m_\varepsilon(y;z_i)$ el conjunto de índices de las restricciones casi-activas que son mínimos locales de las funciones de holgura z_i, para $i = 1,2$, i. e. $m_\varepsilon(y;z_i) = \{s \in S : s$ es un mínimo local de z_i en S y $z_i(y,s) \leq \varepsilon\}$, donde $\varepsilon \geq 0$.*

Hipótesis 3. *Los conjuntos $m_\varepsilon(y;z_i), i = 1,2$, son finitos para todo $y \in \mathbb{R}^{n+1}$ y $\varepsilon \geq 0$.*

Aunque el conjunto $m_\varepsilon(y;z_i)$ podría ser mucho más grande que $\text{constr}_i(y)$, los elementos de $m_\varepsilon(y;z_i)$ tienen un significado específico, para el problema de aproximación, que puede utilizarse para mejorar la eficacia computacional del generador de las direcciones de búsqueda:

Nota. Sea $y = (x_0,x)^T$ una solución factible de (PA2), y sea $E(s) = b(s) - a(s)^T x$ su función de error. Se tiene que $z_1(y,s) = x_0 - E(s)$ y $z_2(y,s) = x_0 + E(s)$, si elegimos $s_1^* \in m_\varepsilon(y;z_1)\backslash \text{constr}_1(y)$ entonces s_1^* es un máximo local para $E(\cdot)$ y $z_2(y,\cdot)$. Y, en consecuencia, $m_\varepsilon(y;z_1) \cap m_\varepsilon(y;z_2) = \emptyset$.

Por otra parte, sabemos que para $s \in \text{constr}(y)$, la función de error $E(s)$ es positiva si $s \in \text{constr}_1(y)$ y negativa si $s \in \text{constr}_2(y)$. Luego, si queremos que el signo del error $E(\cdot)$ en un punto $s_1^* \in m_\varepsilon(y;z_1)$ (respectivamente $s_2^* \in m_\varepsilon(y;z_2)$) sea positivo (negativo), debemos exigir que $x_0 > z_1(y,s_1^*)$ ($x_0 > z_2(y,s_2^*)$, respectivamente). Esta idea sugiere actualizar ε de modo que siempre sea menor que x_0, con lo que se reduce la dimensión de los problemas que generan las direcciones de búsqueda.

El problema (PA1) satisface, por construcción, la cualificación de restricciones de Mangasarian-Fromovitz, pues podemos avanzar a lo largo de la dirección $u = (1,0_n)^T$ sin abandonar el conjunto factible. En consecuencia, las condiciones de optimalidad de tipo Karush-Kuhn-Tucker son necesarias y suficientes para caracterizar la solución óptima del problema.

Teorema 3.1. *Sea ε un escalar no negativo. Una solución factible $(y;z)$ del problema (PA2) es óptima si y sólo si el valor objetivo $v(Q_\varepsilon(y)) = 0$, donde*

$$
\begin{array}{lll}
(Q_\varepsilon(y)) & \text{Min} & d_0 \\
& \text{s.a.} & v_1(s)^T d \geq -z_1(y,s) \quad s \in m_\varepsilon(y;z_1) \\
& & v_2(s)^T d \geq -z_2(y,s) \quad s \in m_\varepsilon(y;z_2) \\
& & -1 \leq d_j \leq 1 \quad j = 0,\ldots,n
\end{array}
$$

Demostración. Sea $(y;z)$ una solución factible tal que $v(Q_\varepsilon(y)) = 0$ y supongamos que no es óptima. Entonces, existe una solución $(y^\#;z^\#)$ de (PA2) tal que $y_0^\# < y_0$, y el segmento $X = (y,y^\#]$ es un subconjunto de soluciones factibles que cumplen que $y_0' < y_0$, para cada $y' \in X$. Podemos elegir $y' \in X$, tal que $\max\left\{\left|y_j' - y_j\right|;\ j = 0,1,\ldots,n\right\} \leq 1$, donde y_j denota la $j-$ésima componente de y, y tomar la dirección $u = y' - y$. Por construcción, $v_i(t)^T u = v_i(t)^T y' - v_i(t)^T y = z_i(y',t) - z_i(y,t) \geq -z_i(y,t)$ para $t \in m_\varepsilon(y;z_i)$ e $i = 1,2$. Luego, u es una solución factible del problema $(Q_\varepsilon(y))$ y $u_0 = y_0' - y_0 < 0$, con lo que se llega a contradicción.

Recíprocamente, sea $(y;z)$ el óptimo del problema (PA2). Nótese que, para $\varepsilon = 0$: $m_0(y;z_i) = \text{constr}_i(y)$ y tenemos exactamente una condición de optimalidad de tipo Karush-Kuhn-Tucker para (PA2) formulada como un programa lineal, entonces $v(Q_0(y)) = 0$. Además, para todo $0 \leq \varepsilon' < \varepsilon$ se tiene que $0 \leq v(Q_{\varepsilon'}(y)) \leq v(Q_\varepsilon(y))$. Por consiguiente $v(Q_\varepsilon(y)) = 0$, pues $d = 0$ es una solución factible del problema lineal $Q_\varepsilon(y)$.

■

Nota. Para determinar la magnitud de avance máxima μ^* a lo largo de una dirección cualquiera d, es fácil ver que debe calcularse $\mu_i = \inf\{-z_i(y,s)/v_i(s)^T d$: para $s \in S$ tal

que $v_i(s)^T d < 0\}$, para $i = 1,2$. Si los subconjuntos $S_i := \{s \in S : v_i(s)^T d < 0\}$ son no vacíos para $i = 1,2$, entonces $\mu^* = \min\{\mu_1, \mu_2\}$ y se asegura que el segmento $[y, y + \mu^* d]$ se mantiene factible. Además, si $S_1 = \emptyset$, existe un $s' \in S$ tal que $v_2(s')^T d < 0$, pues el problema (PA2) es acotado y no pueden existir direcciones de recesión para el conjunto factible. En consecuencia, $\mu^* = \mu_2$. Análogamente, si $S_2 = \emptyset$, resulta que $\mu^* = \mu_1$.

Para calcular la magnitud del avance la formulación anterior no es manejable, así que a efectos prácticos la hemos sustituido por otra equivalente. En lo que resta de sección, y sin pérdida de generalidad, supondremos que las direcciones involucradas cumplen que $S_i \neq \emptyset$ para $i = 1,2$ y que se han construido de forma que no se planteen problemas en los índices activos, i.e. si $z_i(y, s^*) = 0$, entonces d se genera de forma que $v_i(s^*)^T d > 0$, para $i = 1,2$.

Lema 3.2. *Supongamos que $(y;z)$ es una solución factible, no óptima, del problema (PA2). Sea $d \in \mathbb{R}^{n+1}$, definimos $\mu_i := inf\{-z_i(y,s)/v_i(s)^T d : para\ s \in S\ tales\ que\ v_i(s)^T d < 0\}$ para $i = 1,2$. Entonces $\mu_i = 1/\lambda_i$ donde $\lambda_i = sup\{-v_i(s)^T d/z_i(y,s) : s \in S\}$.*

Demostración. Vamos a comprobar que λ_i para $i = 1,2$ no puede alcanzarse en $s \in S$ tal que $v_i(s)^T d \geq 0$. Sea $L_i(s) := -v_i(s)^T d/z_i(y,s)$, entonces, para cualquier $s \in S$ tal que $v_i(s)^T d \geq 0$, se tiene que:

(a) si $z_i(y,s) > 0$, entonces $L_i(s) \leq 0$, y

(b) si $z_i(y,s) = 0$, por construcción se tiene que $v_i(s)^T d > 0$, entonces $L_i(s) = -\infty$.

Sin embargo para $s \in S$ tal que $v_i(s)^T d < 0$, se tiene que $L_i(s) > 0$, puesto que $z_i(y,s) > 0$. Por tanto, $1/\lambda_i = \inf\{-z_i(y,s)/v_i(s)^T d : s \in S$ tales que $v_i(s)^T d < 0\}$, siendo $\lambda_i = \sup\{L_i(s) : s \in S\}$.

■

Para cada solución factible $(y;z)$ y $\varepsilon \geq 0$, denotamos por $d_\varepsilon^*(y)$ la solución óptima del problema lineal $Q_\varepsilon(y)$. Cuando se comprueba que $(y;z)$ no es la solución óptima del (PA2) parece lógico usar la dirección $d_\varepsilon^*(y)$ como dirección de descenso. Sin embargo, hemos comprobado que en numerosas ocasiones no es factible, pues la amplitud de avance asociada vale cero. Para obviar este problema se resuelve un segundo programa lineal:

$$
\begin{array}{lll}
(Q'_\varepsilon(y,d^*)) & \text{Min} & g_0 \\
& \text{s.a.} & v_1(s)^T g \geq -z_1(y,s) - 0.75 d_0^* \quad s \in m_\varepsilon(y;z_1) \\
& & v_2(s)^T g \geq -z_2(y,s) - 0.75 d_0^* \quad s \in m_\varepsilon(y;z_2) \\
& & -1 \leq d_j \leq 1 \quad j = 0, \ldots, n
\end{array}
$$

donde d_0^* es la primera componente del vector $d_\varepsilon^*(y)$. Si denotamos por $g_\varepsilon^*(y)$ su solución óptima se comprueba que $d := d_\varepsilon^*(y) + g_\varepsilon^*(y)$ es una dirección de descenso que, además, siempre es factible.

Lema 3.3. *Sea (y,z) una solución factible del problema (PA2) y $\varepsilon \geq 0$. Sea $d \in \mathbb{R}^{n+1}$ tal que $v_1(s)^T d > -z_1(y,s)$, para $s \in m_\varepsilon(y;z_1)$, y $v_2(s)^T d > -z_2(y,s)$, para $s \in m_\varepsilon(y;z_2)$ y $-1 \leq d_j \leq 1$, para $j = 0,1,\ldots,n$. Entonces, la amplitud de avance máxima a lo largo de d es positiva.*

Demostración. Por el Lema 3.2 sabemos que la amplitud de avance máxima a lo largo de d se calcula como $\mu^* = \min\{\mu_1,\mu_2\}$ y $\mu_i = 1/\lambda_i$ donde $\lambda_i = \sup\{-v_i(s)^T d/z_i(y,s) : s \in S\}$, para $i = 1,2$. Así pues, se trata de asegurar que $\lambda_i < +\infty$, i.e. que las funciones $L_i(s)$ estan acotadas superiormente en S.

Por construcción de d tenemos que $v_i(s')^T d > 0$, para $s' \in \text{constr}(y;z_i)$, luego para cada índice activo s^* existe un entorno abierto $U(s^*)$ en S en el que no se puede alcanzar ese supremo. Para los índices $s \in S$ comprendidos en los subintervalos cerrados restantes, limitados por entornos de índices activos consecutivos, la función $L_i(s)$ está acotada. Y, por la Hipótesis 3, tenemos un número finito de estos subintervalos, luego la función está acotada superiormente en S.

■

Nota. Debido a la estructura del problema (PA2), y siguiendo el lema anterior, puede comprobarse que cualquier vector $d \in \mathbb{R}^{n+1}$, que coincida con $d_\varepsilon^*(y)$ excepto en la primera componente (que podría tomarse como $d_0 = -\gamma d_0^*$, $\gamma \in (0,1)$) serviría como dirección «auxiliar», de modo que sumándola a $d_\varepsilon^*(y)$ permite obtener una dirección factible de descenso. Sin embargo, hemos implementado la construcción de la dirección de búsqueda a partir de $d_\varepsilon^*(y) + g_\varepsilon^*(y)$, pues podrá utilizarse para resolver problemas más generales que el (PA2).

4. DESCRIPCIÓN DE LOS ALGORITMOS

Vamos a describir un esquema de direcciones factibles con un generador de direcciones de búsqueda basado en los resultados de la Sección 3, que nos asegura la factibilidad de todos los iterados si partimos de una solución factible inicial y la disminución del valor de la función objetivo del problema (PA2).

Se pueden conseguir diferentes implementaciones del método dependiendo de la forma de actualizar el parámetro ε. Valores pequeños de ε reducen el número de restric-

ciones que definen el problema $Q_\varepsilon(y)$, y valores suficientemente grandes hacen que el conjunto $m_\varepsilon(y;z_i)$ contenga todos los mínimos locales y el algoritmo no se vea afectado por índices que entran y salen alternativamente de dicho conjunto. Hemos trabajado fijando $\varepsilon = \infty$, así como con otras estrategias que hacen tender a cero el valor de ε mediante actualizaciones adecuadas. Se ha comprobado que la actualización más eficiente, y también más robusta, para variaciones de la solución factible inicial, es la de elegir $\varepsilon = x_0$ para $y = (x_0, x)$. Así, pues, hemos llamado método A1 al que funciona del modo siguiente:

Método A1.

Inicialización. Elegir una solución factible $(y^1; z(y^1, s)) \in \mathbb{R}^{n+1} \times C(S)^2$ de (PA2), $\delta > 0$ y hacer $k = 1$.

Etapa 1. Para $y^k = (x_0, x)$, tomar $\varepsilon_k = x_0$ y calcular $m_{\varepsilon k}(y^k; z_i(y^k, s))$ para $i = 1, 2$.

Etapa 2. Calcular $d^*(y^k) = (d_0^*, d_1^*, \ldots, d_n^*)$ una solución óptima del problema lineal $Q_{\varepsilon k}(y^k)$.
Si $d_0^* < -\delta$, STOP. En otro caso, ir a la Etapa 3.

Etapa 3. Calcular $\mu^*(y^k) = \min\{\mu_1, \mu_2\}$, siendo
$\mu_i = \inf\{-z_i(y^k, s)/v_i(s)^T d^*(y^k)$: para $s \in S$ tal que $v_i(s)^T d^*(y^k) < 0\}$, para $i = 1, 2$.
Si $\mu^*(y^k) > 0$, hacer $y^{k+1} = y^k + \mu^*(y^k) d^*(y^k), k = k + 1$ y volver a la Etapa 1.
En otro caso, ir a la Etapa 4.

Etapa 4. Calcular $g^*(y^k)$, solución óptima del problema lineal $Q'_{\varepsilon k}(y^k, d^*)$.
Hacer $d^*(y^k) = d^*(y^k) + g^*(y^k)$ e ir a la Etapa 3.

Nota. En la Etapa 3 se determina aquel valor μ^* que incorpora al menos una nueva restricción al conjunto de los ceros de las funciones de holgura. Se ha visto en el Lema 3.3 que la dirección generada en la Etapa 4 tiene asociada una amplitud de avance máxima que es positiva. Así pues, el algoritmo no puede ciclarse.

Dado que la solución óptima del problema (PA2) es un punto extremo, nos decidimos a incluir en el anterior esquema de direcciones factibles una etapa de purificación que, utilizando los resultados de la Sección 2, proporciona una solución factible básica. Hemos comprobado que esta etapa interna acelera la disminución del valor objetivo en las etapas iniciales. Cuando la cardinalidad del conjunto $m_\varepsilon(y;z)$ y la localización de los mínimos locales se fija en torno a los valores que corresponden a la solución óptima, el método A1 puede evolucionar directamente de un punto extremo a otro.

Nos planteamos si debía ejecutarse la etapa de purificación para cada iterado que no sea solución factible básica o sólo cada cierto número de iteraciones y hemos comprobado que es más eficiente este segundo criterio. El estudio comparativo realizado en León, Sanmatías y Vercher (1996) muestra claramente la ventaja de purificar cada $n+1$ iteraciones. Los resultados computacionales para una gran cantidad de instancias mostraban un ahorro del 33% en el número de iteraciones al utilizar este criterio. Así pues, se ha implementado un método híbrido, que llamamos método A2, que combina etapas de purificación con el generador de direcciones de búsqueda anterior.

Método A2.

Inicialización. Elegir una solución factible $(y^1;z(y^1,s)) \in \mathbb{R}^{n+1} xC(S)^2$ de (PA2), tomar $\varepsilon = \infty$, $\delta > 0$ y $N > 1$, hacer $Y = y^1$ y $k = 1$.

Etapa Interna. Calcular constr(Y). Si $V(Y) \neq \{0_{n+1}\}$ aplicar un algoritmo de purificación hasta llegar a un punto extremo $(y^k;z(y^k,s))$.

Etapa 1. Para $y^k = (x_0,x)$, tomar $\varepsilon_k = x_0$ y calcular $m_{\varepsilon k}(y^k;z_i(y^k,s))$ para $i = 1,2$.

Etapa 2. Calcular $d^*(y^k) = (d_0^*,d_1^*,\ldots,d_n^*)$, una solución óptima del problema lineal $Q_{\varepsilon k}(y^k)$.
Si $d_0^* < -\delta$, STOP. En otro caso, ir a la Etapa 3.

Etapa 3. Calcular $\mu^*(y^k) = \min\{\mu_1,\mu_2\}$, siendo
$\mu_i = \inf\{-z_i(y^k,s)/v_i(s)^T d^*(y^k)$: para $s \in S$ tal que $v_i(s)^T d^*(y^k) < 0\}$, para $i = 1,2$.
3.1. Si $\mu^*(y^k) > 0$, hacer $y^{k+1} = y^k + \mu^*(y^k)d^*(y^k)$.
Si $k < N$, hacer $k = k+1$ y volver a la Etapa 1.
Si $k = N$ hacer $Y = y^{k+1}$, $k = 1$ y volver a la Etapa Interna.
3.2. Si $\mu^*(y^k) = 0$, ir a la Etapa 4.

Etapa 4. Calcular $g^*(y^k)$, solución óptima del problema lineal $Q'_{\varepsilon k}(y^k,d^*)$.
Hacer $d^*(y^k) = d^*(y^k) + g^*(y^k)$ y volver a la Etapa 3.

En resumen, el método A2 encuentra un punto extremo como solución inicial, ejecuta durante N iteraciones un esquema de direcciones factibles y para proseguir comprueba si el iterado $N-$ésimo es punto extremo, i.e. la etapa interna de purificación se aplica cada N iteraciones del método. Para realizar esta etapa interna hemos extendido un algoritmo de purificación (León y Vercher (1992b)) que estaba diseñado para una única función de holgura y que encuentra un punto extremo del problema (PA2) como máximo en $(n+1)$ iteraciones a partir de cualquier solución factible.

Pensando en la aplicación de este esquema híbrido para la resolución de problemas más generales, hemos programado una versión del método A2 que no actualiza el

parámetro ε, y lo hemos llamado método A3. No hemos considerado necesario introducir su esquema algorítmico, pues la única diferencia con el anterior es que en cada iteración se utilizan todos los mínimos locales para generar las direcciones de búsqueda.

5. RESULTADOS NUMÉRICOS

Todos los métodos se han implementado en lenguaje C en un ordenador personal HP Vectra VL4/100 y funcionan con doble precisión. Respecto a los parámetros introducidos, queremos señalar que: el criterio de parada que hemos utilizado es $\delta = 10^{-10}$; un mínimo local se considera un cero de una función de holgura cuando su valor es menor que la precisión standard 10^{-8}; y hemos tomado $N = n+1$, pues se ha considerado el papel de la etapa interna de purificación como una especie de «restart» del método de direcciones factibles.

En todos los métodos se calcula el conjunto de mínimos locales, $m_\varepsilon(y;z_i)$, utilizando un algoritmo de búsqueda multi-local unidimensional (León, Sanmatías y Vercher (1998)), que resuelve también eficientemente el cálculo de la amplitud de avance en la Etapa 3.

Las direcciones de búsqueda $d_\varepsilon^*(y), g_\varepsilon^*(y)$ y el valor de la función objetivo $v(Q_\varepsilon(y))$ se obtienen resolviendo los programas lineales correspondientes mediante subrutinas de CPLEX Callable Library. Asimismo, las direcciones de búsqueda del algoritmo de purificación se calculan utilizando estas subrutinas de CPLEX. Hay que señalar que en el contador de iteraciones de los métodos A2 y A3 están incluidas las búsquedas-iteraciones realizadas en la etapa interna de purificación.

Para comparar el comportamiento de los métodos introducidos hemos resuelto algunos problemas de aproximación, y presentamos los resultados obtenidos en las tablas siguientes. En la primera columna aparecen las características de cada problema resuelto y la solución inicial utilizada: $y = (M, 0_n)$, siendo M una cota superior de $|b(s)|$ en S. Las otras columnas incluyen: el número de iteraciones (#iter); el número de evaluaciones de las funciones de holgura (#eval_z); el tamaño medio de los problemas lineales resueltos (PL_mean) y el valor de la función objetivo del problema (PA2) en el óptimo (error_máximo). En los problemas de la Tabla 1 el conjunto de funciones aproximantes es el de los polinomios de grado menor que n: $\{1, s, s^2, \ldots, s^{n-1}\}$. Para los problemas de la Tabla 2 se han utilizado diferentes conjuntos de funciones aproximantes: $\{\sin(s), \cos(s)\}$, $\{1, s, \sin(s), \cos(s)\}$, $\{1, s, s^2, s^3, \sin(s), \cos(s)\}$ y $\{1, s, 2s^2 - 1, 4s^3 - 3s, 8s^4 - 8s^2 + 1, 16s^5 - 20s^3 + 5s\}$, respectivamente.

