Qüestiió

Quaderns d'Estadística i Investigació Operativa

Any 2002, volum 26, núm. 1-2
Segona època

Entitats patrocinadores:

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

Entitat col·laboradora:

International Biometric Society

Generalitat de Catalunya
Institut d'Estadística de Catalunya

Any 2002, volum 26, núms. 1-2

SUMARI

EDITORIAL

Amb caràcter extraordinari, el dos primers números del volum 26 (any 2002) s'editen en una única publicació per tal d'accelerar al màxim la publicació d'articles ja acceptats que corresponen al format de la segona època de la revista i, d'aquesta manera, poder inaugurar la nova etapa de la revista a partir del volum 27 (2003) amb els originals que satisfacin els criteris que s'hi han previst. I tot això sense comprometre ni el volum habitual de continguts que equivalguin a dos números de *Qüestiió* ni alterar els terminis per a la tramesa d'exemplars als subscriptors i lectors de la revista. En conseqüència, aquest doble número recull l'edició d'un total de 15 articles (mitjana de dos números d'un volum anual, d'acord amb la tendència dels darrers anys), repartits entre les quatre seccions temàtiques de la revista, acompanyats de dos treballs a la secció docent i dues recensions a l'apartat de comentari de llibres.

D'altra banda, com ja és costum, recollim en aquestes línies algunes referències actualitzades sobre les consultes a l'edició electrònica de *Qüestiió*. En aquest sentit, cal fer esment de l'augment continuat de les consultes al lloc web de l'Idescat sobre *Qüestiió* en els cinc primers mesos de l'any, on els accessos ja han superat les 52.600 peticions http, és a dir, més del 22% de les enregistrades en el mateix període de l'any anterior, el qual ja va experimentar un increment de consultes del 133% respecte al 2000.

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

En aquest número doble 1-2 del volum 26 (2002) s'hi publiquen quinze articles, dels quals n'hi ha set a la secció «Estadística», un a «Investigació Operativa», tres a «Estadística Oficial» i dos a «Biometria», a més de dos articles addicionals de caràcter docent.

Pel que fa a la secció «Estadística», el primer article titulat *A note on the application of integrals involving cyclic products of kernels*, de V. Buldygin, F. Utzet i V. Zaiats, és un estudi de les funcions de moments, la funció de correlació i la funció espectral d'un procés estocàstic mitjançant una integral; els autors també estudien una desigualtat que permet analitzar la convergència de la integral i la distribució asimptòtica de la funció d'impuls del sistema de Volterra.

En el segon article, *Characterizations of inequality orderings by means of dispersive orderings*, de H. M. Ramos i M. A. Sordo, es comparen dos ordres, generalitzat i absolut, de Lorenz i es prova que són equivalents a dos ordres estocàstics. Els autors proposen una caracterització de l'ordre absolut i en fan una aplicació a l'ordenació de distribucions gamma; també es comenten les aplicacions a la distribució de rendes i d'indicadors de benestar social.

L'article *Statistical procedures for spatial point pattern recognition*, de J. Mateu, compara una metodologia basada en models i també en dissenys per tractar estructures espacials representades per un patró de punt, tant des del punt de vista teòric com a partir de dades obtingudes per simulació, i es conclou que la suposició d'un model paramètric és més eficient per reconèixer el patró de les dades que l'aplicació d'un model no paramètric.

El quart article, *Detección de rasgos en imágenes binarias mediante procesos puntuales espaciales marcados*, de J. Mateu i G. Lorenzo, estudia imatges difuses digitalitzades, presentant primer algunes tècniques per proposar, a continuació, una alternativa basada en processos puntuals marcats i mixtura de distribucions, que s'il·lustren amb simulacions i dades reals.

A continuació, *Estimación de la función de distribución sobre poblaciones finitas mediante diseños muestrales bietápicos apropiados*, de J. A. Mayor i M. Martinez, s'utilitza l'estimador de Horvitz-Thompson per estudiar una funció de distribució en poblacions finites i el mostratge en conglomerats en dues etapes, utilitzant una norma més apropiada com a criteri d'optimització, que redueix l'error d'estimació.

El sisè article, *Análisis factorial múltiple como técnica de estudio de la estabilidad de los resultados de un análisis de componentes principales*, d'E. Abascal i M. I. Landaluce, és una contribució metodològica a la naturalesa i l'estabilitat de l'anàlisi factorial múltiple, que les seves autores estudien proposant coeficients i comparant metodologies de diferents autors. En aquest context, les autores estudien l'estabilitat dels salaris de les diferents comunitats autònomes de l'Estat espanyol.

El darrer article de la secció, *Un modelo Poissoniano para predecir la matriculación de vehículos en países europeos*, de M. J. Valderrama, A. M. Aguilera i P. R. Bouzas, és una modelització doblement estocàstica d'un procés de Poisson, on la mitjana no és fixa sinó una distribució normal truncada, il·lustrada amb una predicció de la matriculació de vehicles a la Comunitat Europea.

L'únic article publicat a «Investigació Operativa», titulat *On superlinear multiplier update methods for partial augmented Lagrangian techniques*, d'E. Miganjos, tracta de la minimització de funcions no lineals amb restriccions i acotacions mitjançant una funció Lagrangiana augmentada, amb contribucions a l'estimació dels multiplicadors de Lagrange a còpia de combinar tècniques de reducció de variables amb mètodes de tipus Newton, provant els resultats amb diversos tests i comentant els avantatges de la tècnica proposada.

La secció «Estadística Oficial» publica tres articles relacionats amb l'estadística sobre el turisme i les estimacions sobre petites àrees elaborades actualment per l'Institut d'Estadística de Catalunya, que han estat objecte de recents avenços metodològics. D'una banda, en l'article *Alimentación de modelos cuantitativos con información subjetiva: aplicación Delphi en la elaboración de un modelo de imputación del gasto turístico individual en Catalunya*, de F. Landeta, J. Matey, V. Ruíz i O. Villareal, es presenta una original aplicació del mètode Delphi en un àmbit dominat preferentment per informació provinent de l'enquestació directa i/o d'aprofitament de registres administratius; d'altra banda, també s'il·lustra que les limitacions tradicionals d'aquesta tècnica es redueixen significativament per l'adopció de les noves TI en els processos de captura de dades.

En segon lloc, l'article *Smoothing the Catalan tourism micro-data time series*, de M. Artís, J. L. Carrión, À. Costa i J. Suriñach, exposa un innovadora aproximació a la recreació de dades de sèries temporals amb un elevat component de volatilitat com és el cas del moviment turístic, gràcies a tècniques d'allisament de sèries a través de mitjanes mòbils ponderades estacionals que s'apliquen a les microdades que s'han d'agregar.

El darrer article de la secció, *Estimadores compuestos en estadística regional: una aplicación a la estimación de la tasa de variación de la ocupación en la industria*, d'À. Costa, A. Satorra i E. Ventura, revela les possibilitats dels estimadors compostos en relació a l'estimació de petites àrees tan pròpia de l'estadística regional. Els autors ofereixen una detallada aplicació en l'estimació de les variàncies, on es demostra que la combinació lineal d'estimadors directes i indirectes suposa, generalment, l'opció més adient per a l'estimació d'aquest tipus de paràmetres.

Finalment, la secció «Biometria» acull l'edició de dos articles. En el primer dels articles, *Maintenance policy under unrevealed failures*, de F. G. Badía, M. D. Berrade i C. A. Campos, s'exposa el problema de les fallades de sistemes que són detectades no quan ocorren sinó quan s'enregistra una nova demanda, anàlisi que s'enfoca sota diferents possibilitats. En aquest context, els autors proposen una política de manteniment òptima, les característiques de la qual il·lustren suposant distribucions bivariants Gumbel i Marshall-Olkin.

El segon article, *Inverse sampling and triangular sequential design to compare small proportion with a reference value*, de V. Moreno, I. Martín, F. Torres, M. Horas, J. Ríos i J. R. González, proposa algunes formes de mostratge quan, en dissenys seqüencials, cal comparar una probabilitat petita o molt petita amb un valor conegut, amb especial referència al mostratge invers i al test triangular, que semblen necessitar mides de mostres inferiors a altres mètodes.

Comentari d'altres seccions i apartats

La «Secció Docent i Problemes» inclou, per darrera vegada, la presentació successiva de nous enunciats (la solució dels quals apareixerà en el darrer número d'aquest volum 26) i la resolució dels problemes publicats en el número anterior d'aquest volum. També es publiquen els darrers treballs de caràcter docent, amb la contribució de C. M. Cuadras sobre una original visió de la coneguda distribució de Cauchy des d'una perspectiva estrictament geomètrica i, d'altra banda, una aportació col·lectiva a l'entorn de l'aplicació d'eines ofimàtiques com el fulls de càclcul a la comprensió del comportament de les xarxes neuronals artificials, a càrrec de J. García, A. M. López, J. E. Romero, A. R. García, C. Camacho, J. L. Cantero, M. Atienza i R. Salas.

Seguidament, la secció «Comentaris de llibres» acull una detallada presentació de Joan del Castillo sobre *An Introduction to Statistical Modelling of Extreme Values*, de Stuart Coles, on es destaca l'esforç d'actualització de les aportacions recents a la teoria sobre valors externs i el pragmatisme de les seves prestacions pràctiques. En segon lloc, Tomàs Aluja presenta l'obra *The Elements of Statistical Learning. Data Mining, Inference and Prediction*, de T. Hastie, R. Tibshirani i J. Friedman, de la qual remarca l'encert en la unificació de la nomeclatura estadística a l'entorn del «machine learning», així com l'amplitud de la revisió feta sobre els models estadístics per a la generalització i la predicció de comportaments.

El darrer apartat, dedicat a «Ressenyes d'activitats institucionals», inclou, com ja és costum, una revisió actualitzada d'activitats de la Sociedad Española de Biometría, amb l'anunci de la IX Conferència Espanyola de Biometria que se celebrarà el maig del 2003. En segon lloc, s'ofereix una recensió del «Training for European Statisticians Institute» que *Qüestiió* redifon sistemàticament des del 1997, amb la relació actualitzada dels cursos del *Core Programme 2002* que s'impartiran des del gener fins al novembre del 2002, adreçats principalment als membres dels òrgans d'estadística oficial en l'àmbit comunitari. En tercer lloc, s'ofereix l'anunci del «Workshop on Survival Analysis», que, sota el títol de *Research and Training in Failure Time Methods in the New Millenim* (Barcelona, 12-14 juny 2002), la Universitat Politècnica de Catalunya organitza amb la col·laboració, entre d'altres, de l'Idescat. El número conclou amb la relació de novetats editorials de l'Institut d'Estadística de Catalunya i de la resta de la Generalitat de Catalunya que han estat publicades en el primer semestre d'enguany.

Carles Cuadras, director executiu

Enric Ripoll, editor executiu

Estadística

QÜESTIIÓ, vol. 26, 1-2, p. 3-14, 2002

A NOTE ON THE APPLICATION OF INTEGRALS INVOLVING CYCLIC PRODUCTS OF KERNELS

V. BULDYGIN[1]
F. UTZET[2]
V. ZAIATS[2,3]

In statistics of stochastic processes and random fields, a moment function or a cumulant of an estimate of either the correlation function or the spectral function can often contain an integral involving a cyclic product of kernels. We define and study this class of integrals and prove a Young-Hölder inequality. This inequality further enables us to study asymptotics of the above mentioned integrals in the situation where the kernels depend on a parameter. An application to the problem of estimation of the response function in a Volterra system is given.

Keywords: Integral involving a cyclic product of kernels, cumulant, Young-Hölder inequality, cross-correlogram, asymptotic normality

AMS Classification (MSC 2000): 62M10, 46N30, 62G20

[1] National Technical University of Ukraine. Kyïv Polytechnic Institute, UKRAINE.
[2] Universitat Autònoma de Barcelona, Catalonia, SPAIN.
[3] Universitat de Vic, Catalonia, SPAIN.

- Received May 2001.
- Accepted January 2002.

1. INTRODUCTION

The problem of identification in stochastic linear systems has been a matter of active research for the last four decades. One of the simplest models considers a «black box» which enjoys some input and gives a certain output. The input may be single or multiple, and one can choose between the assumption that it is perfect or noisy. The same applies to the output irrespective of the input. This variety of possibilities generates a great amount of models to be considered, and one still has a choice between a parametric or nonparametric framework. The scope of applications of these models is fairly extensive, ranging from signal processing and automatic control to econometrics (errors-in-variables models). For more details, see Schetzen (1980), Hannan and Deistler (1988), Söderström and Stoica (1989). Each of these books may be a good starting point for further bibliographic references. When only second-order statistics are involved, the question arises about the uniqueness of the solution of a parametric identification problem, see Green and Anderson (1986), Deistler and Anderson (1989). The use of higher order cumulant statistics leads to consistent estimates of the parameters, see Tugnait (1992), and also proves to be useful in nonparametric settings, see Akaike (1966).

Our interest in the topic was stimulated by the problem of estimation of the so-called impulse response function (also called the transfer function in the time domain) of an SISO (single-input single-output) system. One of the usual tools used for this estimation is the discrete-time cross-correlogram between the input and the output, and one of the basic methods applied to prove statistical properties of the estimate is Brillinger's method of cumulants, see Brillinger (1981). Our impression is that the application of this method would always lead us to a certain class of integrals. These integrals also appear in other nonparametric statistical problems, and the peculiarity of this kind of integrals is that they involve cyclic products of kernels, meaning that the internal structure of integrals is always the same. The ability to handle these integrals in general can give a clue to obtaining good upper bounds. If one switches to a spatial setting, the Rosenblatt approximation by quadratic forms is often applied, see Rosenblatt (1985). It is interesting that integrals involving cyclic products of kernels also appear in the cumulants of bilinear forms of Gaussian random vectors, challenging us even more.

Since the weak convergence of estimators is often proved using high-order cumulants (see, e. g., Zhang and Shaman (1991), Grimmett (1992), Brillinger (1996), Haberzettl (1997)), making a closer look at integrals involving cyclic products of kernels worthwhile. Earlier studies of integral representations of cumulants of second-order statistics of stationary stochastic processes were carried out by Lithuanian statisticians R. Bentkus (1972, 1976) and Statulevičius (1977). A later paper by R. Bentkus (1977) deals with cumulants of polynomial statistics. Other integral representations of cumulants of the periodogram of a homogeneous random field were considered by Guyon (1995), Benn and Kulperger (1998), Rosenblatt (2000). Interesting classes of multi-

ple stochastic integrals were introduced by Surgailis (1981), Engel (1982), Fox and Taqqu (1987).

Let us introduce the object we would like to focus on in what follows. Denote by $\mathbb{N}$ the set of positive integers, $\mathbb{Z}$ the set of all integers, $\mathbb{R}$ the set of real numbers, and let $\mathbb{N}_m := \{1,\ldots,m\}$ for $m \in \mathbb{N}$. Let $(\mathbb{T},\mu)$ be a measurable space endowed with a σ-finite measure μ. We define an integral involving a cyclic product of kernels as the following Lebesgue integral:

$$(1)\quad \hat{I}_n(K_1,\ldots,K_n;\varphi_1,\ldots,\varphi_n) := \int\cdots\int_{\mathbb{T}^n}\left[\prod_{p=1}^{n} K_p(t_p,t_{p+1})\varphi_p(t_p)\right]\mu(dt_1)\ldots\mu(dt_n)$$

where $t_{n+1} = t_1$. The functions $K_p := (K_p(s,t),\ s,t \in \mathbb{T})$, $p \in \mathbb{N}_n$, are called kernels. In general, both the kernels K_p, $p \in \mathbb{N}_n$, and the functions $\varphi_p := (\varphi_p(t),\ t \in \mathbb{T})$ are complex-valued.

The remaining part of this paper is organized as follows: Section 2 gives an overview of the situations where integrals (1) appear. Section 3 states some new results on these integrals: a Young–Hölder inequality (Theorem 1) and the convergence to zero of an integral depending on a parameter (Theorem 2). These results are applied to prove asymptotic normality of a nonparametric estimate of the impulse response function in a Volterra system (Theorem 3).