Realmente, el mayor esfuerzo computacional se hace al calcular la amplitud de avance $\mu^*(y)$, pues las direcciones de búsqueda (tanto en la etapa interna como en la Etapa

2) se obtienen como solución de problemas lineales de tamaño menor que $n+2$, lo que se corresponde con la cardinalidad del conjunto de mínimos locales. En las primeras iteraciones hay relativamente pocos mínimos locales, posteriormente este número tiende hacia la cardinalidad del conjunto constr(y^*), donde y^* es el óptimo.

Tabla 1

$b(s),S,n,y^0$	método	#iter	eval_z	PL_mean	error_máximo
$\sin(s),[0,1]$	A1	18	936	3.13	.000155410916
$n=4$	A2	20	1179	3.69	.000155406095
$(\sin(1),0)$	A3	12	699	3.35	.000155407474
$\cos(s),[-1,1]$	A1	23	1843	4.37	.000041877535
$n=5$	A2	25	2124	4.89	.000041879797
$(1,0)$	A3	16	1301	5.16	.000041898024
$1/1+s^2,[0,1]$	A1	29	2328	4.78	.000521989105
$n=5$	A2	24	1797	4.19	.000521989396
$(1,0)$	A3	19	1504	4.45	.000522046387
$\tan(s),[-1,1]$	A1	22	2506	6.24	.002602827389
$n=6$	A2	35	3293	4.87	.002628279523
$(\tan(1),0)$	A3	15	1857	6.04	.002602824791
$1-\exp(-s^2)$	A1	50	5733	5.48	.014256567265
$[-2,2],n=7$	A2	35	4544	6.29	.014256566041
$(b(2),0)$	A3	18	2725	7.61	.014256634097

Tabla 2

$b(s),S,n,y^0$	método	#iter	eval_z	PL_mean	error_máximo
$1/1+s^2,[0,1]$	A1	11	389	2.53	.030554447874
$n=2$	A2	7	303	2.40	.030554446088
$(1,0)$	A3	6	257	3.11	.030554466387
$1/1+s^2,[0,1]$	A1	20	1120	4.00	.002099731524
$n=4$	A2	18	986	3.48	.002099728652
$(1,0)$	A3	13	757	3.70	.002099741233
$1-\exp(-s2)$	A1	12	1227	5.47	.054222163691
$[-2,2],n=6$	A2	25	2446	5.16	.054222163691
$(b(2),0)$	A3	18	1663	5.39	.054222195516
$1-\exp(-s^2)$	A1	40	4227	5.07	.163381197911
$[-3,3],n=6$	A2	64	6328	5.03	.163381193882
$(b(3),0)$	A3	30	2560	5.11	.163381228959

Ya se había indicado que el método A1 era el más eficiente para diferentes actualizaciones del parámetro ε, incluido $\varepsilon = \infty$. Añadirle una etapa interna de purificación no resulta demasiado ventajoso, aunque en algunos problemas pueda dar mejor resultado. Una posible explicación de este comportamiento es que se están aplicando dos criterios que producen efectos contrarios dentro del mismo método A2: por una parte, reducir el número de restricciones casi-activas (por debajo de x_0) y, por otra, completar el número de ceros (en la etapa interna) para alcanzar un punto extremo.

En cambio, el método A3, que genera las direcciones de búsqueda utilizando todos los mínimos locales, y que también aplica la etapa de purificación cada cierto número de búsquedas, es el más eficiente de entre todos los métodos que hemos programado. El método A3 necesita menos iteraciones para llegar a la solución óptima y la encuentra con menor coste computacional. También se ha comportado mejor que una implementación anterior en la que aplicábamos la etapa interna a cada solución factible no básica.

La comparación con los resultados obtenidos por otros algoritmos primales de programación semi-infinita para resolver el problema (PA2) puede ser poco representativa, pues esos métodos incorporan una etapa previa de discretización del conjunto S y luego resuelven el problema finito resultante mediante diferentes esquemas (p.e. punto interior en Ferris y Philpott (1989), refinamiento de la discretización en Hettich y Gramlich (1990)). Otra dificultad adicional radica en la diferencia de precisión utilizada y en el hecho de que no se indican las soluciones iniciales.

Tabla 3

$b(s), S, n, y^0$	método	#iter	eval_z	PL_mean	error_máximo
$s^2, [0,2]$					
$n = 2$	NEW	11	13420	62	0.53824312
$(x_0, 1, 1)$	A3	19	499	1987	0.5382453186
$\sin(s), [0,1]$					
$n = 3$	NEW	7	8920	50	0.00450552
$(x_0, 1, 1, 1)$	A3	9	406	848	0.0045050895

En el trabajo de Zhou y Tits (1996) se introduce un algoritmo de programación cuadrática secuencial, que resuelve problemas continuos minimax para una discretización fina (501 puntos equidistantes) del intervalo S y dos de los problemas planteados son de la forma (PA1). El primero coincide con el Ejemplo 2.2 y el segundo consiste en aproximar la función $\sin(s)$ utilizando los polinomios $\{1, s, s^2\}$. En la Tabla 3 incluimos los resultados obtenidos con su algoritmo NEW, el más eficiente de los presentados, y añadimos el número de evaluaciones de las derivadas de las funciones

de holgura (#eval_zd). El comportamiento del método A3 ha sido muy diferente para estos dos problemas. En el Ejemplo 2.2 ha sido computacionalmente costoso mantener la factibilidad de cada iterado en un entorno del índice $s = 0.40637642$ para la función de holgura z_2 —donde $L_2(s)$ presenta una discontinuidad evitable—. Se ha calculado la amplitud de avance máxima con mucha precisión para evitar la imposibilidad primal, y ese esfuerzo queda reflejado claramente en la columna #eval_zd. Este problema no se presenta cuando se está trabajando con un subconjunto discreto de S.

Pensamos que el método A3 puede ser eficiente también para resolver otros problemas semi-infinitos lineales, pues tanto el esquema que genera las direcciones factibles como la etapa de purificación son generales.

REFERENCIAS BIBLIOGRÁFICAS

[1] **Anderson, E.J.** y **Lewis, A.S.** (1989). «An extension of the simplex algorithm for semi-infinite linear programming». *Mathematical Programming*, **44**, 247–269.

[2] **Anderson, E.J.** y **Nash, P.** (1987). *Linear Programming in infinite-dimensional spaces*. John Wiley and Sons, Chichester.

[3] **Andreasson, D.O.** y **Watson, G.A.** (1976). «Linear Chebyshev approximation without Chebyshev sets». *BIT*, **16**, 349–362.

[4] **Cheney, E.W.** (1966). *Introduction to Approximation Theory*. McGraw-Hill, New York.

[5] **Ferris, M.C.** y **Philpott, A.B.** (1989). «An interior point algorithm for semi-infinite linear programming». *Mathematical Programming*, **43**, 257–276.

[6] **Glashoff, K.** y **Gustafson, S.A.** (1983). *Linear Optimization and Approximation*. Springer Verlag, New York.

[7] **Hettich, R.** y **Gramlich, G.** (1990). «A note on an implementation of a method for quadratic semi-infinite programming». *Mathematical Programming*, **46**, 249–254.

[8] **Hettich, R.** y **Kortanek, K.O.** (1993). «Semi-infinite Programming: Theory, Methods and Applications». *SIAM Review*, **35**, 380–429.

[9] **León, T., Sanmatías, S.** y **Vercher, E.** (1996). «Some semi-infinite programming algorithms for solving the Linear Chebyshev Approximation Problem». *Technical Report*, **#3-96**. Depart. Estadística e Investigación Operativa. Universitat de València.

[10] **León, T., Sanmatías, S.** y **Vercher, E.** (1998). «A multi-local optimization algorithm». Aparecerá en *Top*.

[11] **León, T.** y **Vercher, E.** (1992a). «An optimality test for semi-infinite linear programming». *Optimization*, **26**, 51–60.

[12] **León, T.** y **Vercher, E.** (1992b). «A purification algorithm for semi-infinite programming». *European Journal of Operational Research*, **57**, 412–420.

[13] **León, T.** y **Vercher, E.** (1994). «New descent rules for solving the linear semi-infinite programming problem». *Operations Research Letters*, **15**, 105–114.

[14] **Nash, P.** (1985). «Algebraic fundamentals of linear programming». *Infinite Programming*, Edited by E.J. Anderson and A.B. Philpott, Springer, Berlin, 37–52.

[15] **Panier, E.** y **Tits, A.** (1989). «A globally convergent algorithm with adaptively refined discretization for Semi-Infinite Optimization Problems arising in Engineering Design». *IEEE Transactions on Automatic Control*, **34**, 903-908.

[16] **Polak, E.** y **He, L.** (1991). «Unified Steerable PhaseI-Phase II Method of feasible Directions for Semi-Infinite Optimization». *Journal of Optimization Theory and Applications*, **69**, . 83–107.

[17] **Powell, M.J.D.** (1981). *Approximation Theory and Methods*. Cambridge University Press, Cambridge.

[18] **Zhou, J.L.** y **Tits A.L.** (1996). «An SQP algorithm for finely discretized continuous minimax problems and other minimax problems with many objective functions». *SIAM J. Optimization*, **6**, 461–487.

ENGLISH SUMMARY

A PRIMAL SEMI-INFINITE PROGRAMMING METHOD FOR THE UNIFORM APPROXIMATION PROBLEM*

T. LEÓN
S. SANMATÍAS
E. VERCHER
Universitat de València*

In this paper we develop an algorithm that solves the classic uniform approximation problem formulated as a semi-infinite program. We have studied the algebraic characterization of extreme points and proved some of their properties. We have designed a procedure that generates feasible directions by solving certain linear programs, that also characterize the optimum. The method incorporates a purification algorithm to proceed from a feasible solution to an improved extreme point. Finally, we show the good performance of the algorithm at a number of numerical examples.

Keywords: Semi-infinite programming, uniform approximation, extreme points, feasible direction methods.

AMS Classification: 90C34, 41A50, 49M35

⋆This work has been supported by the Ministerio de Educación y Ciencia, España, DGICYT, PB93-0703.

*Departament d'Estadística i Investigació Operativa. Facultat de Matemàtiques. Universitat de València. C/ Dr. Moliner, 50. 46100 Burjassot (Valencia).

–Received June 1997.
–Accepted March 1998.

1. INTRODUCTION

The uniform approximation problem can be stated as follows: given a continuous function $b(\cdot)$ defined on $[\alpha,\beta]$, our interest is to approximate it by a linear combination of a given set of continuous functions $\{a_1,a_2,\ldots,a_n\}$ also defined on $[\alpha,\beta]$. Under certain conditions it has been shown that the maximum error in the best approximation occurs at least at $n+1$ points in $[\alpha,\beta]$, with alternating sign and that this best approximation is unique. There are some ascent methods that solve it by using the above property (see Powell (1981), for instance).

So, the objective is to minimize the maximum difference between the function $b(\cdot)$ and the approximating function

$$P(x,s) := x_1 a_1(s) + x_2 a_2(s) + \cdots + x_n a_n(s), \quad \text{where} \quad x_1,\ldots,x_n \in \mathbb{R},$$

that is to minimize the maximum error $x_0 := \max |b(s) - P(x,s)|$. It is well known that this problem can be written as a semi-infinite program:

$$\begin{array}{lll} \text{(PA1)} \quad \text{Minimize} & x_0 & \\ \text{s.a.} & x_0 + a(s)^T x \geq b(s) & \\ & x_0 - a(s)^T x \geq -b(s) & \\ & x_0 \in \mathbb{R}, \quad x \in \mathbb{R}^n, \quad s \in [\alpha,\beta] & \end{array}$$

The approximation problems and its applications provide a broad class of semi-infinite programming problems (see for instance Hettich and Kortanek (1993)).

2. EXTREME POINTS

Nash (1985) defines, in an algebraic framework, the fundamental objects and operations of linear programming posed in infinite dimensional spaces. Indeed, he extends the definitions of degenerate and non degenerate basic feasible solution, reduced costs and pivoting of finite linear programming to the infinite dimensional case. We have studied the adaptation of their characterizations of degenerate and non degenerate basic feasible solutions for our problerm and we have derived some properties concerning to them.

Our main result in this section, Theorem 2.3, shows that, under certain mild conditions and $n \geq 2$, the optimal solution of the semi-infinite problem (PA1), stated in the standard LP form, is a degenerate extreme point.

3. SEARCH DIRECTION FINDING SCHEME

We give a feasible directions scheme that constructs a sequence of feasible solutions by solving finite linear programs. The search direction computation takes into account the gradients of the near-binding local minimizer constraints for the current iterate and it uses a *Karush-Kuhn-Tucker* type optimality criterion.

Let us denote by $y = (x_0, x)$, $v_1(s) = (1, a_1(s), a_2(s), \ldots, a_n(s))^T$, $v_2(s) = (1, -a_1(s), -a_2(s), \ldots, -a_n(s))^T$ for $s \in S$. Then we have the following result:

Theorem 3.1. *Let ε be a non negative scalar, then a feasible solution $(y;z)$ is optimal if and only if the objective function value $v(Q_\varepsilon(y)) = 0$, where*

$$
\begin{array}{lll}
(Q_\varepsilon(y)) \quad \text{Min} & d_0 & \\
\text{s.a.} & v_1(s)^T d \geq -z_1(y,s) & s \in m_\varepsilon(y;z_1) \\
& v_2(s)^T d \geq -z_2(y,s) & s \in m_\varepsilon(y;z_2) \\
& -1 \leq d_j \leq 1 & j = 0, \ldots, n
\end{array}
$$

where $z_1(\cdot)$ and $z_2(\cdot)$ are the slack functions defined as: $z_1(y,s) = v_1(s)^T y - b(s)$, $z_2(y,s) = v_2(s)^T y + b(s)$ for $s \in S$, and $m_\varepsilon(y;z_i)$ denotes the set of indices of near-binding constraints that are local minima of the slack function $z_i(\cdot)$ for $i = 1,2$, that is $m_\varepsilon(y;z_i) = \{s \in S : s$ is a local minimizer of zi in S and $z_i(s) \leq \varepsilon\}$.

For computational reasons, we assume that $m_\varepsilon(y;z_i)$ is finite for all $y \in \mathbb{R}^{n+1}$ and $\varepsilon \geq 0$. The analytical functions, for instance, satisfy this assumption.

At each iteration the algorithm solves the LP problem above, for a feasible point $(y;z)$ and $\varepsilon \geq 0$. Let $d_\varepsilon^*(y)$ indicate the optimal solution of $(Q_\varepsilon(y))$. It is accepted that the main drawback to obtain an efficient primal method for semi-infinite linear programming problems is in the computation of the maximum steplenght $\mu^*(y)$. Concerning to this question, we have proved the following result:

Lemma 3.2. *Suppose that $(y;z)$ is a non-optimal feasible solution, and $d_\varepsilon^*(y)$ is an optimal solution of $(Q_\varepsilon(y))$. Let $\mu_i = inf\{-z_i(y,s)/v_i(s)^T d_\varepsilon^*(y)$: for $s \in S$ such that $v_i(s)^T d_\varepsilon^*(y) < 0\}$, then $\mu_i = 1/\lambda_i$ where $\lambda_i = max\{-v_i(s)^T d_\varepsilon^*(y)/z_i(y,s) : s \in S\}$ for $i = 1,2$.*

In some cases, the maximum steplength along $d_\varepsilon^*(y)$ could be zero. So, we must construct a suitable perturbation of the direction in order to guarantee the movement in the neighbourhood of (y,z) without encountering a boundary. This modification is, in fact, the addition of another vector obtained through the solution of a finite LP subproblem.

4. DESCRIPTION OF THE ALGORITHMS

First, we consider a pure feasible directions scheme, based on Theorem 3.1 as the generator of search directions. Depending on the updating rules for ε, we obtain diferent versions of the method. After an empirical study, we decided that the best choice is setting $\varepsilon = x_0$, for $y = (x_0, x)$, to give that we call method A1.

With the aim of deciding if the addition of a purification step implies some improvement over method A1, we have implemented a version of the method that includes it. A purification step can be performed every time we get a non basic feasible solution or after a certain number (fixed or not) of iterations, and its efficacy is quite different. We stated method A2 in such a way that it finds an extreme point through an initial feasible solution, then along N iterations it uses a pure feasible-direction scheme and it checks whether the $N+1$ iterate is an extreme point, otherwise a purification step is performed, and so on. Hence it applies the purification step at most every N iterations, where $N = n+1$ is the dimension of the vector y.

We have called method A3 to a modification of method A2 that sets $\varepsilon = \infty$, that is it does not update ε at each iteration.

5. NUMERICAL RESULTS

In order to compare the performance of the methods we have made some numerical experiments. We present our results at three tables in which some relevant characteristics appear. We could say that method A3 is the most efficient, in the sense that it needs less iterations to find the optimum. Moreover this method could be also useful for solving other semi-infinite linear programming problems, because both the direction generating scheme and the purification step have been stated in a general form.

Estadística Oficial

QÜESTIIÓ, vol. 22, 2, p. 339-361, 1998

MÈTODE PER A L'ESTIMACIÓ DE MIGRACIONS EN ÀREES PETITES A PARTIR DE DADES INCOMPLETES

J.A. SÁNCHEZ CEPEDA
Institut d'Estadística de Catalunya*

La disposició de dades de migració per a àrees petites s'emmarca dins el projecte d'estimacions de població, consistent en determinar la població recent de Catalunya per comarques i municipis majors de 45.000 habitants, desagregada per sexe i edat. La no disposició de dades amb aquest nivell de desagregació per alguns anys ha portat a la implementació d'un mètode per a la seva estimació.

Aquest treball descriu el mètode d'estimació dels efectius migratoris per àrea geogràfica, sexe i edat que s'aplica a l'Institut d'Estadística de Catalunya (Idescat). El mètode consisteix en calcular per separat la intensitat i el perfil tant de l'emigració com de la immigració. El principal problema en l'estimació del nivell o intensitat migratoria és haver de treballar amb àrees menors de 45.000 habitants. En el càlcul del perfil migratori es redueix el nombre de perfils dels 62 inicials a 6, que constitueixen els patrons migratoris observats a Catalunya. Es comenta la tipologia dels patrons migratoris, en base als paràmetres dels mateixos determinats pel model. Finalment, es calcula el nombre d'emigrants i immigrants a partir de la intensitat i el perfil migratoris estimats.

Method for estimating small area migration from incomplete data.

Paraules clau: Àrea petita, model migratori, anàlisi cluster, ajust no lineal.

* Institut d'Estadística de Catalunya. Via Laietana, 58. 08003 Barcelona.
- Rebut el novembre de 1997.
- Acceptat l'abril de 1998.

1. INTRODUCCIÓ

El coneixement de l'evolució anual dels efectius demogràfics és necessari, tant per ser una dada bàsica per a l'assignació de recursos com per ser imprescindible per al càlcul d'indicadors demogràfics i socio-econòmics. L'Institut d'Estadística de Catalunya (Idescat) realitza les estimacions de població, consistents en calcular la població catalana recent pel mètode dels components: la població en un any Y es calcula sumant a la població de l'any anterior els naixements i la immigració, i restant les defuncions i l'emigració. La població es desagrega en 2 sexes, 100 grups d'edat i 62 àrees geogràfiques: els 21 municipis de més de 45.000 habitants i les 41 comarques i restes comarcals (entenem per resta comarcal el total de la comarca menys els seus municipis de més de 45.000 habitants). De les 62 àrees, 23 tenen menys de 45.000 habitants i 25 tenen entre 45.000 i 100.000.

Un dels objectius d'aquest projecte és poder disposar d'una sèrie temporal coherent i amb un important nivell de desagregació geogràfica i d'edat; aquesta sèrie comença l'any 1986. El registre actual de la migració és l'Estadística de Variacions Residencials (EVR), que permet conèixer les dades amb un bon nivell de detall. Es presenten, però, 2 problemes: en primer lloc, l'origen de l'EVR data de 1988, de manera que manquen dades pels anys 1986 i 1987; en segon lloc, en els anys padronals i/o censals, les EVR registren un significatiu descens, de manera que el seu valor no és coherent dins la sèrie temporal. Aquestes 2 situacions han portat a la creació d'un mètode per a l'estimació de la migració en àrees petites.

A continuació, es defineixen els indicadors i les variables que intervenen en el càlcul de la migració. Les definicions són anàlogues per a l'emigració i la immigració, de manera que es parlarà de taxa de migració i índex sintètic de migració, GMR (gross migraton rate). Això no vol dir, però, que es tracti la migració en conjunt, sinó que tot el mètode s'aplica 2 vegades, 1 per a la immigració i 1 per a l'emigració.

Siguin

a = àrea geogràfica

s = sexe

e = edat

M = migrants

P = població

Y = any

$t_Y(a, s, e) = M(a, s, e)/P_Y(a, s, e)$ taxa de migració per àrea, sexe i edat en un any Y

$t_Y(a, e) = M(a, e)/P_Y(a, e)$ taxa de migració per àrea i edat en un any Y

$\mathrm{GMR}_Y(a, s) = \Sigma_{e=0,1,\ldots,99}\, t_Y(a, s, e)$ índex sintètic de migració per àrea i sexe

$\mathrm{GMR}_Y(a) = \Sigma_{e=0,1,\ldots,99}\, t_Y(a, e)$ índex sintètic de migració per àrea i total sexes

$p_Y(a, s, e) = t_Y(a, s, e)/\mathrm{GMR}_Y(a, s)$ pes d'una edat dins el conjunt de taxes

$(p_Y(a, s))_{e=0,1,\ldots,99}$ perfil migratori

Notem que $\|(p_Y(a, s))_{e=0,1,\ldots,99}\|_1 = 1$, fet que permet comparar diferents perfils a la mateixa escala.