2. SOME MOTIVATING EXAMPLES

Integral representation of a cumulant of a set of bilinear forms of Gaussian random vectors

Assume that $m \in \mathbb{N}$, $n_j \in \mathbb{N}$, $j \in \mathbb{N}_m$, and let $\vec{X}_{j,1} := (X_{j,1}(k),\ k \in \mathbb{N}_{n_j})$, $\vec{X}_{j,2} := (X_{j,2}(k),\ k \in \mathbb{N}_{n_j})$, $j \in \mathbb{N}_m$, be real-valued zero-mean jointly Gaussian random vectors. Consider the following set of bilinear forms: $S_j := \sum_{k=1}^{n_j} a_j(k)X_{j,1}(k)X_{j,2}(k)$, $j \in \mathbb{N}_m$, where $a_j(k) \in \mathbb{R}$ are nonrandom numbers for all $k \in \mathbb{N}_{n_j}$, $j \in \mathbb{N}_m$. Consider also the joint simple cumulant (see the definition, for example, in Brillinger (1981), Section 2.3), denoted by cum $(S_1,\ldots,S_m)$, of the random variables $S_1,\ldots,S_m$. By Theorem 2.3.2 in Brillinger (1981) and by some algebra, one can show that

$$(2)\quad \operatorname{cum}(S_1,\ldots,S_m) = \sum_{(\vec{\jmath},\vec{\alpha})\in\{P,2\}_{m-1}}\ \sum_{k_1,\ldots,k_m=1}^{n_1,\ldots,n_m}\left[\prod_{p=1}^{m} a_{j_p}(k_{j_p})C_{j_p,j_{p+1}}^{\overline{\alpha}_p,\alpha_{p+1}}(k_{j_p},k_{j_{p+1}})\right]$$

where $\vec{\jmath} := (j_1,j_2,\ldots,j_m)$, $\vec{\alpha} := (\alpha_1,\alpha_2,\ldots,\alpha_m)$, with the convention that $j_1 = j_{m+1} = 1$, $\alpha_1 = \alpha_{m+1} = 2$, and where

$$C^{\overline{\alpha}_p,\alpha_{p+1}}_{j_p,j_{p+1}}(k_{j_p},k_{j_{p+1}}) := \mathsf{E}X_{j_p,\overline{\alpha}_p}(k_{j_p})X_{j_{p+1},\alpha_{p+1}}(k_{j_{p+1}}), \qquad \overline{\alpha}_k := \begin{cases} 2, & \text{if } \alpha_k = 1, \\ 1, & \text{if } \alpha_k = 2. \end{cases}$$

The notation $(\vec{\jmath},\vec{\alpha}) \in \{P,2\}_{m-1}$ applied to the vectors $\vec{\jmath} = (j_1, j_2, \ldots, j_m)$ and $\vec{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_m)$ is interpreted as follows:

$$j_1 = 1, \quad \alpha_1 = 2, \quad \text{and} \quad ((j_2,\ldots,j_m),(\alpha_2,\ldots,\alpha_m)) \in \mathrm{Perm}\{2,\ldots,m\} \times \{1,2\}^{m-1}$$

where $\mathrm{Perm}\{2,\ldots,m\}$ denotes the set of all permutations of $\{2,\ldots,m\}$. If $\mathrm{card}(A)$ is cardinality of the set A, then it is clear that $\mathrm{card}\left(\mathrm{Perm}\{2,\ldots,m\} \times \{1,2\}^{m-1}\right) = 2^{m-1} \cdot (m-1)!$

Let $j, j' \in \mathbb{N}_m$ and $\alpha, \alpha' \in \{1,2\}$. Assume further that there exists a Lebesgue integrable complex-valued function $G^{\alpha,\alpha'}_{j,j'}(u,v)$, $(u,v) \in \mathbb{R}^2$, such that

$$\mathsf{E}X_{j,\alpha}(k)X_{j',\alpha'}(k') = \int\int_{\mathbb{R}^2} \exp\{i(uk+u'k')\}G^{\alpha,\alpha'}_{j,j'}(u,u')du\,du'.$$

In this case (2) implies that

$$\text{(3)} \qquad \mathrm{cum}\,(S_1,\ldots,S_m) = \sum_{(\vec{\jmath},\vec{\alpha})\in\{P,2\}_{m-1}} \int\cdots\int_{\mathbb{R}^m} \left[\prod_{p=1}^{m} Q^{(\vec{\jmath},\vec{\alpha})}_p(v_p,v_{p+1})\right] dv_1 \ldots dv_m$$

where $v_{m+1} = v_1$ and

$$Q^{(\vec{\jmath},\vec{\alpha})}_p(s,u) := \int_{\mathbb{R}} G^{\overline{\alpha}_p,\alpha_{p+1}}_{j_p,j_{p+1}}(s,t)A_{j_{p+1}}(t,u)dt, \quad p \in \mathbb{N}_m;$$

$$A_j(t,u) := \sum_{k=1}^{n_j} a_j(k)\exp\{i(t+u)k\}, \quad j \in \mathbb{N}_m.$$

Each integral on the right-hand side of (3) involves a cyclic product of kernels and the cumulant $\mathrm{cum}\,(S_1,\ldots,S_m)$ is the sum of these integrals.

Let $j, j' \in \mathbb{N}_m$ and $\alpha, \alpha' \in \{1,2\}$. Assume that there exists a complex-valued measure $M^{\alpha,\alpha'}_{j,j'}$ on $([-\pi,\pi], \mathcal{B}([-\pi,\pi]))$ such that $\mathsf{E}X_{j,\alpha}(k)X_{j',\alpha'}(k') = \int_{-\pi}^{\pi} e^{i(k-k')\lambda} M^{\alpha,\alpha'}_{j,j'}(d\lambda)$. Here and in what follows, $\mathcal{B}(A)$ denotes the Borel σ-algebra on A. Then

$$\text{(4)} \qquad \mathrm{cum}\,(S_1,\ldots,S_m) = \sum_{(\vec{\jmath},\vec{\alpha})\in\{P,2\}_{m-1}} \int\cdots\int_{[-\pi,\pi]^m} \left[\prod_{p=1}^{m} A^{(\vec{\jmath},\vec{\alpha})}_p(\lambda_p,\lambda_{p+1})\right] \cdot$$

$$\cdot \mu^{(\vec{\jmath},\vec{\alpha})}_1(d\lambda_1)\ldots\mu^{(\vec{\jmath},\vec{\alpha})}_m(d\lambda_m)$$

where $\lambda_{m+1} = \lambda_1$ and $A_p^{(\vec{\jmath},\vec{\alpha})}(u,v) := \sum_{k=1}^{n_{jp}} a_{j_p}(k)\exp\{-ik(u-v)\}$, $p \in \mathbb{N}_m$; $\mu_p^{(\vec{\jmath},\vec{\alpha})} := M_{j_p,j_{p+1}}^{\alpha_p,\alpha_{p+1}}$, $p \in \mathbb{N}_m$. Suppose now that for any $(\vec{\jmath},\vec{\alpha})$ and any p we have $\mu_p^{(\vec{\jmath},\vec{\alpha})}(B) = \int_B f_p^{(\vec{\jmath},\vec{\alpha})}(\lambda)d\lambda$, $B \in \mathcal{B}([-\pi,\pi])$. Then

$$(5)\quad \operatorname{cum}(S_1,\dots,S_m) = \sum_{(\vec{\jmath},\vec{\alpha})\in\{P,2\}_{m-1}} \int\cdots\int_{[-\pi,\pi]^m} \left[\prod_{p=1}^{m} A_p^{(\vec{\jmath},\vec{\alpha})}(\lambda_p,\lambda_{p+1}) f_p^{(\vec{\jmath},\vec{\alpha})}(\lambda_p)\right] \cdot d\lambda_1\dots d\lambda_m$$

where $\lambda_{m+1} = \lambda_1$. Formulas (4) and (5) give further representations of the cumulant $\operatorname{cum}(S_1,\dots,S_m)$ as a finite sum of integrals involving cyclic products of kernels.

The book by Mathai and Provost (1992) is a good reference on quadratic forms, and the papers of Mathai (1992) and Holmquist (1996) focus on quadratic and bilinear forms in normal variables.

Inequalities for a cumulant of the cross-correlogram in a Gaussian time series

Let $Y(t), t \in \mathbb{Z}$, be a real-valued zero-mean stationary Gaussian process with covariance function $C(t) := \mathsf{E}Y(t)Y(0)$, $t \in \mathbb{Z}$, and spectral function $F(\lambda)$, $\lambda \in [-\pi,\pi]$. Consider the sample correlogram $\hat{C}(\tau;N) := \sum_{k=1}^{N} a(k;\ \tau,N)Y(\tau+k)Y(k)$, $\tau \in \mathbb{Z}$, $N \in \mathbb{N}$ where $a(k;\ \tau,N)$ are nonrandom real-valued weights for all $k \in \mathbb{N}_N$, $\tau \in \mathbb{Z}$, $N \in \mathbb{N}$. Formula (4) implies the following inequality:

$$(6)\ \left|\operatorname{cum}(\hat{C}(\tau_1;N),\dots,\hat{C}(\tau_m;N))\right|$$

$$\le 2^{m-1}(m-1)!\int\cdots\int_{[-\pi,\pi]^m}\left[\prod_{p=1}^{m}\left|A^{(N)}(\lambda_p-\lambda_{p+1};\ \tau_{p+1})\right|\right] dF(\lambda_1)\dots dF(\lambda_m)$$

where $\lambda_{m+1} = \lambda_1$ and $A^{(N)}(u;\tau) := \sum_{k=1}^{N} a(k;\ \tau,N)\exp\{-iku\}$, $\tau \in \mathbb{Z}$, $u \in [-\pi,\pi]$. If, for example, $a(k;\ \tau,N) \equiv 1/N$, then (6) implies that

$$\left|\operatorname{cum}(\hat{C}(\tau_1;N),\dots,\hat{C}(\tau_m;N))\right|$$

$$\le N^{-m}2^{m-1}(m-1)!\int\cdots\int_{[-\pi,\pi]^m}\left[\prod_{p=1}^{m}\left|\frac{\sin\left(\frac{N(\lambda_p-\lambda_{p+1})}{2}\right)}{\sin\left(\frac{\lambda_p-\lambda_{p+1}}{2}\right)}\right|\right] dF(\lambda_1)\dots dF(\lambda_m).$$

A representation carried out by R. Bentkus of a cumulant of a second-order spectral estimate in a stationary time series

Let $Y(t), t \in \mathbb{Z}$, be a weak sense stationary zero-mean real-valued time series. Consider the second-order spectral estimate:

$$\hat{E}^{(N)}(g_j) = \int_{-\pi}^{\pi} g_j(x) I^{(N)}(x)dx, \quad j \in \mathbb{N}_m, \quad m \in \mathbb{N},$$

where $g_j \in L_1[-\pi,\pi]$, $j \in \mathbb{N}_m$, and where

$$I^{(N)}(x) := (2\pi N)^{-1} \left| \sum_{s=1}^{N} Y(s) e^{-isx} \right|^2, \quad x \in [-\pi,\pi], \quad N \in \mathbb{N},$$

is the so-called periodogram (Brillinger (1981), Section 5.2). Assume that the process $Y(\cdot)$ has n-th order spectral density $\varphi_n\left(\lambda_1,\ldots,\lambda_{n-1},-\sum_{j=1}^{n-1}\lambda_j\right)$, $(\lambda_1,\ldots,\lambda_{n-1}) \in [-\pi,\pi]^{n-1}$, for any $n = 2,\ldots,2m$. Then a result by R. Bentkus (1977) after some algebra implies that

$$\operatorname{cum}(\hat{E}^{(N)}(g_1),\ldots,\hat{E}^{(N)}(g_m))$$
$$= \frac{2}{(2\pi)^m N} \int \cdots \int_{[-\pi,\pi]^{2m}} \prod_{p=1}^{2m} \left[\frac{\sin(N(\nu_p - \nu_{p+1})/2)}{\sin((\nu_p - \nu_{p+1})/2)} \right] G(\nu_1,\ldots,\nu_{2m}) d\nu_1 \ldots d\nu_{2m}$$

where $\nu_{2m+1} = \nu_{2m}$, and where the function $G(\cdot)$ is expressed in terms of $g_1(\cdot),\ldots,$ $g_m(\cdot)$ and $\varphi_1(\cdot),\ldots,\varphi_{2m}(\cdot)$. This formula shows that cum $(\hat{E}^{(N)}(g_1),\ldots,\hat{E}^{(N)}(g_m))$ is reduced to an integral involving a cyclic product of kernels.

Long-range dependence. The Rosenblatt distribution

Let $Y(t)$, $t \in \mathbb{Z}$, be a stationary Gaussian real-valued zero-mean stochastic process with covariance function $C(t), t \in \mathbb{Z}$. Assume that $C(0) = 1$ and $\lim_{|t|\to\infty} \alpha^{-1}|t|^{-\beta}C(t) = 1$, where $\alpha > 0$ and $\beta > 0$. If $\beta \in (0,1/2)$, then $Y(\cdot)$ is called a process with long-range dependence. Consider the sample correlogram $\hat{C}(\tau;N) = N^{-1}\sum_{k=1}^{N} Y(\tau+k)Y(k)$, $\tau \in \mathbb{Z}$, $N \in \mathbb{N}$. For a process with index $\alpha = 1$, the limit distribution as $N \to \infty$ of $N^{\beta}(\hat{C}(0;N) - 1)$ is non-normal, see Rosenblatt (1985), p. 72. This distribution is often called the Rosenblatt distribution. The characteristic function $\varphi(u), u \in \mathbb{R}$, of the Rosenblatt distribution has the following form: $\varphi(u) = \exp\{(1/2)\sum_{k=2}^{N}(2iu)^k \mathbf{I}_{\beta}(k)/k\}$, $u \in \mathbb{R}$, where

$$\mathbf{I}_{\beta}(k) = \int \cdots \int_{[0,1]^k} \left[\prod_{j=1}^{k} |x_j - x_{j+1}|^{-\beta} \right] dx_1 \ldots dx_n$$

for any $k \geq 2$ and where $x_{k+1} = x_1$. It is clear that each of the integrals $\mathbf{I}_{\beta}(k)$, $k \geq 2$, involves a cyclic product of kernels.

3. SOME RESULTS ON INTEGRALS INVOLVING CYCLIC PRODUCTS OF KERNELS

The Young–Hölder inequality for an integral involving a cyclic product of kernels

For $K = (K(s,t),\ s,t \in \mathbb{T})$ we put $\|K\|_{\infty,p} := \mu\text{-ess sup}_{s\in\mathbb{T}} \|K(s,\cdot)\|_p$, where $\|\cdot\|_p$ is the standard norm in $L_p(\mathbb{T})$, $p \in [1,\infty]$. We also introduce another norm: $|||K|||_p := \max\{\|K\|_{\infty,p}, \|K'\|_{\infty,p}\}$ where K' is the dual function of K, that is $K'(s,t) := K(t,s)$ for any $(s,t) \in \mathbb{T}^2$. Put $L_p^{\infty}(\mathbb{T}^2) := \{K:\ |||K|||_p < \infty\}$. An inequality having the form

$$|\hat{I}(K_1,\dots,K_n;\ \varphi_1,\dots,\varphi_n)| \le \prod_{j=1}^{n} (|||K_j|||_{p_j} \cdot \|\varphi_j\|_{q_j}) \tag{7}$$

with $1 \le p_j, q_j \le \infty$, $j \in \mathbb{N}_n$, will be called a Young inequality for integral (1). If each q_j is the real conjugate exponent of p_j, $j \in \mathbb{N}_n$, that is $(1/p_j) + (1/q_j) = 1$ for all $j \in \mathbb{N}_n$, then (7) is called a Young–Hölder inequality for integral (1).