Amb aquestes definicions s'identifiquen i separen els 2 elements constitutius de la migració: el nivell (o intensitat) i el perfil. Donat que el que es vol és estimar les dades en els anys pels quals manca informació, es prendran com a valors de la intensitat i el nivell migratoris en un determinat any Y la mitjana aritmètica dels 4 anys més propers (en particular els 4 anteriors quan es vulgui estimar els valors per a l'últim any de la sèrie històrica).

2. L'ESTIMACIÓ DEL NIVELL MIGRATORI

Donada l'existència d'una sèrie històrica amb dades sobre les migracions i per tant sobre l'indicador GMR, es podria pensar en estimar el valor del GMR en els anys pels quals no es troba disponible a partir de les tendències observades (bé per regressió lineal, sèries temporals o altres mètodes). En aquest sentit, cal fer una observació important: en treballar amb àrees que generalment tenen una població inferior als 45.000 habitants (23 de les 62 àrees), el valor del GMR es troba distorsionat per l'aleatorietat de les xifres petites dels numeradors de les taxes, de manera que l'objectiu principal és calcular uns GMR robustos.

L'objectiu d'aquesta etapa és, doncs, agrupar les àrees de població de manera que tots els grups resultants superin la barrera dels 45.000 habitants. El primer criteri en què es pot pensar és el geogràfic (agrupació de comarques veïnes), però aquest criteri no és vàlid perquè, malgrat el veïnatge, hi ha comarques en expansió junt amb altres en retrocés, fet que es reflecteix clarament en la intensitat de la migració. El criteri emprat és realitzar una anàlisi cluster entre les àrees, agrupant cada àrea de població inferior a 45.000 habitants amb l'àrea que presenti distància mínima dins l'anàlisi cluster. Si el conjunt no arriba al mínim poblacional s'agrupa de nou. Si l'agrupació és d'una àrea petita (menor de 45.000 habitants) amb una de gran, el GMR del cluster només s'aplica a la petita, mentre que el càlcul del GMR de la gran no es modifica. Si l'agrupació és entre 2 o més àrees petites, el GMR del cluster s'aplica a cada una d'elles. Per obtenir un GMR més robust, aquesta anàlisi es realitza sobre les

migracions dels 4 anys més propers a l'any a estimar i sobre el conjunt dels 2 sexes, és a dir sobre $GMR_Y(a)$.

En l'anàlisi cluster s'ha emprat un mètode jeràrquic acumulatiu: partint de 62 àrees, en cada pas s'agrupen les 2 àrees o clusters amb distància mínima. El procés s'itera fins a reduir el nombre de clusters a 2.

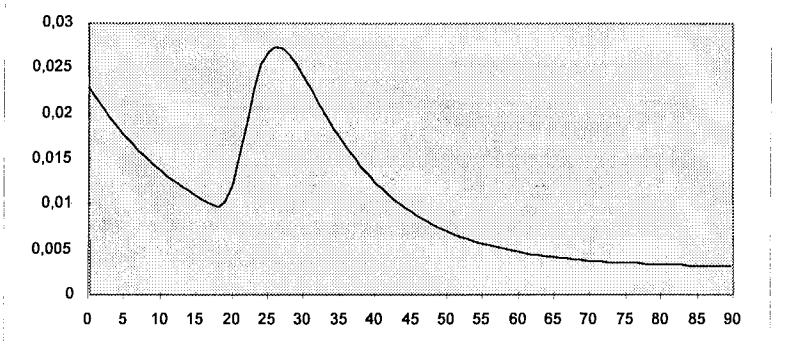
3. L'ESTIMACIÓ DEL PERFIL MIGRATORI

Les dades empíriques mostren que la migració és un fenomen altament selectiu segons l'edat: els joves entre els 20 i els 30 anys exhibeixen les taxes més altes de migració, bàsicament relacionades amb motius laborals i de matrimoni, mentre que els adolescents tenen les taxes de migració més petites. D'altra banda, les taxes migratòries dels fills reflecteixen les dels pares i això explica que els nens tinguin taxes més elevades que els adolescents. S'ha constatat que en algunes àrees pot trobar-se un augment de les taxes a la tercera edat, relacionat amb canvis de residència lligats a la jubilació i, en edats encara més avançades, es poden constatar migracions lligades a canvis en l'estat de salut (ingrés en residències, trasllat a cases dels fills, etc.).

El model matemàtic que descriu la migració a partir de l'edat es coneix com model de Rogers. La seva formulació, prenent l'edat com a variable dependent, és:

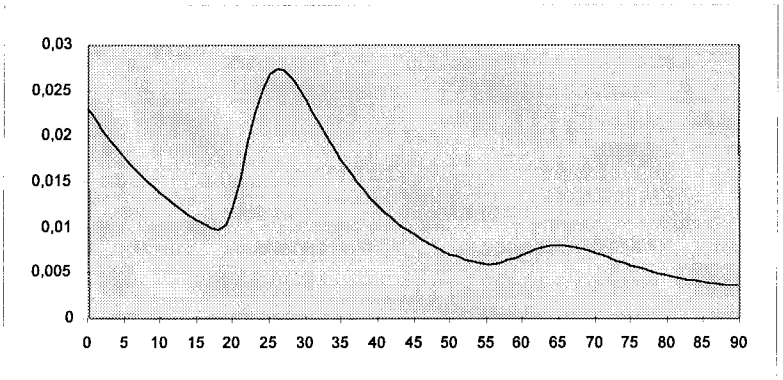
model general:

$$p(x) = a_1 \cdot \exp(-\alpha_1 \cdot x) + a_2 \cdot \exp\{-\alpha_2 \cdot (x-\mu_2) - \exp(-\lambda_2 \cdot (x-\mu_2))\} + c$$



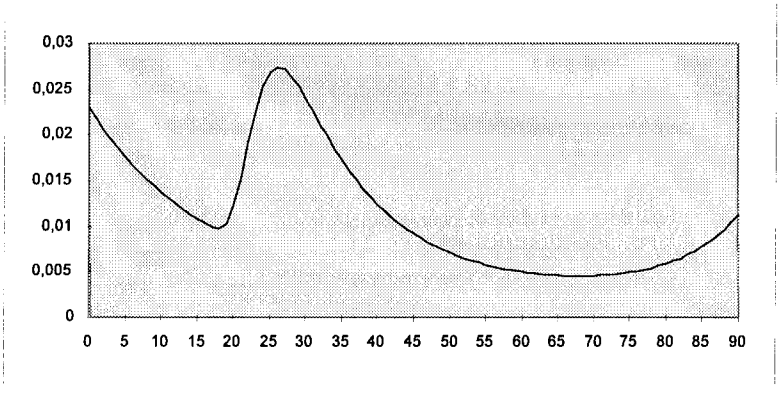
model amb pic de jubilació:

$$p(x) = a_1 \cdot \exp(-\alpha_1 \cdot x) + a_2 \cdot \exp\{-\alpha_2 \cdot (x-\mu_2) - \exp(-\lambda_2 \cdot (x-\mu_2))\} +$$
$$+a_3 \cdot \exp\{-\alpha_3 \cdot (x-\mu_3) - \exp(-\lambda_3 \cdot (x-\mu_3))\} + c$$



model amb pendent de jubilació:

$$p(x) = a_1 \cdot \exp(-\alpha_1 \cdot x) + a_2 \cdot \exp\{-\alpha_2 \cdot (x-\mu_2) - \exp(-\lambda_2 \cdot (x-\mu_2))\} + a_3 \cdot \exp(\alpha_3 \cdot x) + c$$



Per obtenir el perfil migratori de cada àrea i sexe es realitza una anàlisi cluster de les corbes de pesos pel conjunt dels 2 sexes. En aquest sentit, és important treballar amb les corbes normalitzades $p_Y(a)$ en lloc de $t_Y(a)$, per determinar l'efecte de la migració a cada edat, independentment de la intensitat de la migració. Es defineix la distància entre perfils com

$$d(p_Y(a),\, p_Y(a')) = \left\| p_Y(a) - p_Y(a') \right\|_2$$

Un pas previ a l'anàlisi cluster és disposar d'àrees amb prou població com per eliminar l'efecte pertorbador de les poblacions en els numeradors de $t_Y(a, s, e)$. Amb aquest efecte s'agrupen les àrees que no superen els 45.000 habitants, atenent a criteris de veïnatge geogràfic (es considera que el comportament migratori ve influït per factors com l'estructura de la població, el medi geogràfic o el tipus de vida). El nombre d'àrees queda reduït de 62 a 45. A continuació, es realitza l'anàlisi cluster, definint la distància entre 2 clusters com $\|p_Y(k_1) - p_Y(k_2)\|_2$. El resultat d'aquesta etapa és l'obtenció de 6 clusters per a l'emigració i 6 per a la immigració (veure taula 1).

Un cop obtinguts els clusters, per a cada un d'ells se sumen els efectius migratoris i poblacionals de totes les àrees que el componen, obtenint el perfil migratori del cluster per sexe $p_Y(k, s)$ (on k indica el cluster). Aquests vectors tenen com a components valors observats. L'etapa final és procedir a la substitució dels punts $p_Y(k, s)$ per la corba paramètrica $\hat{p}_Y(k, s)$. Aquest pas es realitza amb el paquet estadístic SAS, per aproximació de funcions no lineals; s'empra el mètode de Marquard, que és un compromís entre el de Gauss-Newton i el de màxima pendent.

Els gràfics permeten apreciar la qualitat de l'ajustament de la corba teòrica a les dades observades a diferents comarques i ciutats, amb un efecte de suavització de les corbes. És evident que hi ha un efecte de grandària i que com més població tenen les àrees millor s'ajusten al model. Les gràfiques de la distribució dels errors permeten observar que hi ha diferències en la qualitat de l'ajustament de les dades segons el sexe i l'edat. D'una banda, el model s'ajusta millor al sexe masculí (valors mínims de $\|p_Y(k) - \hat{p}_Y(k)\|_2$) que al femení. Pel que fa a les edats, els valors outliers es concentren als vells i als joves pels homes, mentre que pel sexe femení es troba una major dispersió, apareixent una certa sobrevaloració de la migració en edats laborals i prelaborals.

La introducció de corbes amb pendent de jubilació ha permès millorar la qualitat de l'ajustament en algunes corbes d'immigració, encara que el procediment ha suposat un augment notable del nombre d'iteracions per assolir els criteris de convergència. S'ha introduït també un pic de jubilació en algunes corbes d'emigració amb el mateix resultat positiu (veure gràfics 1 a 4).

De la parametrització dels perfils migratoris es deriva la definició de valors característics que faciliten el seu estudi i comparació. Entre ells destaquem els següents:

$$
\begin{aligned}
x_m &= \Sigma_x \cdot p(x)/\Sigma_p(x) \text{ edat mitjana de la migració} \\
\delta_{1c} &= a_1/c \\
\delta_{12} &= a_1/a_2 \text{ índex de dependència infantil} \\
\beta_{12} &= \alpha_1/\alpha_2 \text{ índex de regularitat parento-infantil} \\
\sigma_2 &= \lambda_2/\alpha_2 \text{ asimetria laboral}
\end{aligned}
$$

$$\delta_{32} = a_3/a_2 \text{ índex de dependència postlaboral}$$
$$\sigma_3 = \lambda_3/\alpha_3 \text{ asimetria postlaboral}$$

El pes de les edats actives i preactives queda reflectit en els paràmetres a_1 i a_2 respectivament. El ratio δ_{21} reflecteix el grau de dominança laboral, mentre que el seu invers δ_{12} mesura el grau de dependència infantil, o taxa amb la qual els adults migren amb fills. L'asimetria laboral, la forma esbiaixada cap a l'esquerra de la migració a les edats laborals, es mesura per σ_2.

Una classificació de les corbes pot obtenir-se segons el caràcter familiar o laboral de la migració (veure taules 2 a 5 i quadres 1 i 2). Les primeres són les corbes migratòries amb dependència infantil, és a dir, amb un gran pes de famílies amb nens. Les corbes de caràcter predominantment laboral són aquelles en les quals es troba un pes relatiu molt important en la migració de la població en edat de treballar, en comparació a la migració de la resta de població. Una mesura del caràcter laboral o familiar de la migració la dóna δ_{12}. El valor d'aquest índex, en els models empírics observats, se situa entorn a 1/3 i es considera que valors superiors a 0,4 indiquen un valor predominantment familiar de la migració. Inversament, valors inferiors a 0,4 indiquen un caràcter predominantment laboral de la migració. De l'estudi dels patrons migratoris es dedueix que les migracions a la Catalunya actual semblen haver substituït el motiu laboral pel residencial, això lligat probablement a les noves facilitats pels transports que permeten un increment de la mobilitat diària per motiu de treball i estudi. L'estudi d'altres paràmetres porta a conclusions com ara l'existència d'un pic de jubilació en un grup de ciutats, que va rebre fortes onades d'immigració la segona meitat del segle XX i que ara experimenta un fenomen de retorn de migrants cap a altres comunitats autònomes. D'altra banda, el pendent de jubilació apareix a determinats grups de caràcter predominantment urbà, explicable per la migració de la gent gran lligada a canvis en l'estat de salut i dirigida cap a centres urbans que disposen d'equipaments i/o cap a la llar dels fills de persones que ja no poden viure soles per l'estat de salut.

4. L'ESTIMACIÓ DELS EFECTIUS MIGRATORIS

Un cop s'han obtingut els GMR estimats i els patrons migratoris, s'estima el nombre de migrants de la següent manera:

Sigui 'a' una àrea que pertany al cluster 'k_1' de GMR i al cluster 'k_2' de patró migratori. Llavors

$$M_Y(a, s, e) = P_{Y-1}(a, s, e-1) \cdot \widehat{GMR}_Y(k_1) \cdot \hat{p}_Y(k_2, s, e) \quad e = 0, 1, \ldots, 99$$

La raó d'aplicar aquesta fórmula és que si a la part dreta se substituís $P_{Y-1}(a, s, e-1)$ per $P_Y(a, s, e)$, $\widehat{GMR}_Y(k_1)$ per $GMR_Y(a)$ i $\hat{p}_Y(k_2, s, e)$ per $p_Y(a, s, e)$ s'obtindria

$$P_Y(a, s, e) \cdot GMR_Y(a) \cdot p_Y(a, s, e) = P_Y(a, s, e) \cdot GMR_Y(a) \cdot t_Y(a, s, e)/GMR_Y(a) = P_Y(a, s, e) \cdot t_Y(a, s, e) = P_Y(a, s, e) \cdot M_Y(a, s, e)/P_Y(a, s, e) = M_Y(a, s, e)$$

Com que els valors observats de $p_Y(a, s, e)$ estan subjectes a més irregularitats se substitueixen per $\hat{p}_Y(k_2, s, e)$, que són més estables. D'altra banda, no es coneix $GMR_Y(a)$ ni $P_Y(a, s, e)$ i se substitueixen per $\widehat{GMR}_Y(k_1)$ i $P_{Y-1}(a, s, e-1)$.

5. CONCLUSIONS I RESULTATS

En aquest article s'ha descrit el mètode emprat a l'Idescat per estimar dades sobre les migracions, tant d'entrada com de sortida, del territori de Catalunya dividit en 62 àrees: 21 municipis i 41 comarques. La disposició d'un registre complet i coherent de la migració es remunta a 1988, any de la creació de l'Estadística de Variacions Residencials (EVR). D'acord amb l'EVR, tota migració queda registrada tant en el municipi de sortida com en el d'arribada. Abans això no era així, de manera que el nombre d'emigrants i el d'immigrants no era coherent. Un altre tema que es planteja és l'exhaustivitat: mentre que esdeveniments demogràfics com són els naixements, les defuncions o els matrimonis queden registrats en la seva totalitat, no passa el mateix amb les migracions. Sovint, la gent canvia la residència d'un municipi a un altre, però no registra el canvi a l'ajuntament, bé sigui per desconeixement, comoditat o interès personal. Aquest registre aflora, sovint, temps després, per exemple: quan cal inscriure els fills a l'escola o quan es realitza la renovació padronal.

L'estudi de la migració separant-ne els dos factors que la componen, el perfil i el nivell, permet fer una estimació acurada de la migració i, per tant, solucionar els problemes de l'inexhaustivitat: d'una banda s'aplica als anys 1986 i 1987 per poder perllongar 2 anys més la sèrie de dades pel seu inici i, de l'altra, per poder disposar de dades més completes als anys de renovació padronal, 1991 i 1996 fins al moment, en què es produeix un subregistre de la migració per quedar absorbit el seu registre al padró.

L'altre punt important del mètode és que dóna una eina per caracteritzar la migració arreu de Catalunya. S'estudia de manera sistemàtica la migració a nivell comarcal, amb un gran nivell de desagregació d'edat, fet que no s'havia realitzat a Catalunya fins ara. La migració és un fet que varia notablement d'una comarca a una altra, no només per la seva intensitat sinó també per la seva distribució per edats. Per primera vegada es fa una anàlisi que separi aquests factors, determinant 6 patrons d'emigració i 6 d'immigració. Cal fer notar que no s'ha treballat amb saldos migratoris, sinó amb emigració i immigració, de manera que una comarca o municipi queda caracteritzat per

2 patrons, un per a les sortides i un altre per a les entrades. Els patrons són diferenciats per sexe i permeten quantificar els components de la migració: el familiar, el laboral, el de jubilació i el d'edats avançades. Donada la influència dels municipis grans en el seu entorn, aquests s'han separat de les comarques, ja que sovint tenen comportaments diferents i fins i tot oposats, que no quedarien ben registrats si s'estudiés la comarca com un tot.

Pel que fa al nivell de migració es poden distingir 4 menes d'àrees: primer les molt dinàmiques o d'alta mobilitat, que presenten valors de GMR elevats, tant per la immigració com per l'emigració; segon, les de poca mobilitat, en ambdós sentits; tercer, les de caràcter marcadament emigratori, és a dir, que presenten un GMR elevat d'emigració combinat amb un GMR mitjà o baix d'immigració, i, finalment, les de caràcter marcadament immigratori.

Destaca el poc dinamisme migratori d'algunes de les ciutats de més de 45.000 habitants, mentre que les posicions amb els nivells més elevats són ocupades per les comarques i restes comarcals. Ciutats com Terrassa, Sabadell, Mataró, Manresa i Lleida destaquen pel seu nivell particularment baix tant d'entrades com de sortides. El caràcter marcadament emigratori de ciutats com Cornellà de Llobregat, Santa Coloma de Gramenet, L'Hospitalet de Llobregat, Barcelona i Badalona contrasta amb la situació de Rubí, Viladecans, Girona, Cerdanyola del Vallès i Reus, de caràcter marcadament immigratori. Entre les comarques amb caràcter marcadament immigratori destaquen el Garraf, el Baix Penedès, el Tarragonès, el Maresme, el Vallès Occidental i el Vallès Oriental.

En conclusió, l'estudi del GMR permet, en comparació amb estudis ja existents, apreciar el dinamisme migratori de les comarques pirinenques que, independentment del saldo migratori, presenten una mobilitat elevada. A l'extrem oposat, destaca la baixa mobilitat a les àrees de l'interior i del sud-oest de Catalunya, amb signes migratoris generalment negatius. Finalment, destaquen per la seva elevada immigració les restes comarcals de les comarques properes a Barcelona i tota l'àrea de les comarques gironines, així com l'entorn de l'eix Reus-Tarragona.

L'estudi dels patrons migratoris ens permet caracteritzar les àrees des d'un altre punt de vista. El grup de ciutats constituït per Cerdanyola del Vallès, Rubí, Sant Adrià de Besòs, Santa Coloma de Gramenet, l'Hospitalet de Llobregat i Viladecans atrauen treballadors i expulsen famílies: presenten un perfil immigratori de tipus predominantment laboral i, en canvi, una emigració amb un caràcter marcadament familiar. A la banda oposada trobem les ciutats de Reus i Tarragona, atractives per a la immigració familiar i amb una emigració predominantment laboral. Pel que fa als moviments de les comarques, destaca el caràcter altament familiar, tant de la immigració com de l'emigració, als municipis menors de 45.000 habitants del Vallès Occidental, Baix Llobregat, Garraf, Baix Camp i Tarragonès, i a la comarca del Baix Penedès. De la resta de comarques cal esmentar el Segrià i la Garrotxa, caracteritzades per una corba

amb un perfil marcadament laboral en la seva emigració, amb la particularitat d'una elevada sobre-emigració del sexe femení en les edats laborals, típica del medi rural a Catalunya.

Taula 1. *Composició dels clusters de perfil migratori, 1988-90, 1992.*

Immigració	*Emigració*
Cluster 1	**Cluster 1**
ALT CAMP	ALT CAMP
CONCA DE BARBERA	CONCA DE BARBERA
ALT PENEDES	BAIX EBRE
ANOIA	BERGUEDA
Resta BAGES	SOLSONES
GARRIGUES	CERDANYA
PRIORAT	RIPOLLES
RIBERA D'EBRE	NOGUERA
TERRA ALTA	PLA D'URGELL
BERGUEDA	SELVA
SOLSONES	SEGARRA
CERDANYA	URGELL
RIPOLLES	Resta BAIX CAMP
Terrassa	Resta TARRAGONES
Esplugues de Llobregat	Girona
OSONA	Manresa
Manresa	Granollers
SELVA	ALT EMPORDA
Mataró	BAIX EMPORDA
Granollers	ALT URGELL
BAIX EBRE	ALTA RIBAGORÇA
Barcelona	PALLARS JUSSÁ
Lleida	PALLARS SOBIRA
MONTSIA	VAL D'ARAN
NOGUERA	Resta GIRONES
PLA D'URGELL	PLA DE L'ESTANY
Sabadell	GARRIGUES
Cluster 2	PRIORAT
Resta BARCELONES	RIBERA D'EBRE
Santa Coloma de Gramenet	TERRA ALTA
Badalona	MONTSIA
Cornellà de Llobregat	**Cluster 2**
Prat de Llobregat	GARROTXA
Viladecans	Resta SEGRIA
Hospitalet de Llobregat	

Taula 1. (cont.) *Composició dels clusters de perfil migratori, 1988-90, 1992.*

Immigració	*Emigració*
Cluster 3	**Cluster 3**
ALT EMPORDA	ALT PENEDES
GARROTXA	OSONA
Resta GIRONES	ANOIA
PLA DE L'ESTANY	Resta BAGES
SEGARRA	Resta VALLES ORIENTAL
URGELL	Sabadell
ALT URGELL	Vilanova i la Geltrú
ALTA RIBAGORÇA	**Cluster 4**
PALLARS JUSSÁ	Resta MARESME
PALLARS SOBIRA	Lleida
VAL D'ARAN	Reus
Girona	Barcelona
BAIX EMPORDA	Badalona
Resta SEGRIA	Terrassa
Cluster 4	Prat de Llobregat
Sant Boi de Llobregat	Tarragona
Cerdanyola del Vallès	**Cluster 5**
Rubí	Resta BAIX LLOBREGAT
Cluster 5	Sant Boi de Llobregat
BAIX PENEDES	Mataró
Resta GARRAF	BAIX PENEDES
Cluster 6	Resta GARRAF
Resta BAIX CAMP	Resta VALLES OCCIDENTAL
Resta TARRAGONES	Rubí
Resta MARESME	Cerdanyola del Vallès
Resta BAIX LLOBREGAT	**Cluster 6**
Resta VALLES ORIENTAL	Resta BARCELONES
Tarragona	Santa Coloma de Gramenet
Reus	Hospitalet de Llobregat
Vilanova i la Geltrú	Cornellà de Llobregat
	Esplugues de Llobregat
	Viladecans

[ ] Agrupacions prèvies al cluster per assolir 45.000 habitants.

Taula 2. Paràmetres de la corba migratòria. Immigració.

	tipus	a1	$\alpha 1$	a2	$\alpha 2$	$\mu 2$	$\lambda 2$	a3	$\alpha 3$	$\mu 3$	$\lambda 3$	c
Catalunya H	0	0,0177	0,0837	0,0430	0,1196	23,37	0,2930	0,0000	0,0000	0,00	0,0000	0,004713
Catalunya D	0	0,0183	0,0847	0,0457	0,1269	22,56	0,2955	0,0000	0,0000	0,00	0,0000	0,004690

(*) Tipus model. Valors possibles: 0=normal, 1=amb pic de jubilació, 2= amb pendent de jubilació.
(**) a1 i $\alpha 1$ determinen la migració infantil; a2, $\alpha 2$, $\mu 2$ i $\lambda 2$ determinen la migració laboral; a3, $\alpha 3$, $\mu 3$ i $\lambda 3$ determinen la migració de jubilació; c és el terme independent de l'edat.
L'equació del model es troba a l'apartat 3 del text.

Taula 3. Valors característics de la corba migratòria. Immigració.

	tipus	xm	%0-14	%15-64	%65+	$\delta 1c$	$\delta 12$	$\delta 32$	$\beta 12$	$\sigma 2$	$\sigma 3$
Cluster1 H	2	37,89	22,95	62,61	18,72	4,46	0,3894	0,000444	0,6436	2,21	0
Cluster1 D	0	35,97	23,02	63,10	17,32	3,53	0,2949	0,000000	0,4967	1,75	0
Cluster2 H	2	51,66	21,09	58,18	35,84	5,18	0,2966	0,000004	0,6834	1,73	0
Cluster2 D	2	41,37	22,91	57,80	26,02	23502,67	0,3337	0,018828	0,5370	1,46	0
Cluster3 H	0	32,11	24,36	66,64	11,03	7,50	0,4842	0,000000	0,7943	3,44	0
Cluster3 D	0	32,30	25,22	64,44	12,84	6,05	0,4151	0,000000	0,6328	2,96	0
Cluster4 H	2	60,90	18,79	55,65	46,48	2,86	0,2461	0,000001	0,5792	1,60	0
Cluster4 D	2	46,07	19,43	58,54	30,52	3,36	0,2455	0,000670	0,5652	1,74	0
Cluster5 H	0	40,75	19,45	64,11	19,72	3,39	0,7484	0,000000	1,0545	11,29	0
Cluster5 D	0	40,73	19,10	64,10	20,37	2,56	0,6794	0,000000	1,0950	13,17	0
Cluster6 H	2	39,59	22,56	62,17	19,92	19014,51	0,6193	0,032861	0,6923	4,67	0
Cluster6 D	2	39,86	23,13	60,36	21,78	20157,32	0,5818	0,038010	0,6487	3,80	0
Catalunya H	0	35,63	22,57	60,80	16,63	3,76	0,4095	0,000000	0,6992	2,45	0
Catalunya D	0	35,02	22,97	60,58	16,45	3,91	0,4016	0,000000	0,6671	2,33	0

(*) Tipus model. Valors possibles: 0=normal, 1=amb pic de jubilació, 2=amb pendent de jubilació.
(**) La definició dels valors característics es troba a l'apartat 3 del text.

Taula 4. Paràmetres de la corba migratòria. Emigració.

	tipus	a1	$\alpha 1$	a2	$\alpha 2$	$\mu 2$	$\lambda 2$	a3	$\alpha 3$	$\mu 3$	$\lambda 3$	c
Catalunya H	0	0,0176	0,0777	0,0331	0,0956	23,27	0,3424	0,0000	0,0000	0,00	0,0000	0,004650
Catalunya D	0	0,0181	0,0780	0,0352	0,1031	22,38	0,3380	0,0000	0,0000	0,00	0,0000	0,004664

(*) Tipus model. Valors possibles: 0=normal, 1=amb pic de jubilació, 2= amb pendent de jubilació.
(**) a1 i α 1 determinen la migració infantil; a2, $\alpha 2$, $\mu 2$ i $\lambda 2$ determinen la migració laboral; a3, $\alpha 3$, $\mu 3$ i $\lambda 3$ determinen la migració de jubilació; c és el terme independent de l'edat.
L'equació del model es troba a l'apartat 3 del text.

Taula 5. Valors característics de la corba migratòria. Emigració

	tipus	xm	%0-14	%15-64	%65+	$\delta 1c$	$\delta 12$	$\delta 32$	$\beta 12$	$\sigma 2$	$\sigma 3$
Cluster1 H	0	32,32	25,33	62,44	12.23	6,61	0,5578	0,000000	0,7129	3,27	0,00
Cluster1 D	0	31,49	25,36	62,48	12,17	6,31	0,4905	0,000000	0,5806	3,70	0,00
Cluster2 H	0	31,31	25,72	62,28	12,00	7,59	0,4159	0,000000	0,7462	1,84	0,00
Cluster2 D	0	30,49	24,50	63,95	11,55	6,44	0,2748	0,000000	0,4124	2,40	0,00
Cluster3 H	1	32,24	25,42	62,95	11,63	7,87	0,4819	0,000042	0,7192	1,96	0,14
Cluster3 D	0	31,97	24,80	62,67	12,53	5,73	0,4041	0,000000	0,4216	3,30	0,00
Cluster4 H	0	35,50	21,88	61,84	16,27	3,36	0,3978	0,000000	0,5274	4,91	0,00
Cluster4 D	0	35,23	21,55	60,97	17,47	3,22	0,2662	0,000000	0,4943	2,15	0,00
Cluster5 H	0	36,56	23,01	59,90	17,10	3,81	0,6865	0,000000	0,9137	5,62	0,00
Cluster5 D	0	35.66	23,68	59,00	17,32	3,56	0,5972	0,000000	0,6832	4,47	0,00
Cluster6 H	1	38.18	22,75	56,31	20,93	52,13	0,7767	0,160768	1,0515	7,89	358740,81
Cluster6 D	1	37,36	22,97	56,14	20,89	22977,77	0,6599	0,161285	0,7610	5,08	175475,06

(*) Tipus model. Valors possibles: 0=normal, 1 =amb pic de jubilació, 2=amb pendent de jubilació.
(**) La definició dels valors característics es troba a l'apartat 3 del text.

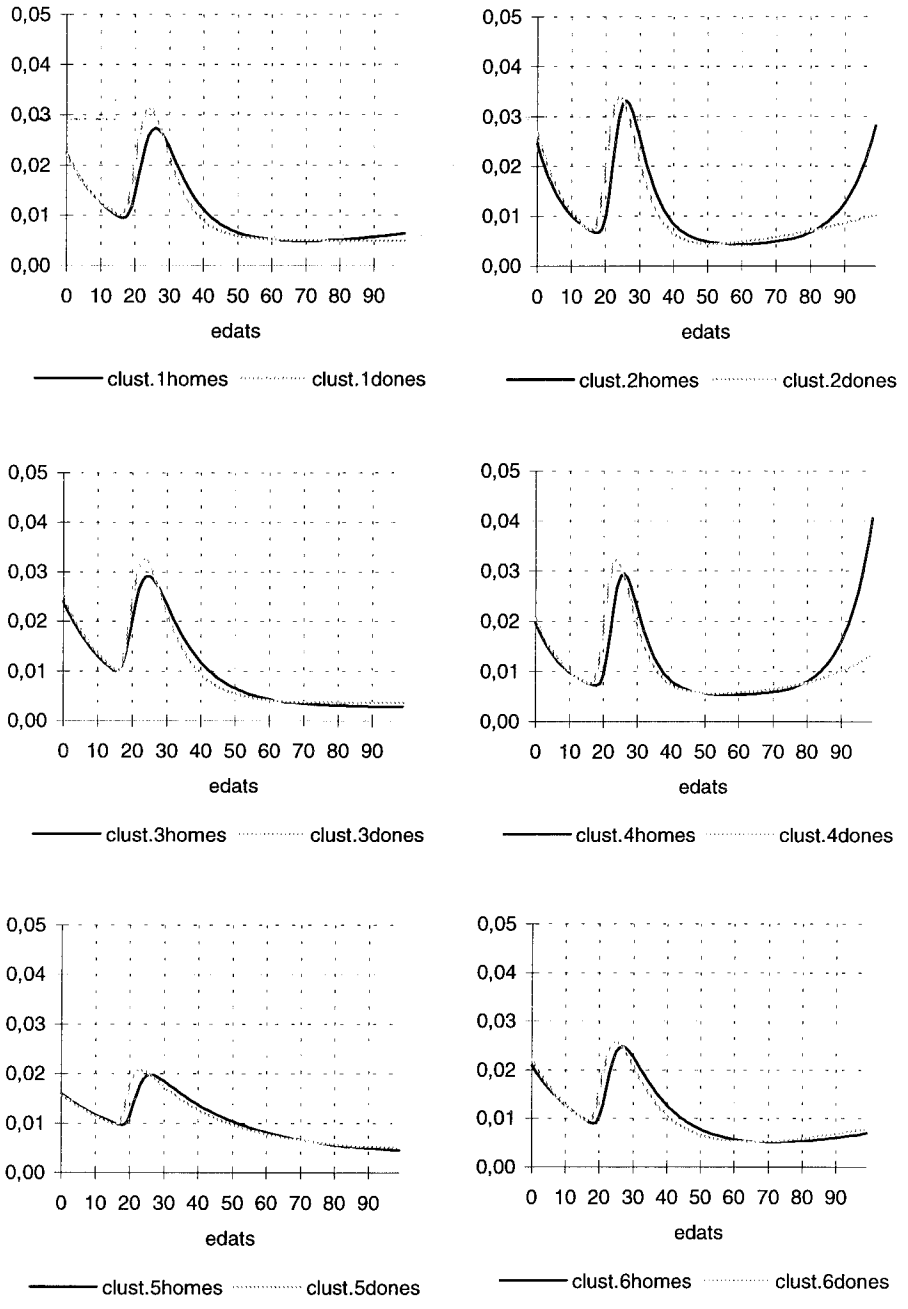
Quadre 1. Classificació de les comarques i restes comarcals segons el tipus d'emigració i d'immigració.

Immigració / Emigració	Laboral Alt	Laboral Mitjà	Familiar Mitjà	Familiar Alt
Laboral Alt		Segrià* Garrotxa		Maresme*
Laboral Mitjà	Anoia Alt Penedès Bages* Osona	Alt Camp, Conca Barberà Ribera d'Ebre Montsià, Baix Ebre Garrigues, Priorat, Terra Alta Noguera, Pla d'Urgell	Vallès Oriental*	
Familiar Mitjà		Berguedà, Solsonès Ripollès, Val d'Aran Alt Ribagorça, Pallars Jussà Alt Urgell, Pallars Sobirà Cerdanya	Baix Empordà Pla Estany Gironès* Alt Enpordà Segarra, Urgell	Tarragonès* Baix Camp*
Familiar Alt				Vallès Occidental* Baix Penedès Baix Llobregat* Garraf*

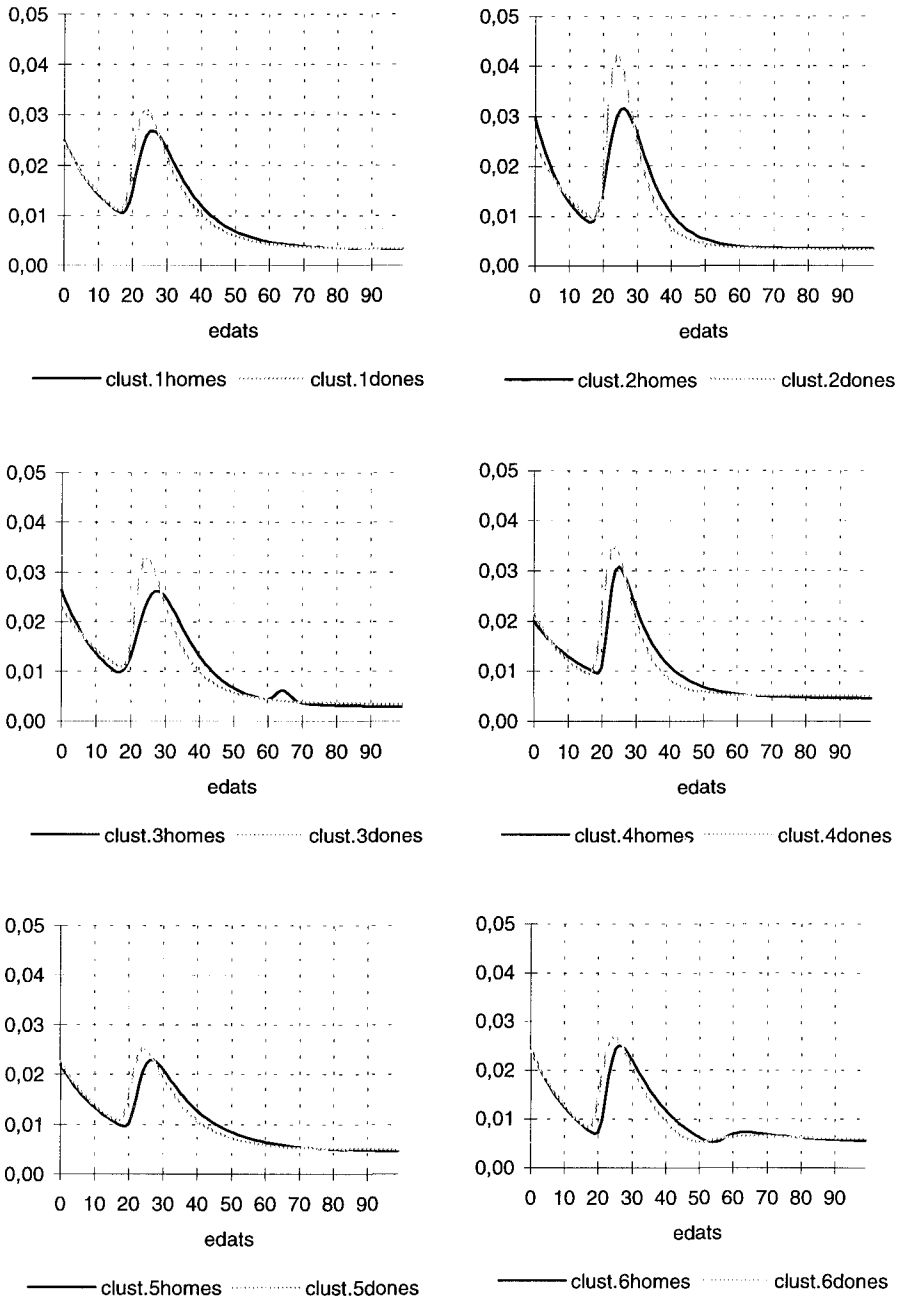
* Restes comarcals

Quadre 2. Classificació de les ciutats de 45.000 habitants i més segons el tipus d'emigració i d'immigració.

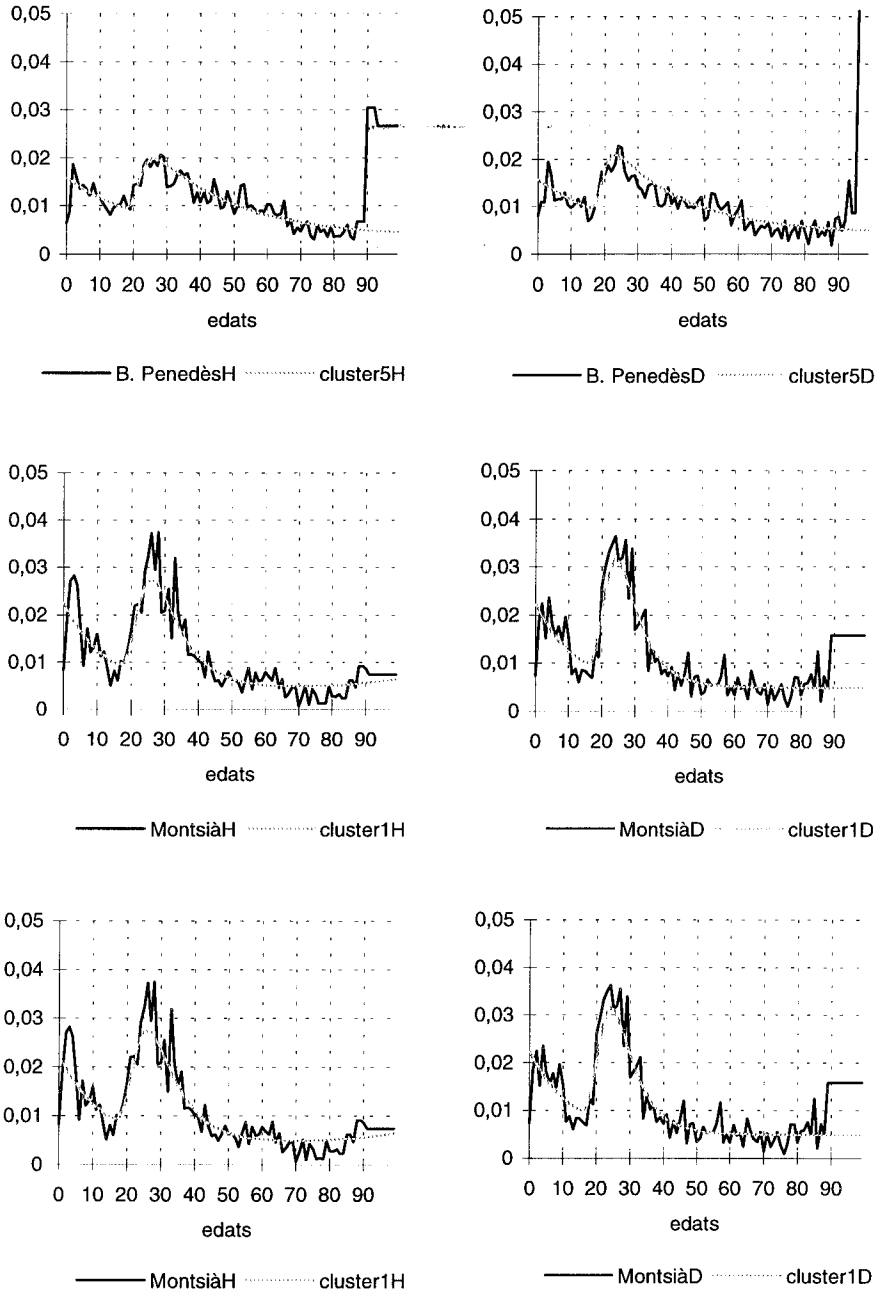
Immigració / Emigració	Laboral Alt	Laboral Mitjà	Familiar Mitjà	Familiar Alt
Laboral Alt	Badalona El Prat Ll.	Lleida Barcelona Terrassa		Tarragona Reus
Laboral Mitjà		Sabadell		Vilanova
Familiar Mitjà		Granollers Manresa	Girona	
Familiar Alt	St. Boi, Rubí Cerdanyola St. Adrià, Sta. Coloma Cornellà, L'Hospitalet Ll. Viladecans	Esplugues Ll. Mataró		