Theorem 1. *Let $n \in \mathbb{N}$, $n \ge 2$. Assume that $1 \le q_1,\dots,q_n \le \infty$ and*

$$\max_{j_1 \ne j_2} ((1/q_{j_1}) + (1/q_{j_2})) \ge 1. \tag{8}$$

If $K_j \in L_{p_j}^{\infty}(\mathbb{T}^2)$, $\varphi_j \in L_{q_j}(\mathbb{T})$ for each $j \in \mathbb{N}_n$, where $p_j^{-1} + q_j^{-1} = 1$, $j \in \mathbb{N}_n$, then

$$\left|\hat{I}_n(K_1,\dots,K_n;\ \varphi_1,\dots,\varphi_n)\right| \le \prod_{j=1}^{n} |||K_j|||_{p_j} \|\varphi_j\|_{q_j}. \tag{9}$$

Sketch of the proof

Let

$$j_1 \in \operatorname{Argmax}\left(\frac{1}{q_j},\ j \in \mathbb{N}_n\right), \qquad j_2 \in \operatorname{Argmax}\left(\frac{1}{q_j},\ j \ne j_1\right). \tag{10}$$

Then we obtain by (8)

$$\frac{1}{q_{j_1}} + \frac{1}{q_{j_2}} \ge 1. \tag{11}$$

Put

$$p_j = q_j',\quad j \in \mathbb{N}_n, \tag{12}$$

and consider the following collections:

$$(\vec{p}_{j_1,j_2-1},\vec{q}_{j_1+1,j_2-1}) = (q'_{j_1}q_{j_1+1}q'_{j_1+1}\cdots q_{j_2-1}q'_{j_2-1}),$$
$$(\vec{p}_{j_2,j_1-1},\vec{q}_{j_2+1,j_1-1}) = (q'_{j_2}q_{j_2+1}q'_{j_2+1}\cdots q_{j_2-1}q'_{j_1-1}).$$

By (10) and (12), we have

$$\frac{1}{p_{j_1}} = \frac{1}{q'_{j_1}} = 1 - \frac{1}{q_{j_1}} = \min_{j\in\mathbb{N}_n} \frac{1}{p_j}$$

and

$$\frac{1}{p_{j_2}} = \frac{1}{q'_{j_2}} = 1 - \frac{1}{q_{j_2}} = \min_{j\in[j_2,j_1-1]} \frac{1}{p_j},$$

that is

$$p_{j_1} = \max_{j\in[j_1,j_2-1]} p_j, \qquad p_{j_2} = \max_{j\in[j_2,j_1-1]} p_j. \tag{13}$$

Furthermore

$$\frac{1}{q_j} + \frac{1}{p_j} = \frac{1}{q_j} + \frac{1}{q'_j} = 1, \qquad j \in \mathbb{N}_n, \tag{14}$$

and moreover

$$\left(\Sigma_{j\in[j_1,j_2-1]} \frac{1}{p_j}\right) + \left(\Sigma_{j\in[j_1+1,j_2-1]} \frac{1}{q_j}\right) = \frac{1}{q'_{j_1}} + \Sigma_{j\in[j_1+1,j_2-1]} \left(\frac{1}{q_j} + \frac{1}{q'_j}\right) =$$
$$= \frac{1}{q'_{j_1}} + (|j_2 - j_1| - 1).$$

Therefore

$$\frac{1}{a_{j_1,j_2-1}} = \frac{1}{q'_{j_1}} \in [0,1]. \tag{15}$$

In a similar manner, we can prove that

$$\frac{1}{a_{j_2,j_1-1}} = \frac{1}{q'_{j_2}} \in [0,1]. \tag{16}$$

We also have

$$\left(\sum_{j=1}^{n} \frac{1}{p_j}\right) + \left(\sum_{j=1}^{n} \frac{1}{q_j}\right) = \sum_{j=1}^{n} \left(\frac{1}{q'_j} + \frac{1}{q_j}\right) = n. \tag{17}$$

Now Theorem 1 can be obtained from Theorem 9.5.1 in Edwards (1965), and formulas (11), (13)–(17) using the techniques employed in Section 5 of Buldygin *et al.* (2000). □

Corollary 1. *Let $n \geq 2$. Assume that $1 \leq q_1, \ldots, q_n \leq \infty$ and* (8) *holds. Then there exist numbers $p_j \in (1, \infty]$, $j \in \mathbb{N}_n$, satisfying $\sum_{j=1}^{n}(p_j^{-1} + q_j^{-1}) = n$ and such that* (9) *holds whenever $K_j \in L_{p_j}(\mathbb{T}^2)$ for each $j \in \mathbb{N}_n$.*

Convergence to zero of an integral involving a cyclic product of kernels, with the kernels depending on a parameter

Assume that the kernels $K_1, \ldots, K_n$ appearing in the integrals $\hat{I}_n$ depend on a parameter σ, that is $K_j = K_j^{(\sigma)}$ for all $j \in \mathbb{N}_n$, $\sigma \in \mathfrak{S}$. We then write $\hat{I}_n^{(\sigma)} := \hat{I}(K_1^{(\sigma)}, \ldots, K_n^{(\sigma)}; \varphi_1, \ldots, \varphi_n)$. Consider the following majorant condition: *for any $p \in (1, \infty]$, there exists a constant $C_p \geq 0$ not depending on $\sigma \in \mathfrak{S}$ such that $\max_j |||K_j^{(\sigma)}|||_p \leq C_p[\rho(\sigma)]^{1/2-1/p}$ where $\rho(\sigma) > 0$, $\sigma \in \mathfrak{S}$.*

Theorem 2. *Let $n \in \mathbb{N}$, $n \geq 3$. Assume that:* (i) *the kernels $K_j^{(\sigma)}$, $j \in \mathbb{N}_n$, satisfy the majorant condition;* (ii) *among the functions $\varphi_1, \ldots, \varphi_n$, there exist $n_1 \geq 0$ functions belonging to the space $L_1(\mathbb{R}^d)$, $n_\infty = n_1$ functions belonging to the space $L_\infty(\mathbb{R}^d)$, and $n_2 = n - 2n_1 \geq 0$ functions belonging to the space $L_2(\mathbb{R}^d)$. Then $\lim_{\rho(\sigma)\to\infty} I_n^{(\sigma)} = 0$.*

The proof of Theorem 2 follows the lines of that of Part B) of Theorem 5.3 in Buldygin *et al.* (2000) with slight modifications related to the form of the majorant condition above.

Asymptotic normality of a cross-correlogram estimate of the response function in an SISO system

Take a time-invariant continuous SISO (single-input single-output) linear system with a real-valued impulse response function $H := (H(\tau), \tau \in \mathbb{R})$, also called the transfer function in the time domain. This means that the system is an input-output type «black box», and the response of the system that enjoys an input $x(t)$, $t \in \mathbb{R}$, has the following form: $y(t) = \int_{-\infty}^{\infty} H(t-s)x(s)ds$, where the function $H(\cdot)$ is unknown. The problem is to estimate H from observations after the input and the output interpreted as stochastic processes X and Y, respectively. Suppose that a family of continuous spectral densities f_Δ, $\Delta > 0$, satisfies conditions (1a)–(1g) in Buldygin *et al.* (1998). Let $X_\Delta := (X_\Delta(t), t \in \mathbb{R})$, $\Delta > 0$, be a family of separable stationary zero-mean Gaussian processes with spectral

densities $f_{X_\Delta} = f_\Delta$, $\Delta > 0$. Consider a stochastic process Y_Δ representing the response of the system to the input X_Δ. Define the sample cross-correlogram:

$$\hat{H}_{\Delta,N}(\tau) := \frac{1}{N} \sum_{n=1}^{N} Y_\Delta(nh(\Delta)) X_\Delta(nh(\Delta) - \tau)$$

where $h(\Delta) := \pi/(2\lambda_0(\Delta))$, $\lambda_0(\Delta) := \sup\{\lambda > 0 : \ f_\Delta(\lambda) > 0\} < \infty$, and $\lambda_0(\Delta) \to \infty$ as $\Delta \to \infty$, see condition (1b) in Buldygin *et al.* (1998). By (6) and Theorem 2, the following statement can be proved.

Theorem 3. *If $H \in L_2(\mathbb{R})$, then for any $n \geq 1$ and all $\tau_1, \ldots, \tau_n \in \mathbb{R}$ one has*

$$\mathsf{E}\left[\prod_{j=1}^{n} \hat{Z}_{\Delta,N}(\tau_j)\right] \to \mathsf{E}\left[\prod_{j=1}^{n} Z(\tau_j)\right] \quad \text{as } (\Delta, Nh(\Delta)) \to \infty.$$

In particular, all finite-dimensional distributions of the process $\hat{Z}_{\Delta,N} = (\hat{Z}_{\Delta,N}(\tau), \tau \in \mathbb{R})$ converge weakly to the corresponding finite-dimensional distributions of the Gaussian process Z. Here, $\hat{Z}_{\Delta,N}(\tau) := \sqrt{Nh(\Delta)}\,[\hat{H}_{\Delta,N}(\tau) - \mathsf{E}\hat{H}_{\Delta,N}(\tau)]$, $Z = (Z(\tau), \ \tau \in \mathbb{R})$ is a zero-mean Gaussian process whose correlation function is

$$\mathsf{E}Z(\tau_1)Z(\tau_2) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \left[e^{i(t_1 - t_2)\lambda} |H^*(\lambda)|^2 + e^{i(t_1 + t_2)\lambda} (H^*(\lambda))^2\right] d\lambda, \quad \tau_1, \tau_2 \in \mathbb{R},$$

and H^* is the L_2 Fourier transform of H.