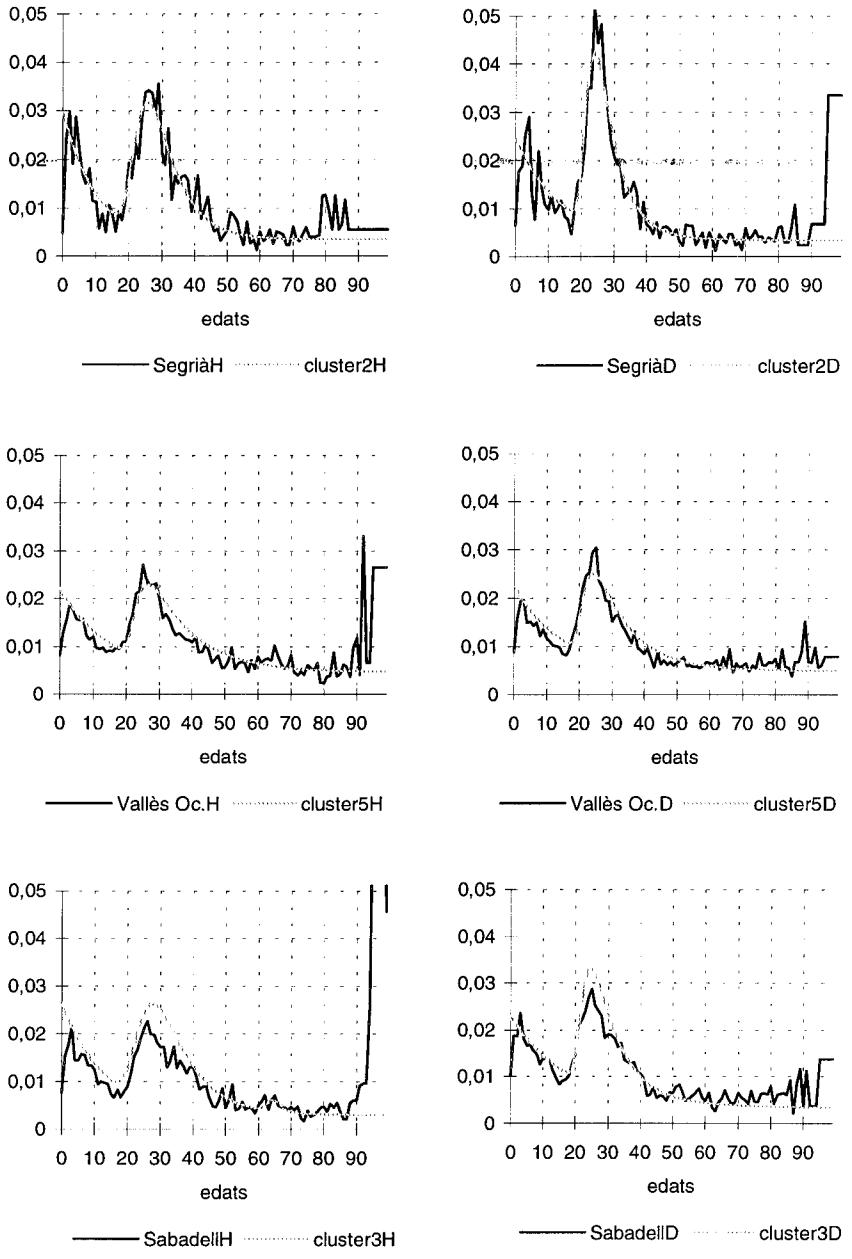
Gràfic 1. Corbes dels clusters* de perfil migratori. Immigració 1988-90, 1992.



Gràfic 2. Corbes dels clusters de perfil migratori. Emigració 1988-90, 1992.



Gràfic 3. Comparació per edad i sexe de les dades observades amb el model ajustat. Immigració 1988-90, 1992. Baix Penedès, Montsià i L'Hospitalet.



Gràfic 4. Comparació per edad i sexe de les dades observades amb el model ajustat. Emigració 1988-90, 1992. Segrià*, Vallès Occidental* i Sabadell.

REFERÈNCIES

[1] **Aldenferden** and **Blashfield** (1990). «Cluster analysis». *Quantitative applications in the social sciences*, **44**. Sage publications.

[2] **Boden, P., Stillwell, J.** and **Rees, P.** (1991). «Internal migration projection in England: the OPCS/DOE model examined» in Stillwell,J. and Congdom, P. *Migration models*. Belhaven Press.

[3] **Castro, L.** and **Rogers, A.** (1986). «Migration» in Rogers, A. and Willekens, F.J. *Migration and settlement. A multiregional comparative study*. IIASA.

[4] **Dûchene, J.** (1989). *Téchniques auxiliares en démographie II: Analyse des donées en démographie*. Institut de Démographie, Université Catholique de Louvain.

[5] **Institut d'Estadística de Catalunya**. *Estimacions de població 1986-1995*. Generalitat de Catalunya.

[6] **Lewis-Beck, M.S.** (1990). «Applied regression». *Quantitative applications in the social sciences*, **22**. Sage publications.

[7] **Longfod, N.T.** (1996). «Small-area estimation using adjustment by covariates». *Qüestiió*, **20, 2.**

[8] **SAS/STAT.** *User's guide*, version 6 (1990). SAS Institute.

ENGLISH SUMMARY

METHOD FOR ESTIMATING SMALL AREA MIGRATION FROM INCOMPLETE DATA

J.A. SÁNCHEZ CEPEDA
Institut d'Estadística de Catalunya*

The knowledge of small area migration data is a part of the project of Population estimates, which calculates recent Catalan population for towns above 45.000 inhabitants and small areas, broken down by age and sex. The difficulty, for some years, of having figures with such level of detail has led to the implementation of a method for its estimation.

This article describes the method applied by the Institut d'Estadística de Catalunya (Idescat) for estimating migration broken down by geographic area, sex and age. The method calculates separately the level and the profile of migration. Working with areas below 45.000 inhabitants is the main problem that arises in estimating the level of migration. The number of profiles is reduced from the initial 62 small areas into 6, which define the migration patterns observed in Catalonia. The typology of this migration patterns is commented on the base of parameters determined by the model. Finally, the number of immigrants and emigrants is calculated from the estimated migration levels and profiles.

Keywords: Small area, migration model, cluster analysis, non-linear adjustment.

* Institut d'Estadística de Catalunya. Via Laietana, 58. 08003 Barcelona.
- Received November 1997.
- Accepted April 1998.

The knowledge of the yearly evolution of the population is necessary as it is a basic tool for resources allocation and it is essencial for calculating demographic and social-economic indicators. L'Institut d'Estadística de Catalunya (Idescat) estimates recent catalan population using the components method. Population at the end of a year Y is calculated adding biths and inmigration to population at the end of the previous year, and substracting deaths and outmigration registered during year Y. Population is broken down by sex, 100 age groups and 62 geographic areas, 23 of which have less than 45.000 inhabitants, 25 have between 45.000 and 100.000 inhabitants and the rest have more than 100.000 inhabitants. For years 1986 and 1987, data for outmigration and inmigration are not available and they were estimated from data belonging to other years.

Let 'a' be a geographic area, 's' a sex, 'e' an age, 'M' the number of migrations, 'P' the population and 'Y' a year. Then,

$t_Y(a, s, e) = M(a, s, e)/P_Y(a, s, e)$ migration rate by area, sex and age in a year Y

$t_Y(a, e) = M(a, e)/P_Y(a, e)$ migration rate by area and age in a year Y

$\mathrm{GMR}_Y(a, s) = \Sigma_{e=0,1,\ldots,99}\, tY_(a, s, e)$ gross migration rate by area and sex

$\mathrm{GMR}_Y(a) = \Sigma_{e=0,1,\ldots,99}\, t_Y(a, e)$ gross migration rate by area

$p_Y(a, s, e) = t_Y(a, s, e)/\mathrm{GMR}_Y(a, s)$ weight of an age in the set of rates

$(p_Y(a, s))_{e=0,1,\ldots,99}$ migration profile

Notice that $\|(p_Y(a, s))_{e=0,1,\ldots,99}\|_1 = 1$.

The two elements of migration are the level of migration (the gross migration rate) and the profile. They are separately estimated.

The main problem with the level of immigration is its definition: as it is calculated as the sum of rates by age, when populations are under 45.000 inhabitants the value of GMR is statistically poor. In order to calculate better values for GMR, a cluster analysis is carried out. Areas are joined in order to have more than 45.000 inhabitants and $\mathrm{GMR}_Y(a)$ is calculated for them.

Migration rates behave under Rogers'age dependant model,

$$p(x) = a_1 \cdot \exp(-\alpha_1 \cdot x) + a_2 \cdot \exp\{-\alpha_2 \cdot (x - \mu_2) - \exp(-\lambda_2 \cdot (x - \mu_2))\} + c$$

This is known as the general model. The first term is the pre-labor force component, the second one the labour component and the third one a non-age dependant component. In some populations, an additional component for older ages, or retirement

peak, is observed,

$$p(x) = a_1 \cdot \exp(-\alpha_1 \cdot x) + a_2 \cdot \exp\{-\alpha_2 \cdot (x-\mu_2) - \exp(-\lambda_2 \cdot (x-\mu_2))\} + \\ + a_3 \cdot \exp\{-\alpha_3 \cdot (x-\mu_3) - \exp(-\lambda_3 \cdot (x-\mu_3))\} + c$$

whereas in other populations a post-retirement slope component is observed,

$$p(x) = a_1 \cdot \exp(-\alpha_1 \cdot x) + a_2 \cdot \exp\{-\alpha_2 \cdot (x-\mu_2) - \exp(-\lambda_2 \cdot (x-\mu_2))\} + a_3 \cdot \exp(\alpha_3 \cdot x) + c$$

A cluster analysis is made for the set of arrays $p_Y(a)$, using the distance between clusters $d(p_Y(a), p_Y(a')) = \|p_Y(a) - p_Y(a')\|_2$. It is important to analize $p_Y(a)$ instead of $t_Y(a)$, as $\|(p_Y(a, s))_{e=0,1,\ldots,99}\|_1 = 1$. The number of profiles is reduced from 62 to 6; these 6 profiles define the migration patterns observed in Catalonia. For every cluster k, $p_Y(k,s)$ is computed from the areas belonging to k. This profile is substituted for a smooth one, $\hat{p}_Y(k,s)$, using nonlinear functions aproximation.

Finally, once GMR and profile patterns have been obtained, the number of migrants is obtained as follows:

Let 'a' be an area belonging to the GMR cluster 'k,'and the profile cluster 'K'; then $M_Y(a,s,e) = P_{Y-1}(a,s,e-1) \cdot \widehat{GMR}_Y(k_1) \cdot \hat{p}_Y(k_2,s,e)$ $e = 0,1,\ldots,99$.

The reason to use this formula is that substituting $P_{Y-1}(a,s,e-1)$ for $P_Y(a,s,e)$, $\widehat{GMR}_Y(k_1)$ for $GMR_Y(a)$ and $\hat{p}_Y(k_2,s,e)$ for $p_Y(a,s,e)$,

$P_Y(a, s, e) \cdot GMR_Y(a) \cdot p_Y(a, s, e) = P_Y(a, s, e) \cdot GMR_Y(a) \cdot t_Y(a, s, e)/GMR_Y(a) =$
$P_Y(a, s, e) \cdot t_Y(a, s, e) = P_Y(a, s, e) \cdot M_Y(a, s, e)/P_Y(a, s, e) = M_Y(a, s, e)$

Notice that observed values for $p_Y(a,s,e)$ are much more irregular that the ones for $\hat{p}_Y(k_2,s,e)$; besides, $\mathrm{GMRY}_Y(a)$ and $p_Y(a,s,e)$ are unknown and they are substituted for $\widehat{GMRI}_Y(k_1)$ and $P_{Y-1}(a,s,e-1)$.

As the result of applying this method, a tool to determine migration in Catalonia has been obtained. Migration in small areas is studied by sex and age: 6 inmigration and 6 outmigration patterns are obtanied. These patterns allow to determine the level of migration in every area due to family, work related, retirement causes. To work with retirement peak and post-retirernent slope models has improved the results.

Biometria

QÜESTIIÓ, vol. 22, 2, p. 365-378, 1998

ANALYSIS OF SURVEY DATA INVESTIGATING THE MALARIAL ENDEMICITY OF A MIXED TRIBAL POPULATION OF BIHAR, INDIA

T.K. BASU
S. GANGULY
S.K. SARKAR
Indian Statistical Institute*
R. ARNAB
University of Durban**

A section of the mixed tribal population of the Singhbhum district, Bihar, India is declared malaria epidemic zone. The tribal population of several generation is known to be suffering from malaria. A survey based on cross-sectional data analysis, was conducted on the mixed tribal population for one month. The purpose of this study was to investigate the health status using collected blood samples. The main focus is on the comparative roles of the «defense mechanism» and «vitality factor» of the human system in context to the malarial infection. By gradual elimination of the blood parameters by statistical analyses, the «vitality» parameters probing malarial endemicity are assessed with a view to predicting the epidemic.

The main findings from this survey are (i) of the selected twenty two parameters of blood, albumin, total cholesterol, total protein, β-Globulin, γ-Globulin, Immuno globulin G seem to have some predictive capacity with respect to the malarial endemicity of the tribal people and (ii) Categorical variables like blood groups and sex are comparatively less important for prediction of malaria.

Keywords: Vitality factors; defense mechanism; two-way Anova; discriminant function.

* Biometry Research Unit. Indian Statistical Institute. 203 Barrackpore Trunk Road. Calcutta 700 035, India.
** Department of Statistics. University of Durban, South Africa.
- Received January 1997.
- Accepted November de 1997.

1. INTRODUCTION

The Indian Council of Medical Research, a leading medical research concern of India, identified a few villages of tribal habitat of the Singhbhum district of Bihar (India) as a malarial epidemic zone. The people of these villages show evidence of malarial suffering for several generations. A study was undertaken to investigate the malarial endemicity by checking the blood samples obtained from a mixed tribal population of the Singhbhum district. The investigation was executed by the Indian Council of Medical Research in conjunction with the Biometric Unit, Indian Statistical Institute.

Malaria depends upon two principal factors- (a) vitality and (b) defense mechanism, regulated by the circulation of blood. Vitality is controlled by total, free and ester cholesterol, albumin, red blood corpuscles (RBC) and haemoblobin, and the defense mechanism is controlled by immunoglobulins (IgG, IgA and IgM), white blood corpuscles (WBC), monocytes, nutrophils, eosinophils and lymphocytes. Lymphocytes are subdivided into T and B lymphocytes. In addition, the role of globulin-containing fractions is also prominent on this aspect. Under malarial conditions, RBC, haemoglobin and albumin decrease whereas WBC, neutrophils and globulins tend to increase. This is a very broad overview of this complex mechanism. In this context, a brief description of the biology of malaria is presented.

1.1. Biology of Malaria

Malaria is caused by a parasite protozoan infection. Four different species of the genus *Plasmodium* are known to infect humans. In the tropics, *Plasmodium falciparum* is most prevalent. The life cycle of the parasite is an interaction between a female mosquito of the genus *Anopheles* (the vector) and the human host. Transmission of disease occurs by a bite of the infectious mosquito. The parasite migrates to the liver, remains in the latent stage for several days while replicating. Ultimately, there occurs a penetration of host RBC with an asexual replication within the parasites which results in the lysis of the cells. Malarial symptoms are occurred by this asexual parasites in the blood. When there is a fresh case of mosquito biting, a sexual stage, called *gametocytes* develops (from the asexual parasites) which is responsible for transmission of parasites from the host. These transmitted parasites fertilize in mosquito gut, replicate and a new cycle of transmission starts. These types of epidemiological surveys are generally based on blood smears in which one can observe both the *asexual* parasites and *gametocytes*. Classifications of blood smears are generally done in terms of the appearance/absence of parasite or in terms of its density. According to above, there are actually three classes of population related to the malarial status- susceptibles, the proportion that are uninfected (Class 1); infecteds, the proportion with infection (Class 2); and immunes, the proportion with asymptotic infection which is not prominent

(Class 4). The Febrile Class i. e. Class 3 represented subjects with unidentified fever, not malaria (as confirmed by slide test).

1.2. Main Findings

Subsequent analysis of data gives some indication that the levels of albumin, total cholesterol, total protein, β - Globulin, γ -Globulin and IgG in the circulating system may be used for the prediction of malarial endemicity of the tribal population. The categorical variables like blood groups and sex are less important for the prediction of the malarial incidence.

2. SAMPLING DESIGN AND COLLECTION OF DATA

A simple random sample of 88 persons were selected from the list of volunteers from the tribal population. Among them, 45 were male and 40 were female. The selected individuals were invited at the Clinics of Kokda and Khanderber, organized by the Gram Vikas Kendra, a rural development centre of Tata Engineering and Locomotive Works, Jamshedpur, Bihar. The list of volunteers was prepared earlier by the Centre. 10 persons were invited daily in the Clinics serially from the chart. The average daily attendance was six. This is because most of the females did not respond owing to low literacy level, social-taboos or customs. The nature of nonresponse is quite random in nature and so this will not cause any bias in the experiment. However, the efficiency of the estimators will reduce because of the reduction of sample size. Relevant history regarding family members, total income of the family and other socio- economic data of the volunteers were collected by the Centre and were sent to the Clinics for record. They were mostly nonvegetarians residing in huts as usually observed in tribal villages. According to their family income, they belonged to the «below poverty-level» group.

2.1. Experimental protocol

Subjects, as instructed earlier, came in post-absorptive stage (overnight fasting) in the Clinics between 9 to 10 am. The attending physician checked the weight, pulse and blood pressure of each individual and drew blood from their bracial vein for subsequent testing. 10ml of blood was taken in an aseptic condition for the biochemical, haematological and immunological tests. Out of this 10ml blood, 4ml was kept in a sterilized glass vial with the sequestering agent. Blood smear was prepared on two glass slides, one thick (for the detection of Malarial parasite) and the other thin

(for differential count of the blood cells). The remaining portion of blood was taken in a sterilized test tube and was allowed to clot for release of serum.

The electrophoretic separation of different protein fractions was carried out in the line of Smithies(1955) with 1% agar and 0.1M Veronal Buffer (pH 8.6). The fractions of Protein- Albumin and Globulins were scanned by using a Densitometer. The relative percentage of different Protein fractions was measured by using a Planimetre.

The Immunodiffusion study was performed according to the standard method (Mancini et al, 1965). The diametre of the precipitating ring was measured in mm by using the 'Immunomeasure' scale. The standard curve was plotted on a double log graph paper using standard antigens supplied by M/s Immunodiagnostics Pvt. Ltd., Delhi. The Immunoglobulin fractions were calculated from the standard curve. The conventional methods for the estimation of Total protein (Lowry et al, 1951), Total, Free and Ester cholesterol(Wootton,1974), Haemoglobin, RBC,WBC,Blood groups and Differential counts i.e.estimation of Leucocytes, Monocytes, Eosinophiles, Basophils (Davie and Lewis,1984), T- and B- Lymphocytes (Blood, 1977) were followed (Summary tables 1A-1C). On the basis of the above pathological tests the population was post stratified into four Classes according to their malarial status.

Class 1. Normal: Those who had no fever currently and no past history of malaria within the last couple of years. Total number 20 (17 male and 3 female)

Class 2. Malarious: Those who turned up with symptoms of malaria and showed the malarial parasite positive in the 'slide test'. Total number 15 (11 male and 4 female)

Class 3. Febrile: Those who emerged with fever but no malarial parasite was identified in the 'slide test'. Total number 25 (15 male and 4 female).

Class 4. Dormant: Those who had recurring malaria during the last two years but showed no current malarial symptoms. Total number 28 (21 male and 7 female).

3. STATISTICAL ANALYSIS

A: Analyses of cardiovascular and blood parameters

The means of the following twenty one blood parameters (viz. Total Protein (TP), Total Chlesterol (Tch), Free Chlesterol (Fc), Ester Cholesterol (Ec), Red Blood Corpuscles (RBC), White Blood Corpuscles (WBC), Neutrophils (N), Lymphocytes (L), Eosinophils (E), Monocytes (M), Haemoglobin (Hb), Immuno- globulin G (IgG), Immunoglobulin A (IgA), Immunoglobulin M (IgM), Albumin (Alb), α_1 -Globulin (α_1), α_2 -Globulin (α_2), β -Globulin, γ -Globulin, T-cell (Tc), B-cell (Bc) and two cardio-

vascular parameters, namely, Pulse (P) and Blood Pressure, Systolic/Diastolic [Bp (Sys/Dia)] for the four classes of people (Normal, Malarious, Febrin and Dormant) were estimated through the sample means which are given in the tables 1A, 1B and 1C. The estimated standard errors (SR) of the estimates are also presented in these tables. The following table gives an idea of the general health condition of the tribal community under consideration.