The proof presents no technical difficulties and can be obtained similarly to the proof of Theorem 4.1 in Buldygin *et al.* (2000) with the reference to Part B) of Theorem 5.3 replaced by that to the above stated Theorem 2.

ACKNOWLEDGEMENTS

The authors thank an anonymous referee for helpful comments. The first author acknowledges support from the Ministry of Education and Culture of Spain under the grant SAB1998–0169 for a sabbatical stay at the Universitat Autònoma de Barcelona. The second author was supported by DGICYT under the grant BFM2000-0009 and by CIRIT under the grant 2001SGR 00174. The third author was partially supported by DGICYT under the grant BFM2000–0002 and by the Universitat de Vic.

REFERENCES

Akaike, H. (1966). «On the use of non-Gaussian process in the identification of a linear dynamic system», *Ann. Inst. Statist. Math.*, 16, 269-276.

Benn, A. G. and Kulperger, R. J. (1998). «A remark on a Fourier bounding method of proof for convergence of sums of periodograms», *Ann. Inst. Statist. Math.*, 50, 1, 187-202.

Bentkus, R. (1972). «The asymptotic normality of an estimate of the spectral function», *Litovsk. Mat. Sb.*, 12, 3, 5-18.

Bentkus, R. (1976). «Cumulants of estimates of the spectrum of a stationary sequence», *Lithuanian Math. J.*, 16, 4, 501-518.

Bentkus, R. (1977). «Cumulants of multilinear forms of a stationary sequence», *Litovsk. Mat. Sb.*, 17, 1, 27-46.

Brillinger, D. R. (1981). *Time series: Data analysis and theory*. San Francisco: Holden-Day.

Brillinger, D. R. (1996). «Some uses of cumulants in wavelet analysis», *J. Nonpatametr. Statist.*, 6, 2-3, 93-114.

Buldygin, V., Utzet, F. and Zaiats, V. (1998*b*). *Preprint no. 8/1998*, Barcelona: Departament de Matemàtiques, Universitat Autònoma de Barcelona.

Buldygin, V., Utzet, F. and Zaiats, V. (2000). *Preprint no. 12/2000*, Barcelona: Departament de Matemàtiques, Universitat Autònoma de Barcelona.

Deistler, M. and Anderson, B. D. O. (1989). «Linear dynamic errors-in-variables models: some structure theory», *J. Econometrics*, 41, 39-63.

Edwards, R. E. (1965). *Functional analysis. Theory and applications*. New York: Holt, Rinehart and Winston.

Engel, D. D. (1982). «The multiple stochastic integral». *Mem. Amer. Math. Soc.*, 265.

Fox, R. and Taqqu, M. S. (1987). «Multiple stochastic integrals with dependent integrators». *J. Multivar. Anal.*, 21, 105-127.

Green, M. and Anderson, B. D. O. (1986). «Identification of multivariable errors-in-variables models with dynamics», *IEEE Trans. Automat. Control*, AC-31, 467-471.

Grimmett, G. (1992). «Weak convergence using high-order cumulants», *J. Theoret. Probab.*, 5, 4, 767-773.

Guyon, X. (1995). *Random fields on a network. Modeling, statistics, and applications*. New York: Springer-Verlag.

Haberzettl, U. (1997). «On estimating linear functionals of the covariance function of a stationary process», *Stochastic Anal. Appl.*, 15, 5, 759-782.

Hannan, E. J. and Deistler, M. (1988). *The statistical theory of linear systems*. New York: John Wiley.

Holmquist, B. (1996). «Expectations of products of quadratic forms in normal variables», *Stochastic Anal. Appl.*, 14, 2, 149-164.

Mathai, A. M. (1992). «On bilinear forms in normal variables», *Ann. Inst. Math. Statist.*, 44, 4, 769-779.

Mathai, A. M. and Provost, S. B. (1992). *Quadratic forms in random variables*. New York: Marcel Dekker.

Rosenblatt, M. (1985). *Stationary sequences and random fields*. Boston: Birkhäuser.

Rosenblatt, M. (2000). *Gaussian and non-Gaussian linear time series and random fields*. New York: Springer.

Schetzen, M. (1980). *The Volterra and Wiener theories of nonlinear systems*. New York: John Wiley.

Söderström, T. and Stoica, P. (1989). *System identification*. London: Prentice-Hall.

Statulevičius, V. A. (1977). «Application of semi-invariants to asymptotic analysis of distributions of random processes». In: *Multivariate Analysis*–IV (Krishnaiah, P. R., ed.) Amsterdam: North-Holland, 325-337.

Surgailis, D. (1984). «On multiple Poisson stochastic integrals and associated Markov semigroups». *Probab. Math. Statist.*, 3, 2, 217-239.

Tugnait, J. K. (1992). «Stochastic system identification with noisy input using cumulant statistics», *IEEE Trans. Automat. Control*, AC-37, 4, 476-485.

Zhang, Hong-Ching and Shaman, P. (1991). «On the calculation of cumulants of estimators arising from a linear time series regression model», *J. Multivar. Anal.*, 37, 2, 135-150.

QÜESTIIÓ, vol. 26, 1-2, p. 15-28, 2002

CHARACTERIZATIONS OF INEQUALITY ORDERINGS BY MEANS OF DISPERSIVE ORDERINGS

H. M. RAMOS ROMERO
M. A. SORDO DÍAZ
Universidad de Cádiz*

The generalized Lorenz order and the absolute Lorenz order are used in economics to compare income distributions in terms of social welfare. In Section 2, we show that these orders are equivalent to two stochastic orders, the concave order and the dilation order, which are used to compare the dispersion of probability distributions. In Section 3, a sufficient condition for the absolute Lorenz order, which is often easy to verify in practice, is presented. This condition is applied in Section 4 to the ordering of generalized gamma distributions with different parameters.

Keywords: Generalized Lorenz order, absolute Lorenz order, concave order, dilation order

AMS Classification (MSC 2000): 60E10, 60E15

* Departamento de Estadística e I. O. Universidad de Cádiz. Duque de Nájera, 8. 11002 Cádiz. Spain. Corresponding author: Tel: 34-956015406; fax: 34-956-015385; e-mail: hector.ramos@uca.es.

- Received October 2000.
- Accepted November 2001.

1. INTRODUCTION

The basic concepts of inequality and dispersion arise in many and diverse fields, so it is difficult to give brief definitions that will command universal acceptance. Loosely speaking, inequality in income distributions is seen as a particular aspect of variability when the variables considered are non-negative and represent quantities that can be transferred from one unit to another. In economics, inequality is usually used in connection with concepts such as injustice or social welfare. Champernowne and Cowell (1998) provide a convenient reference on this topic.

Several authors have approached the problem of ranking income distributions by seeking a dominance relationship between concentration curves. In this context, the generalized Lorenz curve and the absolute Lorenz curve have been used to compare two income distributions, in terms of social welfare and inequality. One of the purposes of this paper is to show that the partial orderings of income distributions induced by such curves are equivalent to two other stochastic orders used to compare probability distributions in terms of dispersion: the concave order and the dilation order, respectively (the definition of these orders is given below). These results are stated in Section 2, and some aspects of economic inequality and the usual concept of dispersion are connected.

The definition of the generalized Lorenz curve $GL_X(p)$ corresponding to the non-negative random variable X, which represents the income of a society or community, with distribution function $F(x)$ is (Shorrocks, 1983):

$$GL_X(p) = \int_0^p F_X^{-1}(t)\,dt, \quad p \in [0,1],$$

where F_X^{-1} denotes the inverse of F_X :

$$F_X^{-1}(a) = \inf\{x : F_X(x) \geq a\}, \; a \in [0,1].$$

The generalized Lorenz curve can be used to define a partial ordering (denoted $\leq_{gl}$) on the class of non-negative random variables as follows:

$$X \leq_{gl} Y \text{ if and only if } GL_X(p) \geq GL_Y(p) \text{ for all } p \in [0,1].$$

Then, we say that X exhibits more welfare in the generalized Lorenz sense than Y. Generalized Lorenz ordering reflects a desire for both greater equality and higher incomes. Some recent results on this ordering can be found in Ramos *et al.* (2000).