Table 1

Parameters	Normal Range Indian Standard	mean	standard deviation	standard error(se)	Proportion below normal range	Proportion above normal range	Proportion outside normal range	sample range min-max
P(c/min)	65-85	94.898	15.834	1.688	0	70.11	70.11	68-124
BP/sys	100-140	110.932	13.118	1.398	17.04	0	17.04	78-150
BP/dia	70-90	76.33	10.783	1.149	20.45	4.54	24.94	54-100
WBC(m/cmm)	4000-10000	7185.8	931.853	99.336	0	0	0	5000-9500
α_1(% Tp)	52-68	3.555	1.351	0.144	26.13	23.86	50	0.9-8.7
α_2(%Tp)	6.1-10.1	3.838	1.697	0.181	89.7	0	89.7	1-9.9
β(% Tp)	8.5-14.5	6.619	2.495	0.266	84.1	0	84.1	2.4-14.5
γ(% Tp)	10-21	18.991	5.38	0.537	4.54	27.27	31.88	6-42.9
IgG(mg%)	700-1500	2453.69	598.515	63.802	0	85.22	85.22	1205-3477
IgA(mg%)	90-450	193.818	18.698	1.993	0	0	0	152-237
IgM(mg%)	40-250	128.352	28.642	3.053	0	0	0	69-192
Alb(% Tp)	52-68	67.436	7.046	0.751	2.27	51.13	53.46	45.7-85.1
N(%)	60-65	56.125	5.858	0.624	65.9	6.81	72.72	42-68
L(%)	20-40	33.318	4.797	0.511	0	7.95	7.95	24-48
E(%)	1-3	8.148	4.268	0.455	0	87.45	87.45	1-19
M(%)	2-8	2.614	1.309	0.14	13.63	0	13.63	1-6
Hb(gm%)	14-16	11.063	1.231	0.131	100	0	100	8-12.8
Tc(%)	70-75	63.5	6.447	0.687	80.68	0	80.68	42-75
Bc(%)	15-20	24.08	5.126	0.546	1.13	70.45	71.59	12-38
Tp(gm%)	6-8	7.142	0.387	0.041	0	0	0	6.5-8
Tch(mg%)	150-280	126.773	22.951	2.447	82.95	0	82.95	80-184
Fc(mg%)	50-70	42.045	13.597	1.449	78.4	2.27	80.68	20-94
Ec(mg%)	95-210	85.886	22.21	2.368	69.31	0	69.31	26-140
RBC(m/cmm)	4000-10000	3.794	0.474	0.051	98.86	0	98.86	2.31-4.78

From the above table, we briefly comment on the average health condition as follows:

BP: Due to the simplicity of living and diet, the community under study maintained acceptably good BP levels.

RBC, Hb: Lower values indicate that most of the people are anaemic.

Tp, Tch, Fc, Ec: Lower values than the normal range is indication of low fat diets coupled with normal protein levels.

E: High concentration of eosinophils is generally correlated with parasitic infection in the people.

WBC, N, L, M: The levels of these parameters are acceptable in terms of facilitating immnuno-defence mechanism.

Summary Table: 1A

Class	Sample size	Statistic	P (count/min)	BP/sys (mm Hg)	BP/dia (mm Hg)	TP (gm%)	Tch (mg%)	Fc (mg%)	Ec (mg%)	RBC (m/cumm)	WBC (no/count)
		mean	91.6	112.6	77.7	7.08	136.2	39.2	102.5	4.062	7707.5
1	20	sd	16.28005	8.9241	8.97273	0.32031	21.7154	11.26765	14.42047	0.28971	816.13035
		se	3.64033	1.9955	2.00636	0.07162	4.85572	2.51952	3.22451	0.06478	182.19229
		mean	94.06667	108.67	7606	6.94666	96	38.63333	57.46667	3.72133	6736.66667
2	15	sd	15.48533	13.656	12.04326	0.23907	10.0133	14.50455	13.9421	0.48780	772.97405
		se	3.99829	3.526	3.10955	0.06172	2.58543	3.74506	3.59983	0.12595	199.58104
		mean	100.4	106.36	72.92	7.156	136.16	45.72	90.12	3.6264	7084
3	25	sd	15.09967	14.982	10.9724	0.43458	20.9908	13.92413	19.17982	0.43472	953.90984
		se	3.01993	2.9965	2.19448	0.08691	4.19816	2.74482	3.83596	0.08694	190.78197
		mean	92.78571	115.04	78.25	7.27857	128.143	42.67857	85.45429	3.79071	7144.64286
4	28	sd	15.065	11.975	10.35659	0.39402	14.8413	13.40589	17.97713	0.52002	898.54438
		se	2.84701	2.263	1.95721	0.07446	2.80475	2.53347	3.39735	0.09827	169.80892

Summary Table: 1B

Class	Sample size	Statistic	N (%)	L (%)	E (%)	M (%)	Hb (gm%)	Tc (%)	Bc (%)
		mean	55.45	32.85	9.65	2.05	11.82	65	23.1
1	20	sd	4.225	3.454	3.825	0.805	0.621	4.278	4.3
		se	0.945	0.772	0.855	0.18	0.139	0.957	0.962
		mean	53.667	36.4	5.667	4.267	10.647	63.733	23.6
2	15	sd	5.907	4.514	3.32	1.289	1.447	5.234	5.414
		se	1.525	1.165	0.857	0.333	0.374	1.351	1.398
		mean	588	31.56	7.64	2.64	10.704	62.16	25.32
3	25	sd	6.203	5.375	4.906	1.353	1.144	7.22	5.732
		se	1.241	1.075	0.981	0.271	0.229	1.444	1.146
		mean	55.536	33.571	8.857	2.107	11.064	63.5	23.929
4	28	sd	5.635	4.346	3.71	0.673	1.257	7.287	4.705
		se	1.065	0.821	0.701	0.127	0.238	1.377	0.889

Summary Table: 1C

Class	Sample size	Statistic	IgG (mg%)	IgA (mg%)	IgM (%)	Alb (%)	α_1 (%)	α_2 (%)	β (%)	γ (%)
		mean	1519.3	199.65	114.35	73.985	3.095	3.135	4.83	14.885
1	20	sd	199.158	19.132	28.079	5.513	1.28	1.042	1.36	4.62
		se	44.533	4.278	6.279	1.233	0.286	0.233	0.304	1.003
		mean	2870.333	182.6	129.067	58.74	3.68	4.353	8.267	24.867
2	15	sd	277.906	19.231	24.845	6.706	1.574	2.513	3.374	6.182
		se	71.755	4.965	6.415	1.732	0.407	0.649	0.871	1.596
		mean	2777.96	190	135.84	67.844	3.792	3.74	5.084	19.48
3	25	sd	410.31	15.773	33.308	3.86	1.406	1.484	1.606	3.417
		se	82.062	3.155	6.662	0.772	0.281	0.297	0.321	0.683
		mean	2608.393	199.07	131.286	67.054	3.604	4.136	6.971	18.339
4	28	sd	267.147	16.873	22.247	5.046	1.125	1.527	2.125	3.693
		se	50.486	3.189	4.204	0.954	0.213	0.289	0.402	0.698

B: Analysis of Variance

In this section, we will study whether or not the means of 21 blood parameters, BP (Systol/Diastol) and pulse rates are different for 4 classes of people described above. For this study, we consider the following two way analysis of a variance model using class and sex as two classifying factors for each of the 21 blood parameters, pulse rate and BP (Systol/Diastol):

$$(1) \qquad y_{i,j,k}(t) = m + a(i) + b(j) + c(i,j) + e_{i,j,k}$$

where $y_{i,j,k}(t)$ = response obtained from the k_{th} individuals for the j_{th} $(j = 0,1)$ sex of the i_{th} $(i = 1,2,3,4)$ class, corresponding to the parameter, t.

$$\begin{aligned} m &= \text{general effect} \\ a(i) &= \text{effect due to } i_{th} \text{ class } (i = 1,\ldots,4) \\ a(j) &= \text{effect due to } j_{th} \text{ sex } (j = 0,1) \\ a(i,j) &= \text{interaction effect between } i_{th} \text{ class and } j_{th} \text{ sex} \end{aligned}$$

Assumptions of the Analysis of variance model:

$$(a) \quad \sum_i a(i) = \sum_j b(j) = \sum_i c(i,j) = \sum_j c(i,j) = o$$

(b) $e(i,j,k)$ = independently and identically distributed as normal variate with mean zero and variance σ^2.

The assumptions of the model imply that $y_{i,j,k(l)}$'s are independently normally distributed with mean $m+a(i)+b(j)+c(i,j)$ and variance σ^2. Since the number of observations of the 4 classes (20, 15, 28) are different, unbalanced analysis of varince techniques are used. This is clearly explained with an example by Kshirsagar (1983). Following the analysis of variance tables, the effects of β, γ, IgG, Alb, Tp, Tch and Ech were found to be significant. This implies that the mean of each of the parameters β, γ, IgG, Alb, Tp,Tch and Ech is different for 4 classes. The computations for these parameters are shown in the tables (2A -2G).

Table 2A: ANOVA (β)

Source	ss	df	ms	F-value	P-value
Class	123.106	3	41.035	8.141	0
Sex	0.303	1	0.303	0.06	0.807
Interaction	1.754	3	0.585	0.116	0.951
Error	403.26	80	5.041		

Table 2B: ANOVA (γ)

Source	ss	df	ms	F-value	P-value
Class	627.839	3	209.28	10.509	0
Sex	39.157	1	39.157	1.966	0.165
Interaction	35.099	3	11.7	0.588	0.625
Error	1593.121	80	19.914		

Table 2C: ANOVA (IgG)

Source	ss	df	ms	F value	P value
Class	14318260	3	4772753.3	47.721	0
Sex	68669.954	1	68669.954	0.687	0.41
Interaction	65112.199	3	3360328.2	33.599	0.844
Error	8001077.2	80	100013.47		

Table 2D: ANOVA (Alb)

Source	ss	df	ms	F value	P value
Class	1236.765	3	412.255	14.108	0
Sex	0.011	1	0.011	0	0.984
Interaction	31.364	3	10.455	0.358	0.784
Error	2337	80	29.222		

Table 2E: ANOVA (TP)

Source	ss	df	ms	F-value	P-value
Class	2.212	3	0.737	3.892	0.012
Sex	0.033	1	0.033	0.172	0.679
Interaction	0.465	3	0.155	0.817	0.488
Error	15.159	80	0.189		

Table 2F: ANOVA (TCH)

Source	SS	df	ms	F-value	P-value
Class	13386.909	3	4462.303	13.466	0
Sex	1.986	1	1.986	0.006	0.938
Interaction	1477.278	3	492.426	1.486	0.225
Error	26509.82	80	331.373		

Table 2G: ANOVA (EC)

Source	ss	df	ms	F-value	P-value
Class	14398.88	3	4799.627	15.454	0
Sex	272.46	1	272.46	0.877	0.352
Interaction	249.905	3	83.302	0.268	0.848
Error	24846.723	80	310.584		

C. Correlation Matrix

The correlation matrix for 13 selected important parameters (viz. Alb, N,B,E, M, Hb, Tc, Tp, Tch, Ec), based on 88 observations obtained after combining 4 classes, are given in table 3. From table 3, we note that 'r' values of Ester Cholesterol (Ec) and Total Cholesterol (Tch), Bc and Alb are fairly high. But there is moderately high negative correlation between eosinophils (E) and neutrophils (N) and between neutrophils (N) and lymphocytes (L). Other correlations are quite low.

Table 3: Correlation Matrix

	Alb	N	L	E	M	Hb	Tc	Bc	Tp	Tch	Fc	Ec
Alb	1											
N	0.0244	1										
L	0.2996	−0.6653	1									
E	0.2996	−0.5307	−0.1506	1								
M	−0.452	0.06	0.1844	−0.2399	1							
Hb	0.2225	0.623	−0.1233	0.0466	−0.1339	1						
Tc	0.0945	0.571	−0.2196	0.0221	0.0337	0.1507	1					
Bc	0.0172	0.512	0.0443	−0.1387	0.0181	0.0709	0.0095	1				
Tp	0.1128	0.199	−0.005	0.2262	−0.1798	−0.0862	−0.053	−0.0983	1			
Tch	0.446	0.0666	−0.2467	0.2879	−0.4295	0.1672	0.087	−0.027	0.2972	1		
Fc	−0.125	0.0743	−0.0461	0.0439	−0.0533	−0.0817	−0.168	0.0101	0.0146	0.3247	1	
Ec	0.6435	0.1071	−0.2366	0.2906	−0.4317	0.2567	0.0279	−0.0307	0.2676	0.7868	−0.226	1

D. Categorical Data Analysis

In this section we test whetheror not, there is any association between (i) blood group and class and (ii) sex and class. Both the computed chi-squares for class vs sex were found to be insignificant (details are given in tables 4A and 4B). Thus neither the blood group nor sex has been found to be significantly associated with population class.

Table 4A: Contingency table (Marginal subtable)

Class	BG				
	o	a	b	ab	Total
Normal	4	7	1	8	20
Malarious	8	0	1	6	15
Febrile	8	7	3	7	25
Dormant	6	7	3	12	28
Total	26	21	8	33	88

Computed $\chi^2 = 10.6316 < 16.919 = \chi^2\ (0.05, 9)$, BG = Blood Group

Table 4B: Contingency table (Marginal subtable)

Class	Sex		
	males	females	Total
Normal	3	17	20
Malarious	4	11	15
Febrile	10	15	25
Dormant	7	21	28
Total	24	64	88

Computed $\chi^2 = 6.361 < 7.815 = \chi^2\ (0.05, 3)$

E. Predicting malaria infection

In this section, we try to find a method of classifying an individual in any of the 4 classes (Normal, Malarious, Febrile and Dormant) on the basis of the data collected on 23 blood parameters, Pulse rate, Blood pressure as well as sex and Blood group. From the categorical data analysis mentioned in section we note that Blood group and sex have no association with the class. In other words, sex and blood group cannot classify an individual in any of the four classes. But from the Analysis of variance technique given in section B, we note that Alb,Tp, Tch, Ech, β, γ, IgG are different for different classes. We already mentioned in section C that Tch and Ech are highly correlated. So we use Tch along with β, γ, IgG, Alb and Tp as classification variables. To classify an individual in any of the four classes, we compute L(x), discriminant function or Fisher's classification function for each of the four classes (Seber, 1984) for some given value of x= (β, γ, IgG, Alb, Tp, Tch). So for each of the four classes, the discriminant function was computed by using the STATISTICA (discriminant analysis programme) assuming prior probabilities are the same for all the four classes (details are given in tables 5A and 5B). The discriminant functions are as follows:

$$\text{Class 1: } L_1(x) = -904.862 + 22.633\beta + 15.701\gamma + 0.006\,\text{IgG} + 16.115\,\text{Alb} + \\ + 44.378\,\text{Tp} - 0.386\,\text{Tch}$$

$$\text{Class 2: } L_2(x) = -928.602 + 23.594\beta + 16.262\gamma + 0.022\,\text{IgG} + 16.099\,\text{Alb} + \\ + 41.704\,\text{Tp} - 0.450\,\text{Tch}$$

$$\text{Class 3: } L_3(x) = -905.102 + 22.649\beta + 15.714\gamma + 0.021\,\text{IgG} + 15.867\,\text{Alb} + \\ + 41.172\,\text{Tp} - 0.321\,\text{Tch}$$

$$\text{Class 4: } L_4(x) = -922.172 + 23.180\beta + 15.820\gamma + 0.019\,\text{IgG} + 16.012\,\text{Alb} + \\ + 43.170\,\text{Tp} - 0.364\,\text{Tch}$$

Table 5A

STAT DISCRIM. ANALYSIS		Classification Functions	Grouping	CL (malaria status)
Variable	G1:1	G2:2	G3:3	G4:4
β	2.633	23.594	22.649	23.180
γ	5.701	16.262	15.714	15.820
IgG	0.006	0.002	0.021	0.019
Alb	16.115	16.099	15.867	16.012
Tp	44.378	41.704	41.172	43.170
Tch	−0.386	−0.450	−0.321	−0.364
Constant	−904.862	−928.602	−905.102	−922.172

Table 5B

STAT. DISCRIM. ANALYSIS		Classification Matrix (malaria. sta) Rows observed classifications Columns: Predicted classifications			
	Percent	G1:1	G2:2	G3:3	G4:4
Group	Correct	p=0.22727	p=0.17045	p=0.28409	p=0.31818
G1:1	100.0000	20	0	0	0
G2:2	86.6667	0	13	1	1
G3:3	76.0000	0	0	19	6
G4:4	75.0000	0	1	6	21
Total	82.9545	20	14	26	28

The method involves the assignment of assigning an individual with $x = (\beta, \gamma,$ IgG, Alb, Tp, Tch) to the class with the longest value of the discriminant function. For example, when β =1, γ =8, IgG = 2000, Alb =60, Tp = 6.5 and Tch =130, we have $x = (1, 8, 2000, 60, 6.5, 130)$ and $L_1(x)$ = 460.556, $L_2(x)$ = 447.604, $L_3(x)$ = 463.269, $L_4(x)$ = 459.573. So, an individual having β =1, γ = 8, IgG = 2000, Alb = 60, Tp = 6.5 and Tch = 130 will be classified in class 3 (Febrile) since $L_3(x)$ = 463.269 and is the maximum. In the process of classification, there is the possibility of misclassification. The estimated percentages of correct classification were computed by STATISTICA and it yielded 82.9545% for the total and 100%, 86.6%, 76% and 75% respectively for class 1, class 2, class 3 and class 4 respectively.

Remark 1: Seshadri *et al.* (1983), Sharma *et al.* (1983) indicated the significant role of Albumin and Total Cholesterol in detecting malaria infected cases.

□

F. Multiple regression

In this section, we have presented multiple regression for some of blood parameters discussed above.

Dependent Variable	Independent Variable	Multiple regression
WBC	N,L,E,M	WBC=10594.4−35.69N−44.03L +46.33E−120.83M
RBC	Hb, Tc, Bc, IgG, IgM, IgA	RBC=0.13+0.365 Hb* −0.0006 Tc +0.002 Bc +0.00002 IgG −0.00009 IgM −0.0009 IgA
Tp	Alb,α_1, α_2, β, γ	Tp =2.303+0.051 Alb* +0.035α_1+0.045α_2 +0.046β =0.042γ*
Tch	Fc,Ec	Tch=14.260+0.874Fc* +0.8823 Ec*

* indicates significant (at 5% level) regression coefficient corresponding to the respective independent variable.

From the above table, we note that none of the independent variables N, L, B, M can be used effectively for predicting size of WBC since regression coefficients corresponding to the variables are independent. In a similiar fashion we we can conclude that only Hb, Alb, γ and both Ec and Bc can be used as reliable predictor for RBC, Tp, and Tch respectively.

4. CONCLUSION

The survey based on cross-sectional data analysis on the tribal population described in this paper revealed that the 5 blood parameters viz. Tp, Alb, Tch, IgG, β and γ, out of the 21 blood parameters considered, may be used for predicting malarial endemicity of the tribal community. The factors like blood group and sex do not take important roles for the prediction of malaria. There are fairly high positive correlations between Tch and Ec: Alb and Ec. But there are also moderately high negative correlations between R and N and N and M. The study is based on limited data. More accurate model making require the collection of further data, especially of females who are recalcitrant to respond due to their «social taboos».

ACKNOWLEDGEMENTS

We would like to thank the referee for his useful suggestions that greatly improved this work. We also thank Mr. A. Bose, Mr. P.P. Chakravorty and Mr. C. Kundu for statistical computations. Financial support was obtained from the Indian Council of Medical Research and the Indian Statistical Institute, Calcutta for conducting the survey.

REFERENCES

[1] **Blood, B. Ed.** (1977). *In vitro methods in cell-mediated immunity.* Academic Press, New York.

[2] **Davie, J.V.** and **Lewis, S.M.** (1984). *Practical H œmatology.* Churchill Livingstone.

[3] **Kshirsagar, A.M.** (1983). *A course in Linear Models.* Marcell Dekkar Inc., New York, 305–307.

[4] **Lowry, O.H., Rosenbrough, N.J., Farr, A.L.** and **Randall, R.J.** (1951). «Protein measurement with Folin Phenol Reagent». *Journal of Biological Chemistry,* **193**, 265–275.

[5] **Mancini, G., Carbonara, A.O.** and **Horemans, J.F.** (1965). «Immunochemical quantitation of antigens by single radial immuno-diffusion». *Immunochemistry,* **2**, 235.

[6] **Seber, G.A.F.** (1984). *Multivariate Observations.* John Wiley & Sons, USA 332–335.

[7] **Sheshadri, C., Shetty B. Ramkrishna, Gouri, N., Sitaraman, R., Revathi, R., Venkatraghaban, S.** and **Chari, M.V.** (1983). «Biochemical changes at different levels of parasitaemia in *Plasmodium vivax* Malaria». *Indian Journal of Medical Research,* **77**, 437–442.

[8] **Sharma, S., Shrivastava, I.K., Amarnath Dutta, G.P.** and **Agarwal S.S.** (1983). «Serum proteins and Immunoglobulin changes in Human Malaria». *Indian Journal of Malariology,* **20**, 15–19.

[9] **Smithies, O.** (1955). «Zone electrophoresis in starch gels: Group variations in the serum protein of normal human adults». *Biochemical Journal,* **61**, 629.

[10] **Wootton, I.D.P.** (1974). *Micro-analysis in Medical Biochemistry.*4th ed. (J. and A. Churchill Ltd.) 83–86.

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 22 N. 1

PROBLEMA N. 68

Els viatges a B, G, T del viatjant s'ajusten a una cadena de Markov amb tres estats B, G, T i matriu de transició homogènia:

$$\begin{array}{c} \\ \mathrm{B} \\ \mathrm{G} \\ \mathrm{T} \end{array} \begin{array}{c} \begin{array}{ccc} \mathrm{B} & \mathrm{G} & \mathrm{T} \end{array} \\ \begin{pmatrix} 0.5 & 0.3 & 0.2 \\ 0.5 & 0.5 & 0 \\ 0.5 & 0 & 0.5 \end{pmatrix} \end{array} = \boldsymbol{P}$$

1) Si un dilluns està a B, la distribució inicial és $(1 \quad 0 \quad 0)$ i la distribució del sistema dos dies després és

$$(1 \quad 0 \quad 0)\boldsymbol{P}^2 = (1 \quad 0 \quad 0)\begin{pmatrix} 0.5 & 0.3 & 0.2 \\ 0.5 & 0.4 & 0.1 \\ 0.5 & 0.15 & 0.35 \end{pmatrix} = \begin{array}{c} \begin{array}{ccc} \mathrm{B} & \mathrm{G} & \mathrm{T} \end{array} \\ (0.5 \quad 0.3 \quad 0.2) \end{array}$$

Resposta: La probabilidad de que el dimecres estigui a Girona és 0.3.

2) La matriu $\boldsymbol{P}$ és regular ja que $\boldsymbol{P}^2$ té tots els elements positius. Per tant, té distribució estacionària $(p \quad q \quad r)$ que verifica:

$$(p \quad q \quad r)\boldsymbol{P} = (p \quad q \quad r)$$

Resolent el sistema

$$\begin{aligned} 0.5p + 0.5q + 0.5r &= p \\ 0.3p + 0.5q &= r \\ 0.2p + 0.5r &= r \\ p + q + r &= 1 \end{aligned}$$

obtenim $p = 0.5$, $q = 0.3$, $r = 0.2$. La repartició del presupost $A = 1.000.000$ serà, per tant:

Barcelona: $0.5A = 500.000$
Girona: $0.3A = 300.000$
Tarragona: $0.2A = 200.000$

C.M. Cuadras
Universitat de Barcelona

PROBLEMA N. 69

1) La mitjana de $Y = T(1-X)$, on T té mitjana $\mu(T)$ i X és uniforme (0, 1) és:

$$E(Y) = E(T(1-X)) = E(T) \cdot E(1-X) = \frac{1}{2} E(T)$$

$$\Longrightarrow \quad E(T) = 2E(Y)$$

Resposta: $\mu(T) = 2\mu(Y)$.

2) Si T és exponencial negativa amb $\alpha = 1$, la densitat bivariant de (X, T) és:

$$f(x, t) = \mathrm{e}^{-t} \quad 0 < x < 1, \quad t > 0,$$

doncs X i T són independents.

Fem el canvi $\quad y = t(1-x) \quad z = x$

El canvi invers i el jacobià són

$$t = y/(1-z) \quad x = z$$

$$J = \begin{vmatrix} \dfrac{\partial t}{\partial y} & \dfrac{\partial t}{\partial z} \\ \dfrac{\partial x}{\partial y} & \dfrac{\partial x}{\partial z} \end{vmatrix} = \begin{vmatrix} \dfrac{1}{(1-z)} & \dfrac{y}{(1-z)^2} \\ 0 & 1 \end{vmatrix}$$

$$|J| = \frac{1}{1-z}$$

Resposta: La densitat bivariant de (X, Y) és

$$f(x, y) = \frac{1}{1-x} \cdot \mathrm{e}^{-\frac{y}{1-x}} \quad 0 < x < 1, \quad y > 0.$$

C.M. Cuadras
Universitat de Barcelona

PROBLEMES PROPOSATS

PROBLEMA N. 70

La funció de distribució de les variables aleatòries (U, V) és:

$$H(u, v) = (u\,v)^{(1-\theta)}\,(\min\{u, v\})^{\theta} \quad 0 < u, v < 1$$

on $0 \leqslant \theta \leqslant 1$ és un paràmetre.

1) Demostra que les distribucions marginals de U i V són uniformes $(0, 1)$.
2) Comprova que si $\theta = 0$, U i V són estadísticament independents, i que si $\theta = 1$, el coeficient de correlació entre U i V és $\rho = 1$.
3) Prova que el paràmetre θ verifica:

$$\theta = \lim_{u=v\to 1} \left(\frac{H(u, v) - uv}{\min\{u, v\} - uv} \right)$$

C.M. Cuadras
Universitat de Barcelona

PROBLEMA N. 71

La matriu de correlacions de les variables $X = (X_1, X_2, \ldots, X_p)$ és R. Sigui Y_1 la primera component principal de X, obtinguda a partir de $\boldsymbol{R}$.

Siguin $\rho(X_i, Y) \quad i = 1, \ldots, p$ els coeficients de correlació entre cada X_i i una variable $Y = a_1 X_1 + \cdots + a_p X_p$ combinació lineal de X.

Demostra que

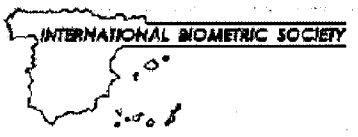
$$\max \sum_{i=1}^{p} \rho^2(X_i, Y) = \sum_{i=1}^{p} \rho^2(X_i, Y_1),$$

és a dir, la primera component principal Y_1 fa màxima la suma de correlacions al quadrat entre una variable composta i les variables X.

C.M. Cuadras
Universitat de Barcelona

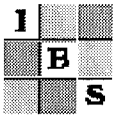
Cursos i congressos d'estadística

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

Presidente:	Emilio A. Carbonell Guevara
Vicepresidente:	Guadalupe Gómez i Melis
Secretario y Tesorero:	Fernando López Santoveña
Vocal en calidad de Miembro del Consejo de la IBS:	Carles Cuadras Avellana
Vocales:	Rosa Estarelles Rodríguez
	Martin Ríos Alcolea
	Juan Luis Chorro Gascó
	José Luis González Andújar
	Alex Sánchez Pla
	María Jesús Bayarri García
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 el primer semestre de 1998 se ha celebrado un curso sobre «Regresión Logística» y otro sobre «Modelos Mixtos con S-Plus».

Actualmente se está estudiando la programación de dos nuevos cursos: «Modelos de Supervivencia» y «Métodos Gráficos de Análisis de Datos».

VII Conferencia Española de Biometría
Reunión bianual de la Sociedad Española de Biometría
10-12 de marzo, 1999. Palma de Mallorca.
http://iris1.uib.es/people/biome99/biometria99@clust.uib.es

The TES Institute
Training of European Statisticians

TES OR THE ENHANCEMENT OF STATISTICS

The TES Institute is a non-profit association created in November 1996 by the Directors General of the National Statistical Institutes of ten Member States of the European Union and of the four Member States of the European Free Trade Association.

As an international post-graduate vocational training institute for statisticians, the mission of the TES Institute is to create truly European vocational training and staff development opportunities at post-graduate level. The annual training programmes are designed for target groups ranging from young statisticians to executives of National Statistical Institutes.

Over the last five years, the objectives of the TES Institute were *(a)* to provide training courses and seminars of short duration at post-graduate level, thus offering statisticians opportunities to perfect their skills, *(b)* to provide a forum for mutual consultation on vocational training for statisticians, *(c)* to give assistance in the training for new European statistical projects, thus disseminating standards, European methods and classifications, and to help future members to prepare for their integration into the European Statistical System, and *(d)* to promote exchanges of skills and experiences.

TYPES OF TRAINING

The TES Institute offers two types of training: the *Core Programme* and the *Special Courses Programme*. Both of them are covering the following specialisation areas:

(1) *Data Collection and Survey Methodology*, (2) *Economic Statistics*, (3) *Social Statistics* and (4) *Publication, Dissemination and Use of Statistics*, complemented by the *Common Courses* (general «statistical» culture) and the *Support Courses* (statistical analysis, computing techniques and statistical management).

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 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.

TRAINING ORGANISATION

Obviously, vocational training programmes of this kind imply the involvement of different groups of people.

- The training staff of the TES courses consists of *an international pool of trainers* (highly qualified practitioners and university professors). This pool is responsible for the design and the execution of the courses and the development of the course material, which are annually reassessed and updated where necessary. Each year the organisation of the programme requires the intervention of about hundred trainers.

- The National Statistical Institutes of the European Union (EU), the European Free Trade Association (EFTA), the Central and Eastern European Countries (ECO[1]) and the Mediterranean Basin countries[2] have nominated *TES Correspondents* for the TES programme. This network of TES Correspondents is responsible for the communication between their Institute and the TES Institute, for the co-ordination of the registration of candidates and for the dissemination of information about TES courses.
- The main task of the staff of the TES Institute is to assist in the design, to organise and to give logistic support to the international vocational training programme and to provide information about existing training possibilities in the various countries in Europe. To carry out these activities, the TES Institute comprises a permanent team of eleven people.

CURRENT ACTIVITIES

Training Core and Special Courses Programmes

Over the last five year, the TES Institute organised some 110 courses in the framework of the *Core Programme* open to EU, EFTA and ECO statisticians. Besides, *Special Courses* have also been organised especially for statisticians from ECO countries.

On the whole, about 600 participants have been trained every year through these two kind of training programme.

Moreover, since 21st May 1997, when the first Eurostat *Task Force on Training* for MEDSTAT countries was organised, the TES Institute officially extended its training activities to the Mediterranean Basin countries.

Ad hoc projects

Recently, a seminar on «Education Statistics» has been organised at the request of Eurostat (Vienna, January 1998).

Moreover, two ad hoc courses (5 days) for the statistical service of Malta: the course «Enterprise Statistics» and the course «National Accounts Statistics in Practice». Both courses were executed end 1997.

---

[1] Albania, Bulgaria, Czech Republic, Estonia, Hungary, FYROM (Former Yugoslav Republic of Macedonia), Latvia, Lithuania, Poland, Romania, Slovakia and Slovenia.

[2] Algeria, Cyprus, Egypt, Israel, Jordan, Lebanon, Malta, Morocco, Palestine, Syria, Tunisia, Turkey.

In a near future, the TES Institute will carry out:

- Two seminars on «Statistics and Policy Making» at the request of the Palestinian Statistical Training Centre (Gaza, Westbank). These seminars will be financed through UN funds and are carried out in co-operation with UNITAR, the UN Institute for Training and Research, Geneva. Again this seminar may be introduced, in a modified form, in the future regular programme.
- The first part of a «Summer School on Social Statistics» in July 1998 ($5\frac{1}{2}$ days) in Siena. The other two parts are planned for summer 1999 and 2000.
- Specific training will be provided to statisticians from the Ukrainian Statistical Office. The training will cover *Business Registers, Enterprise Statistics and National Accounts.*

Consulting

Apart from the well-known training activities, the TES Institute has recently set up a procedure for consulting activities.

In this context, the TES Institute is requested to provide support to the set-up of a training centre for statisticians in a specific country in Ukraine. This support will consist of *(a)* training of future trainers, as above mentioned and *(b)* consulting for curriculum development for their future training programmes.

Publication

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 decided to set up a TES Manual series for the manuals to be developed in the specialisation areas covered by the vocational training programme (i.e. *Data Collection and Survey Methodology, Economic Statistics, Social Statistics and Publication, Dissemination and Use of Statistics*).

By the end of this year, the following publications will be available at the TES Institute:

- «*Introduction to Demographic Data Analysis*» by Mr Otto Andersen from Statistics Denmark. This manual will cover subjects like: data in demography, lexis diagram, fertility, mortality, etc. As the manual will be available at the TES Institute.

- «*Index Number Theory and Practice*» by Prof. Marco Martini from the University of Milano.

Please inform the TES Institute if you are interested in the above mentioned publications (articles and/or manuals) and detailed information will be forwarded to you.

GENERAL INFORMATION

For further information about courses, new brochure, publications, specific projects and so on, please contact Ms Valérie Vandewalle:

by phone:	+352/29 85 85 34
fax:	+352/29 85 29/30
e-mail:	vvandewalle@tes-institute.lu
website:	http://www.tes.institute.lu.institute
for mail:	5, rue Guillaume Kroll
	L - 1882 Luxembourg

Activities of the TES Institute from September to December 1998

The new 1998-1999 training programme is planned from September 1998 to June 1999 and comprises 31 courses organised whole over Europe. In its framework, nine courses will be held between September and December 1998.

From 21 to 25 September, Mr Wright (ONS Consultant) will deal with how to make an *effective presentation of official statistics to the news media.* The course will be organised in the premises of the Office for National Statistics in London. This course should provide the participants with a better knowledge of the interaction between the news media and the NSIs. They will be able to select topics for news releases and to prepare plans for special announcements.

A course on the *Revised European system of accounts focused on quarterly accounts* will be organised in the premises of the TES Institute in Luxembourg from 28 to 30 September. Through the course, Mr Mazzi (Eurostat) will provide experts on national accounts of NSIs, central banks, research institutes,... with an understanding of all the characteristics of Quarterly Accounts and current methodologies.

The theory and practice on index numbers will be presented by Prof. Martini (University of Milan). The course will take place in the University of Milan from 12 to 16 October. Mr Martini will help the participants in getting the ability to choose and use the techniques appropriate to the practical problems related to the construction of index numbers.

After having organised a course on the systems of health statistics in the previous programme, Mr Bonte (Eurostat) will train participants in *health statistics based on interview surveys*. This course will take place in Heerlen from 26 to 29 October and will help participants in developing, managing, analysing and publishing results of a comprehensive HIS that covers national information needs and of how and when to use internationally recommended and accepted methods and instruments to enlarge the survey's capacities for improving international comparisons..

The construction and use of the employment cost index will be organised in Brussels from 2 to 4 November. The course will provide the knowledge on how the US employment cost index was developed and how it could be constructed in the EU.

The theory and practice of National Accounts will be presented by a training team from Statistics Netherlands from 2 to 13 November. After the course, participants will have a general understanding of the system of national accounts and the problems in relations between data sources and national accounts.

A course on *the measurement of statistics* will be held in Luxembourg from 16 to 18 November. The teaching will be executed in association between Mr Depoutot (Eurostat) and Ms Elvers (Statistics Sweden). Trough lectures, guidelines presentations, case studies and group discussions, the participants will be acquainted with the requirements of the European Statistical System and quality reports and their implementation in practice. They will also have the opportunity to exchange their experience on methods used in other organisations regarding the measurement of the quality of statistics.

Another important topic will be dealt from 23 to 26 November in Voorburg and concerns the *statistical disclosure control.* Mr Willenborg (Statistics Netherlands) will provide the participants with an understanding of statistical disclosure control methods and to train them in using the ARGUS software.

A second course on the *revised system of accounts* will be organised focused this time on *sector accounts*. The course will be held in the premises of the TES Institute and provide a good understanding of structures, concepts and definitions of the non-financial accounts of ESA.

Please note that all the courses offered by the TES Institute are divided between both theoretical and practical sessions. The lectures will provide you the theoretical background necessary for further developments and the practical sessions will offer you the opportunity to discuss and exchange experience with other colleagues.

FIRST ANNOUNCEMENT:

Fifth International Conference of

«Forum for Interdisciplinary Mathematics»

International Conference on Combinatorics, Statistics, Pattern Recognition and Related Areas

December 28 - 30, 1998

Vanue: Mysore University,

Mysore - 570006, Karnataka, India

«Organizing Committee»

Professor Bhu Dev Sharma, Convener (Past President of the Forum; Vice President), Mathematics Department, Xavier University of Louisiana, New Orleans, LA 70125, USA; phone (504)483-7463; fax (504)485-7928; bsharma@mail.xula.edu

Professor N.R. Mohan (Local Secretary), Chairman, Department of Statistics, University of Mysore, Mysore 570006, India; fax: 011(91-821)421-263; chairman@giasbg01.vsnl.net.in

Professor Sat N. Gupta (Vice President of the Forum), Department of Mathematics & Statistics, University of Southern Maine, Port Land, Maine 04103, USA; phone (207)780-4278; fax (207)780-4933; rxm381@usm.maine.edu

Professor Satya N. Mishra (Vice President of the Forum), Department of Mathematics, University of South Alabama, Mobile, AL 36688; phone (334)460-6264/6474; fax (334)460-7969; smishra@jaguarl.usouthal.edu

«International Organizing Committee»

(Incomplete list)

Professor C.R. Rao (Past President of the Forum), Department of Statistics, Pennsylvania State University, University Park, PA 16802, USA, phone (814)865-3194; e-mail: crrl@psuvm.psu.edu

Professor M.M. Rao (President of the Forum), Dept. of Mathematics, University of California, Riverside, CA 92921, phone (909)787-5944; e-mail: rao@ucrmath.ucr.edu

Professor C.M. Cuadras, Dept. d'Estadística, University of Barcelona, Av. Diagonal, 645, 08028 Barcelona, Spain; phone (34)93 4021565; fax (34) 93 4110969; e-mail: carlesm@porthos.bio.ub.es

Professor P. Nagabhushan, Chairman, Dept. of Mathematics, University of Mysore.

Professor Padmavathamma, Chairman, Dept. of Comp. Sci, University of Mysore.

Professor E. Sampathkumar, Dept. of Mathematics, University of Mysore.

Call for papers

Abstracts of papers, in any area of the conference, for presentation and registration from persons outside India be addressed to Drs. Sharma, Gupta or Mishra at their addresses, phone/fax/e-mail addresses given above. Those from India these be sent to Professor N.R. Mohan (address above).

Symposia:

Several symposia in the fields of Combinatorics, Statistics, Computer Science and related areas are being planned, and shall be soon announced.

Registration:

US $125.00 for persons from outside India (before September 1; add $25 after that).
Rs. 600.00 for persons from India (before September 1; add Rs. 25.00 after that).

Mysore:

Mysore, a beautiful city, has been the capital of an earlier major princely state of India. It has world famous palaces, Vrindavan Gardens, Zoo with rare animals, and museum, etc. It well connected by air, rail or road transport.

WORKSHOP
GIE

Jornades internacionals sobre

Generació d'informació estadística: qualitat i limitacions

Barcelona,
30 de novembre - 1 de desembre de 1998

XARXA TEMÀTICA

ENQUESTES I QUALITAT DE LA INFORMACIÓ ESTADÍSTICA

Col·laboren:

- Institut d'Estadística de Catalunya (IDESCAT)
- CIRIT, Generalitat de Catalunya
- Service pour la Science et la Technologie de l'Ambassade de France

Organitzen:

- Departament d'Econometria, Estadística i Economia Espanyola, Universitat de Barcelona
- Departament d'Estadística i Investigació Operativa, Universitat Politécnica de Catalunya

Generalitat de Catalunya
Departament de la Presidència
Comissionat per a Universitats i Recerca

II Pla de Recerca de Catalunya 1997/2000

Generació de la informació estadística: qualitat i limitacions

La preocupació per la qualitat de les dades en les enquestes i en la recollida d'informació estadística és essencial en la producció de resultats fiables. L'estudi dels fenòmens socio-econòmics no pot ignorar la rellevància de la generació de les dades i els límits de la utilització de la informació estadística recollida per sondeig o per qualsevol altra font.

Es pretén donar al Workshop un enfocament pràctic i afavorir la sinergia entre investigadors i professionals del sector públic i de l'empresa.

A qui s'adreça el Workshop?

Aquestes jornades s'adrecen a estadístics i usuaris de l'estadística, investigadors i professionals del món institucional o empresarial, preocupats per la qualitat i fiabilitat de la informació estadística.

Comitè de programa:

Manuela Alcañiz (UB, Barcelona)
Tomàs Aluja (UPC, Barcelona)
Joan M. Batista (URL, Barcelona)
Mónica Bécue (UPC, Barcelona)
Germà Coenders (UG, Girona)
Michael Greenacre (UPF, Barcelona)
Montse Guillén (UB, Barcelona)
Enric Ripoll (IDESCAT, Barcelona)
Frederic Utzet (UAB, Bellaterra)
Sylvie Viguier-Pla (U. Perpignan, Perpignan)

Informació i inscripcions:

GIE Mónica Bécue
Dept. d'Estadística i Investigació Operativa
Facultat de Matemàtiques i Estadística (UPC)
Pau Gargallo, 5; 08028 Barcelona
Fax +34 93 401 51 55; e-mail: monica@eio.upc.es

Ponents convidats:

Àlex Costa (IDESCAT, Barcelona)
Reyes Crespo (Dirección General de Aduanas, Madrid)
François Héran (INSEE-INED, París)
Ludovic Lebart (CNRS-ENST, París)
Valentín Martínez (CIS, Madrid)
Juana Porras (INE, Madrid)
Enric Renau (Institut DEP-UPF, Barcelona)
William Saris (Universitat d'Amsterdam)
Yves Tillé (ENSAI-INSEE-CREST, Rennes)
Laurent Toulemon (INSEE, París)

Programa provisional:

Es realitzaran quatre sessions monogràfiques i l'exposició de ponències.

1. Disseny i elaboració de mostres estadístiques.
2. Dades longitudinals i dades de panell.
3. Operacions censals i explotació de registres administratius.
4. Disseny de qüestionaris i recollida de dades.

Els participants que desitgin presentar una comunicació n'hauran d'enviar un resum d'una pàgina abans del 15 de setembre de 1998. L'acceptació es comunicarà abans del 15 d'octubre.