The welfare judgements embodied in the generalized Lorenz ordering are not universally accepted (see, e.g., Kolm, 1976). If one assumes an alternative concept of «efficiency preference», which corresponds to a preference for higher incomes, while keeping the same absolute differences between incomes, then absolute Lorenz ordering obtains (Moyes, 1987). For a non-negative random variable X with finite mean μ_X,

let $X-\mu_X$ be the mean-centred distribution obtained from X, and denote $F_{X-\mu_X}$ as its distribution function. The absolute Lorenz curve corresponding to X (Moyes, 1987) is defined as:

$$LA_X(p) = \int_0^p F^{-1}_{X-\mu_X}(t)\,dt, \quad p \in [0,1]. \tag{1}$$

$LA_X(p)$ represents the average income short-fall of the $100p\%$ poorest individuals, i.e., the average income that would be necessary in order to provide to anyone of them the society's mean income. The absolute Lorenz curve induces a partial ordering (denoted $\leq_{la}$) on the class of non-negative random variables, as follows:

$$X \leq_{la} Y \text{ if and only if } LA_X(p) \geq LA_Y(p) \text{ for all } p \in [0,1].$$

If $X \leq_{la} Y$, then X is said to exhibit less inequality in the absolute Lorenz sense than Y.

Although, for any finite population, there is no problem evaluating the absolute and generalized Lorenz curves, for a continuous distribution an analytic expression for these curves is rarely available. Ramos *et al.* (2000) gave a sufficient condition for the generalized Lorenz order that does not involve the explicit form of the generalized Lorenz curve. In Section 3, we complete the study they began by obtaining sufficient conditions for the absolute Lorenz ordering of random variables. These do not require a direct comparison of the involved absolute Lorenz curves. These results are applied in Section 4 to ordering of generalized gamma distributions with different parameters.

In the literature there are many partial orderings of probability distributions (e.g. Lewis and Thompson, 1981; Stoyan, 1983; Hickey, 1986). Some of them are defined by requiring

$$E[\Phi(X)] \leq E[\Phi(Y)] \tag{2}$$

to hold for all functions Φ in some class of functions. The concave order (Stoyan, 1983) is defined by saying that X is smaller than Y in the sense of the concave order (denoted by $X \leq_{cv} Y$) if (2) holds for all non-decreasing and concave functions Φ for which these expectations exist. The concave order has both properties of ordering by size and/or variability: if $X \leq_{cv} Y$ then X is both smaller and/or more variable than Y in some stochastic sense (see Chapter 3 in Shaked and Shanthikumar, 1994). It is known (see Stoyan, 1983) that

$$X \leq_{cv} Y \iff \int_{-\infty}^{x} F_X(t)\,dt \geq \int_{-\infty}^{x} F_Y(t)\,dt \text{ for all } x \in \mathbb{R} \tag{3}$$

provided the integrals exist.

Now let X and Y be random variables with finite means μ_X and μ_Y, respectively. Following Hickey (1986), we say that the random variable X is smaller than Y, in the dilation

sense, (denoted by $X \leq_{dil} Y$) if

$$E[\Phi(X-\mu_X)] \leq E[\Phi(Y-\mu_Y)] \tag{4}$$

for all convex functions Φ, provided these expectations exist. Clearly, dilation generalizes the use of variance to compare distributions in terms of dispersion.

For non-negative random variables, we show in Section 2 that X exhibits less welfare (*inequality*) than Y in the generalized (*absolute*) Lorenz sense if and only if X is smaller than Y in the concave (*dilation*) sense.

Several authors (Shorrocks, 1983; Lambert, 1993; Yitzhaki, 1999) have studied connections between generalized Lorenz order and increasing concave functions. Actually, in these papers, the authors restricted themselves to discrete distributions (Shorrocks, 1983) or absolutely continuous distributions having finite support $[a,b]$, $0 \leq a < b < \infty$ (Lambert, 1993; Yitzhaki, 1999). Nevertheless, from the result of Section 2 it follows that these restrictions are not needed. Our measure-theoretic approach permit us to handle, in one framework, both discrete and continuous distibutions, as well as combinations thereof.

The characterizations are obtained as an easy consequence of the theory of submajorization, as applied to decreasing rearrangements of functions. This notion has been discussed by several authors (see, e.g., Hardy *et al.*, 1929; Ryff, 1963; Chong, 1974). We first recall some definitions and results. Denote by $M(\Omega,\mu)$ the set of all extended real-valued measurable functions on a measure space (Ω,Λ,μ). For each $f \in M(\Omega,\mu)$ consider the function $D_f : \overline{\mathbb{R}} \longrightarrow [0,\mu(\Omega)]$ defined by

$$D_f(t) = \mu(\{x : f(x) > t\}), \, t \in \overline{\mathbb{R}}, \tag{5}$$

with $\overline{\mathbb{R}}$ denoting the extended real line. The decreasing rearrangement of f is defined by

$$\delta_f(t) = \inf\{s \in \mathbb{R} : D_f(s) \leq t\}, \, t \in [0,\mu(\Omega)].$$

Let $f,g \in M(\Omega,\mu) \cup M(\Omega',\mu')$, where $\mu(\Omega) = \mu'(\Omega') = a < \infty$ and denote by m the Lebesgue measure on $\mathbb{R}$. Then, we write $f \ll g$ whenever

$$\int_0^t \delta_f dm \leq \int_0^t \delta_g dm \text{ for all } t \in [0,a]$$

and $f \prec g$ whenever $f \ll g$ and

$$\int_0^a \delta_f dm = \int_0^a \delta_g dm.$$

If a is infinite, then the order relations $\prec$ and $\ll$ are defined for non-negative integrable functions $f,g \in L^1(\Omega,\mu) \cup L^1(\Omega',\mu')$ analogously.

We have the following results from Chong (1974).

Theorem 1. *$f \ll g$ if and only if $\int_u^{+\infty} D_f dm \leq \int_u^{+\infty} D_g dm$ for all $u \in \mathbb{R}$.*

Theorem 2. *If $f \in L^1(\Omega,\mu)$, $g \in L^1(\Omega',\mu')$ where $\mu(\Omega)$ and $\mu'(\Omega')$ are finite and equal, then $f \prec g$ if and only if*

$$\int_\Omega \Phi(f) d\mu \leq \int_{\Omega'} \Phi(g) d\mu'$$

for all convex functions $\Phi : \mathbb{R} \longmapsto \mathbb{R}$.

2. THE CHARACTERIZATIONS

Theorem 3. *Let X and Y be non-negative random variables with finite means μ_X and μ_Y, respectively. Then*

(i) $X \geq_{gl} Y$ if and only if $X \leq_{cv} Y$

(ii) $X \leq_{la} Y$ if and only if $X \leq_{dil} Y$.

Proof

(i) Let F_X and F_Y be the distribution functions of X and Y, respectively, and let $(\Omega, \mathcal{B}, m)$ be the measure space defined by $\Omega = \mathbb{R}^+$, $\mathcal{B}$ the Borel algebra of Ω and m the Lebesgue measure. Define $a(t) = 1 - F_X(t)$ and $b(t) = 1 - F_Y(t)$ for all $t \in \mathbb{R}^+$. Then $a(t)$ and $b(t)$ are non-increasing integrable functions and are equal to their respective decreasing rearrangements. It is easy to see that

$$D_a(t) = m\{x \in \mathbb{R}^+ : a(x) > t\} = F_X^{-1}(1-t), \text{ for all } t \in [0,1]$$

and, analogously,

$$D_b(t) = F_Y^{-1}(1-t) \text{ for all } t \in [0,1].$$

From Theorem 1 it follows that

$$\int_0^x (1 - F_X(t))\, dt \leq \int_0^x (1 - F_Y(t))\, dt \text{ for all } x > 0 \tag{6}$$

if and only if

$$\int_u^1 F_X^{-1}(1-t)\, dt \leq \int_u^1 F_Y^{-1}(1-t)\, dt \text{ for all } u \in [0,1]. \tag{7}$$

Clearly, (6) holds if and only if

$$\int_0^x F_X(t)\,dt \geq \int_0^x F_Y(t)\,dt \text{ for all } x > 0,$$

which is $X \leq_{cv} Y$ from (3). Now, by a change of variable it is seen that (7) is equivalent to

$$\int_0^u F_X^{-1}(t)\,dt \leq \int_0^u F_Y^{-1}(t)\,dt \text{ for all } u \in [0,1],$$

which is $X \geq_{gl} Y$ and the result holds.