Lloc:

Palau de les Heures
Universitat de Barcelona
Campus de la Vall d'Hebron
Metro: L3 (Estació Montbau)
Autobusos: 27, 60, 73, 76, 85, 173

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 certificat de rebut corresponent a la data de presentació, la qual figurarà com a «data de rebut» en la publicació final de l'article. Per la seva banda, la «data d'acceptació» de l'article que figurarà en la publicació 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 als 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 LaTeX 2ε [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, PCX o BMDP. 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 LaTeX 2ε 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
Telèfon: +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 wellcomes for their publication 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 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 sould 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 sended in event of not being possible their autoedition.

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 2ε 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, PCX or BMDP 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. On the other hand, 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 that the original shows).

The Secretary of *Qüestiió* could send, by request of the authors, the LaTeX 2ε style of *Qüestiió* for manuscript preparation and the appropriate AMS classification references.

QÜESTIIÓ
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona

Telèfon: +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
Telèfon: +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 the sponsor entities, collaborating institutions, other public and private bodies and the entire scientific community related to statistics or operational research to submit institutional advertisements on courses, seminars, congress and similar activities that will be held preferably in our country 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 these events and therefore does not publish a comprehensive list of such activities.

Once its insertion is approved the advertisements will be 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 editing process to the printed version that the advertiser has sent. If the original advertisements are sent in the same electronic terms requested by the articles (please see «Guidelines for the submission of articles for *Qüestiió*») the journal will proceed to its autoedition. If a better quality than the simple reproduction or autoedition or a colour publication of the adverts it's wanted, the authors should contact the Secretary of *Qüestiió* in order to send the photolits of the corresponding original texts.

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 people 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 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 in the agreed terms.

Mònica M. Jaime
Secretaria de *Qüestiió*
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona

Telèfon: +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 COMERCIALS 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* la, 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.

Pagament:

Les tres modalitats admeses són la transferència bancària, el pagament amb targeta de crèdit o la tramesa postal d'un xec bancari a la secretaria de *Qüestiió*, per l'import que s'indiqui a la factura corresponent (on hi figurarà l'import del cost dels fotolits si la filmació de l'anunci ha estat a càrrec de la revista). En el cas que l'anunciant desitgi una factura proforma cal que ho sol·liciti amb l'antelació suficient.

Compte bancari:

2013-0100-53-0200698577 (Institut d'Estadística de Catalunya)
Caixa de Catalunya
c. Comte d'Urgell, 182; 08036 Barcelona

Correspondència:

Mònica M. Jaime
Secretaria de *Qüestiió*
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona

Telèfon: +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:

The three accepted ways of payment are by a bank transfer, charge on a credit card or remittance by mail of a bank cheque to the *Qüestiió* secretary for the amount shown at the invoice (where it will be shown 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 propper address.

Account number:

2013-0100-53-0200698577 (Institut d'Estadística de Catalunya)
Caixa de Catalunya
c. Comte d'Urgell, 182; 08036 Barcelona

Mail address:

Mònica M. Jaime
Secretaria de *Qüestiió*
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona

Telèfon: +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 1998.
El preu de la subscripció és de 3.000 PTA (IVA inclòs).

Forma de pagament

☐ Transferència al compte de la Caixa de Catalunya 2013-0100-53-0200698577
Comte d'Urgell 162, 08036 Barcelona

☐ Domiciliació bancària al compte número

☐ Taló nominatiu a l'Institut d'Estadística de Catalunya

☐ Gir postal

☐ En efectiu

Retornar aquesta butlleta (o una fotocòpia) a:

Qüestiió:
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona

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

SOLAS *l'anàlisi de les dades no disponibles*

COM TRACTAR LES DADES NO DISPONIBLES?

Són molts els estudis avançats que han estat perjudicats per l'ús de mètodes puntuals en el tractament de dades no disponibles, com ara per exemple les "anàlisis de cas global" o les "anàlisis de cas disponible". Els organismes de control consideren aquests mètodes cada cop més esbiaixats. D'acord amb l'índole de dades que no es troben disponibles (ja siguin acotades, univariants o multivariants), dins d'aquest conjunt de dades, Solas permet escollir entre les tècniques més adients per aplicar a un subconjunt de les dades o a totes en conjunt.

SOLAS, l'anàlisi de les dades no disponibles, és un paquet de programari estadístic completament integrat dins l'entorn Windows i inclou les últimes tècniques reconegudes per la indústria en els tractaments dels valors no disponibles. Les tècniques d'imputació inclouen:

Imputació múltiple

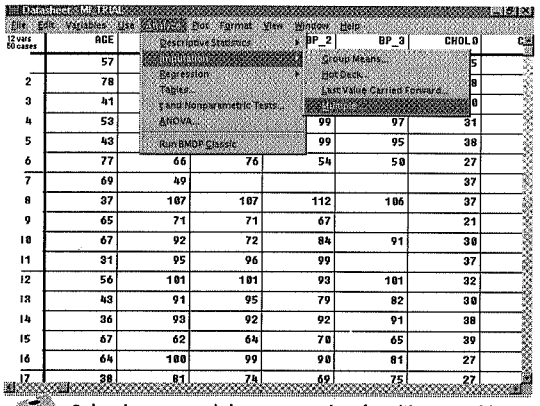
- Basada en tècniques desenvolupades per Rubin et altri.
- Genera imputacions múltiples en utilitzar unes puntuacions de propensió, així com la rutina bayesiana d'inicialització més adient.
- Permet l'anàlisi de les dades longitudinals.

Imputació senzilla

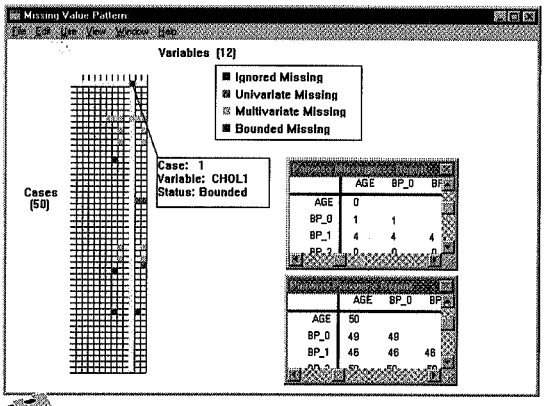
- Últim valor transferit: permet a l'usuari imputar el valor que falta en funció de l'últim valor observat.
- Mitjana de grup: permet la imputació d'una mesura de grup (o d'una modalitat en cas que les dades siguin categòriques).
- Hot decking: permet a l'usuari especificar i prioritzar totes les combinacions de variable per a una imputació del tipus "correspondència més propera".

Patró SOLAS de dades no disponibles

El patró de dades no disponibles és una exclusiva de Solas i ofereix a l'usuari una visualització clara i global de la quantitat, la localització i els tipus de dades no disponibles a cadascuna de les cel_les. Aquesta prestació ofereix la possibilitat d'aïllar, encara més, les cel_les individuals d'un conjunt de dades i identificar-ne les configuracions per defecte d'observació i d'estat. Un cop acabada la imputació aquest dispositiu gràfic, codificat amb colors, identifica amb facilitat el mètode de tractament seleccionat.



Seleccioneu un dels quatre mètodes d'imputació disponibles amb SOLAS per a l'anàlisi de les dades no disponibles.



Determineu a quin tipus pertanyen els valors no disponibles utilitzant l'Informe de Patró de Dades No Disponibles.

Statistical Solutions Ltd.
8 South Bank
Crosse's Green, Cork, Irland

Tel. + 353 21 319629 • Fax + 353 21 319630
Email: sales@statsol.ie
Website: http://www.statsol.ie*

Statistical Solutions
Stonehill Corporate Center Suite 104
999 Bradway, Saugus, MA 01906

Tel. (781) 2317680 • Fax (781) 2317684
Email: info@statsolusa.com
Website: http://www.statsolusa.com

solas *For Missing Data Analysis 1.0*

HOW DO YOU HANDLE MISSING DATA?

Many advanced studies have suffered from the use of ad-hoc methods of handling missing data such as "complete-case-analysis" or " available-case analysis". These methods are becoming increasingly viewed as biased by regulatory agencies. According to the types of missing data (bounded, univariate or multivariate) in your dataset SOLAS allows you to choose the most appropriate techniques to apply to a subset or all of your data.

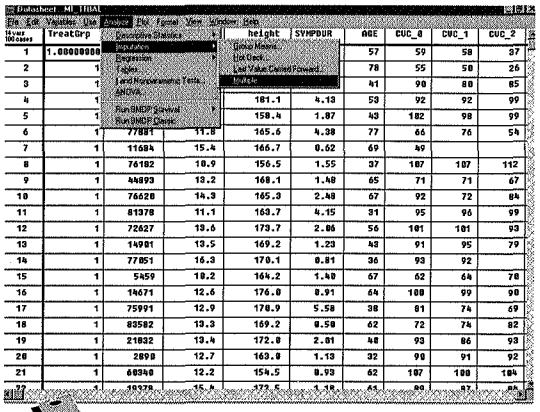
solas *For Missing Data Analysis* , a fully Windows based statistical software package, incorporates the latest industry accepted techniques for treating missing values. Imputation techniques include :

Multiple Imputation

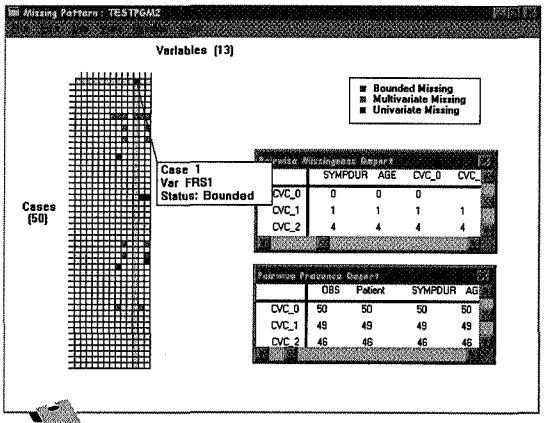
- Based on techniques developed by Rubin et Al.
- Generates multiple imputations using propensity scores and the approximate Bayesian bootstrap.
- Allows for analysis of longitudinal data.

Single Imputation

- Last Value Carried Forward - allows the user to impute Missing Value based on last value observed
- Group Means - Imputation of group mean (or mode in the case of categorical data).
- Hot Decking - Allows the user to specify and prioritize all combinations of variables for 'closest match' imputation



Select one of the four imputation methods available in SOLAS for Missing Data Analysis



Identify the type of missing values using the Missing Pattern Report

Solas Missing Data Pattern

The unique missing data pattern in Solas provides the user with a clear over-view of the quantity, positioning and types of missing values in each cell. This feature allows you to further isolate particular cells in a data set and identify observation and status defaults. Following imputation this colour coded graphics feature easily identifies the treatment method selected.

Statistical Solutions Ltd.
8 South Bank
Crosse's Green, Cork, Irland

Tel. + 353 21 319629 • Fax + 353 21 319630
Email: sales@statsol.ie
Website: http://www.statsol.ie*

Statistical Solutions
Stonehill Corporate Center Suite 104
999 Bradway, Saugus, MA 01906

Tel. (781) 2317680 • Fax (781) 2317684
Email: info@statsolusa.com
Website: http://www.statsolusa.com

nQuery Advisor® Versió 2.0

Programari de Planificació d'Estudis

Determinació de la dimensió i la potència de la mostra

nQuery Advisor és una eina molt útil per a investigadors i estadístics a l'hora de planificar els estudis d'investigació i de garantir un ús eficaç dels recursos.

La correcta determinació de la dimensió i de la potència de la mostra representa un aspecte important dins de la planificació dels estudis.

nQuery Advisor proporciona uns mètodes simples, fiables i eficaços per determinar aquests valors, a més d'una experta assistència en computació i una àmplia cobertura quant als problemes relacionats amb la dimensió i la potència de les mostres, així com les respostes sobre la dimensió de les mostres per a les anàlisis de l'interval de confiança i d'equivalència. En la major part dels problemes que es presenten permet que els grups posseeixin una dimensió igual o desigual.

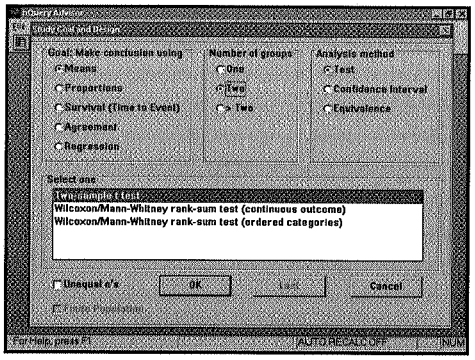
La versió 2.0 de nQuery Advisor compta amb notables millores dissenyades per tal de facilitar, encara més, la planificació dels seus estudis d'investigació. S'han inclòs noves taules amb opcions expandides per a la planificació de les dimensions de les mostres en els estudis de supervivència, el mostreig de poblacions finites, el mostreig per grups i, també, les anàlisis que utilitzen proves no paramètriques. Tant el manual com l'ajuda en pantalla s'han ampliat per proporcionar més detalls sobre l'ús de nQuery Advisor en mesures repetides i dissenys encreuats, demostracions d'equivalència i anàlisis de supervivència.

Per aquells investigadors que estan planificant assajos clínics o altres estudis amb supervivència del temps a un esdeveniment com a punt final, nQuery Advisor versió 2.0 introdueix una nova taula de dimensions de mostres per a les anàlisis de supervivència. Aquesta taula permet a l'usuari especificar unes corbes de supervivència que posseeixen una forma no exponencial o bé ràtios de risc que no es mantenen sempre constants, a més d'especificar uns patrons complexos d'acumulació o d'abandonament.

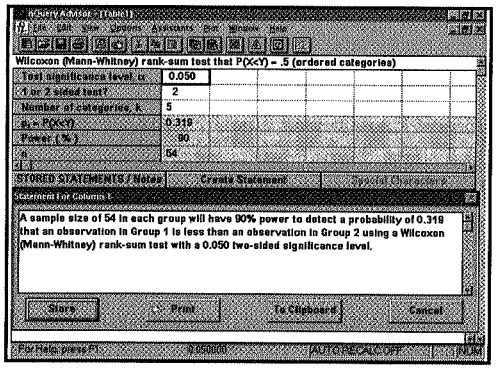
Els investigadors de ciències socials sabran, sens dubte, apreciar les noves opcions per a la planificació dels estudis basats en una població finita. La versió 2.0 de nQuery Advisor ajuda a estimar la desviació estàndard de grups per poder determinar el nombre de grups sobre els quals cal fer el mostreig.

La versió 2.0 de nQuery Advisor incorpora a més computacions de dimensions de mostra en pretendre utilitzar la proba de rang-suma de Wilcoxon/Mann-Whitney per analitzar els resultats d'un estudi de dos grups.

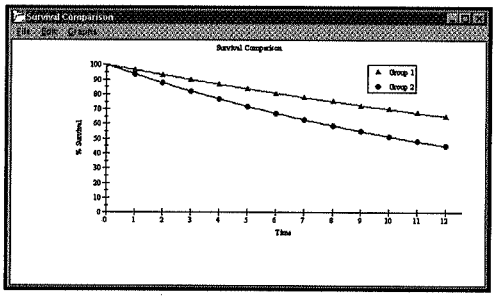
Quan s'utilitza en un entorn Windows nQuery Advisor, proporciona: Un índex estructurat per tal de facilitar l'especificació de l'objectiu de la investigació; Entrades i visualització de la dimensió i potència de mostres realitzades similars a les dels fulls de càlcul; Fitxes guia per ajudar els usuaris amb les seves preguntes estadístiques i pràctiques sobre les entrades; Eines de computació per especificar les dimensions d'efecte i les estimacions de variabilitat; Traces que mostren la relació entre la potència i la dimensió de les mostres. Quant als resultats computats proporciona informes de justificació de les dimensions de la mostra preparats per a la seva protocol_lització; facilitat en comparar els resultats sota diferents condicions o anàlisis; fórmules i algoritmes seleccionats d'acord amb les avaluacions comparatives i accés a les funcions de distribució per als usuaris experts.



Les noves prestacions introduïdes en la versió 2.0 inclouen proves no paramètriques i mostreigs en funció de les poblacions finites.



Per als resultats computats, s'han preparat informes de justificació de la dimensió de la mostra per a la seva protocol_lització.



nQuery Advisor versió 2.0 permet a l'usuari especificar ja les corbes de supervivència.

nQuery Advisor és un producte desenvolupat per la Dra. Janet Elashoff de Dixon Statistical Associates, Los Angeles, EE.UU.

Statistical Solutions Ltd.
8 South Bank
Crosse's Green, Cork, Irland

Tel. + 353 21 319629 • Fax + 353 21 319630
Email: sales@statsol.ie
Website: http://www.statsol.ie*

Statistical Solutions
Stonehill Corporate Center Suite 104
999 Bradway, Saugus, MA 01906

Tel. (781) 2317680 • Fax (781) 2317684
Email: info@statsolusa.com
Website: http://www.statsolusa.com

nQuery Advisor® Release 2.0

STUDY PLANNING SOFTWARE

Sample size and Power determination

nQuery Advisor is a very useful aid to investigators and statisticians in planning research studies to ensure efficient use of resources. Determination of the correct sample size and power are important aspects of study planning.

nQuery Advisor provides simple, reliable, and efficient methods for determining these values. It provides expert computational assistance and extensive coverage of sample size and power determination problems as well as sample size answers for confidence interval and equivalence analyses. It allows for either equal or unequal group sample sizes for most problems.

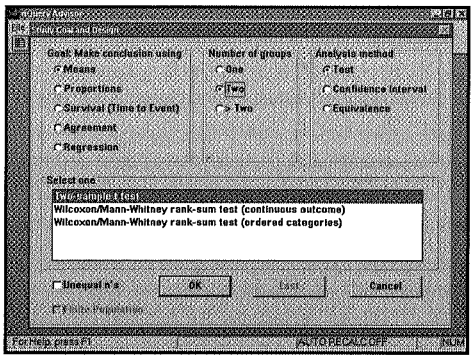
nQuery Advisor Release 2.0 features major new enhancements designed to make planning your research studies even easier. New tables have been included with expanded options for planning sample sizes for survival studies, sampling from finite populations, cluster sampling, and for analyses using nonparametric tests. The manual and on-line help have been expanded providing more details on the use of nQuery Advisor for repeated measures and crossover designs, equivalence demonstrations, and survival analysis.

For researchers planning clinical trials or other studies with survival or time to an event as an endpoint, nQuery Advisor Release 2.0 introduces a new sample size table for Survival Analysis. This table allows user specification of survival curves which are not exponential in shape, or hazard ratios which are not constant throughout, and user specification of complex patterns of accrual or dropout.

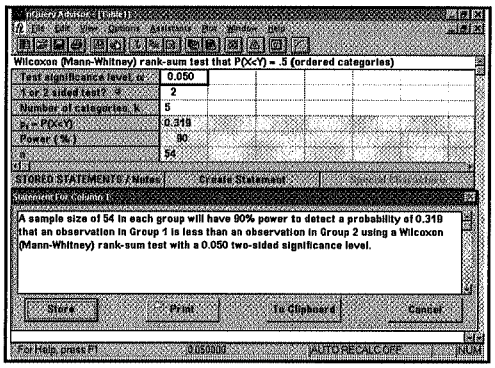
Social science researchers will appreciate the new options for planning studies for sampling from a finite population. nQuery Advisor Release 2.0 now provides sample size tables with a finite population correction for a range of study designs with testing and confidence interval goals.

The design of community intervention studies often requires randomization of clusters of subjects. nQuery Advisor Release 2.0 assists in estimation of the cluster standard deviation to use in determining the number of clusters which need to be sampled.

nQuery Advisor Release 2.0 also includes sample size computations when the Wilcoxon/Mann-Whitney Rank-Sum test will be used to analyze results of a two-group study.



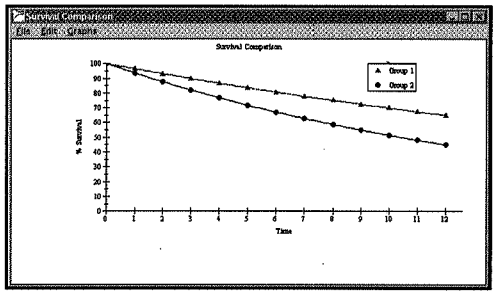
Among the new features introduced in version 2.0 are Non-Parametric tests, and sampling from a finite population.



Protocol-ready sample size justification statements for the computed results.

Operating in a Windows environment nQuery Advisor® provides:

- A structured index to assist in specifying the research goal
- Spreadsheet-style tabular entry and display of sample size and power
- Guide cards to assist users with statistical and practical questions on input
- Computational aids for specifying effect sizes and estimates of variability
- Plots showing the relationship between power and sample size
- Protocol-ready sample size justification statements for the computed results
- Ease of comparison of results under different conditions or analyses
- Formulas and algorithms chosen on the basis of comparative evaluations
- Access to distribution functions for expert users



nQuery Advisor® Release 2.0 now allows user specification of survival curves.

nQuery Advisor® is developed by Dr. Janet Elashoff Ph.D.of Dixon Statistical Associates, Los Angeles, U.S.A.

Statistical Solutions Ltd.
8 South Bank
Crosse's Green, Cork, Irland

Tel. + 353 21 319629 • Fax + 353 21 319630
Email: sales@statsol.ie
Website: http://www.statsol.ie*

Statistical Solutions
Stonehill Corporate Center Suite 104
999 Bradway, Saugus, MA 01906

Tel. (781) 2317680 • Fax (781) 2317684
Email: info@statsolusa.com
Website: http://www.statsolusa.com