(ii) Now let $(\Omega_X, \mathcal{B}_X, P_X)$ and $(\Omega_Y, \mathcal{B}_Y, P_Y)$ be the probability spaces on which X and Y, respectively, are defined. Define $a(\omega) = X(\omega) - \mu_X$ for all $\omega \in \Omega_X$ and $b(\omega) = Y(\omega) - \mu_Y$ for all $\omega \in \Omega_Y$. Then

$$D_a(t) = P_X\{\omega \in \Omega_X : a(\omega) > t\} = 1 - F_{X-\mu_X}(t), \text{ for all } t \in \mathbb{R},$$

and, analogously,

$$D_b(t) = 1 - F_{Y-\mu_Y}(t), \text{ for all } t \in \mathbb{R}.$$

The decreasing rearrangements of a and b are given, respectively, by $\delta_a(t) = F_{X-\mu_X}^{-1}(1-t)$ and $\delta_b(t) = F_{Y-\mu_Y}^{-1}(1-t)$, for all $t \in [0,1]$. From Theorem 2 it follows that

$$\int_0^u F_{X-\mu_X}^{-1}(1-t)\,dt \leq \int_0^u F_{Y-\mu_Y}^{-1}(1-t)\,dt \text{ for all } u \in [0,1] \tag{8}$$

if and only if (4) holds for all convex functions $\Phi : \mathbb{R} \longmapsto \mathbb{R}$, which is $X \leq_{dil} Y$. Since $X - \mu_X$ and $Y - \mu_Y$ have the same mean, it is not hard to see that (8) can be written as

$$\int_0^{1-u} F_{X-\mu_X}^{-1}(t)\,dt \geq \int_0^{1-u} F_{Y-\mu_Y}^{-1}(t)\,dt \text{ for all } u \in [0,1] \tag{9}$$

which is $X \leq_{la} Y$. Hence, the proof is complete. □

3. SUFFICIENT CONDITIONS FOR ABSOLUTE LORENZ ORDERING

Let X and Y be non-negative random variables with respective means μ_X and μ_Y, having continuous distribution functions F and G, respectively. The following theorem provides a sufficient condition for X and Y to be ordered in the absolute Lorenz sense, by means of a «single-crossing property» on the distribution functions of the random variables $X - \mu_X$ and $Y - \mu_Y$.

Theorem 4. *Suppose that $F(x+\mu_X) - G(x+\mu_Y)$ has at most one sign change (from $-$ to $+$) on $\mathbb{R}$. Then $X \leq_{la} Y$.*

Proof

Let $F_{X-\mu_X}$ and $G_{Y-\mu_Y}$ be the distribution functions of $X-\mu_X$ and $Y-\mu_Y$, respectively, defined by

$$F_{X-\mu_X}(x) = F(x+\mu_X) \text{ and } G_{Y-\mu_Y}(x) = G(x+\mu_Y).$$

Since the difference $F_{X-\mu_X} - G_{Y-\mu_Y}$ changes sign at most once with sequence $-,+$, by the assumptions on F and G, then the difference between the corresponding inverse distribution functions $F^{-1}_{X-\mu_X} - G^{-1}_{Y-\mu_Y}$ changes sign at most once with sequence $+,-$. Since $X-\mu_X$ and $Y-\mu_Y$ have the same mean, we have that

$$\int_0^1 F^{-1}_{X-\mu_X}(t)dt = \int_0^1 G^{-1}_{Y-\mu_Y}(t)dt$$

and it follows that

$$\int_0^p \left[F^{-1}_{X-\mu_X}(t) - G^{-1}_{Y-\mu_Y}(t)\right] dt \geq \int_0^1 \left[F^{-1}_{X-\mu_X}(t) - G^{-1}_{Y-\mu_Y}(t)\right] dt = 0$$

for all p in $[0,1]$. Hence, $LA_X(p) \geq LA_Y(p)$ holds for all p in $[0,1]$ and, consequently, $X \leq_{la} Y$. □

Suppose now that X and Y are absolutely continuous random variables with density functions f and g, respectively. Let $f_{X-\mu_X}$ and $g_{Y-\mu_Y}$ respectively denote the density functions of the random variables $X-\mu_X$ and $Y-\mu_Y$. The next result provides a convenient sufficient condition for the absolute Lorenz comparison of two random variables.

Corollary 1. *Assume that* $\text{supp}(X-\mu_X) \subseteq \text{supp}(Y-\mu_Y)$. *If* $f(x+\mu_X)/g(x+\mu_Y)$ *is unimodal for* x *restricted to* $\text{supp}(Y-\mu_Y)$, *where the mode is a supremum, then* $X \leq_{la} Y$.

Proof

Let $S(h)$ be the number of sign changes of the function $h(t)$. Since

$$f_{X-\mu_X}(x) = f(\mu_X + x), \quad g_{Y-\mu_Y}(x) = g(\mu_Y + x),$$

and $f(x+\mu_X)/g(x+\mu_Y)$ is unimodal on $\text{supp}(Y-\mu_Y)$, so is $f_{X-\mu_X}(x)/g_{Y-\mu_Y}(x)$, with the mode yielding a supremum. Hence

$$S\left(f_{X-\mu_X} - g_{Y-\mu_Y}\right) = S\left(\frac{f_{X-\mu_X}}{g_{Y-\mu_Y}} - 1\right) \leq 2$$

and the sign sequence is $-,+,-$, in the case of equality. This condition implies that the difference $F_{X-\mu_X} - G_{Y-\mu_Y}$ changes sign at most once with sequence $-,+$. From Theorem 4 it follows that $X \leq_{la} Y$. □

4. ABSOLUTE LORENZ ORDERING OF GENERALIZED GAMMA DISTRIBUTIONS

Let $GG(p,\beta,\gamma,a)$ be the four-parameter generalized gamma distribution with density

$$(10) \qquad f(x) = \frac{a(x-\gamma)^{ap-1}\exp\left[-\left(\frac{x-\gamma}{\beta}\right)^{a}\right]}{\beta^{ap}\Gamma(p)}, \quad x \geq \gamma,\ a>0,\ \beta>0,\ p>0,$$

where $\Gamma(\cdot)$ denotes the complete gamma function. If $\gamma = 0$, we have the three-parameter generalized gamma distribution considered, among others, by McDonald (1989), as descriptive model for the distribution of income. The four-parameter generalized gamma distribution includes Weibull distributions ($p = 1$), half-normal distributions ($p = 1/2$, $a = 2$, $\gamma = 0$) and, of course, ordinary gamma distributions ($a = 1$). The moment of order r about γ of $GG(p,\beta,\gamma,a)$ can be found in Johnson *et al.* (1994):

$$(11) \qquad E[(X-\gamma)^r] = \frac{\beta^r\Gamma\left(p+\frac{r}{a}\right)}{\Gamma(p)}.$$

McDonald (1989) fitted the gamma generalized distribution to U.S.A. family nominal income for 1970-1980 and obtained estimates of a, β and p for 1970, 1975 and 1980. The results of these estimations show large variations on β for the period under consideration, whereas variations on a and p are very small. This suggests to consider the impact upon absolute Lorenz curve of variations on β, when the other parameters are fixed. The next result characterizes the parameter space of the four-parameter generalized gamma distribution in terms of absolute Lorenz ordering, when a and p are fixed.

Corollary 2. *Let* $X_1 \sim GG(p,\beta_1,\gamma_1,a)$ *and* $X_2 \sim GG(p,\beta_2,\gamma_2,a)$. *If* $(a-1)\cdot(ap-1) \geq 0$ *then*

$$X_1 \leq_{la} X_2 \Longleftrightarrow \beta_1 \leq \beta_2.$$

Proof

It is clear from (1) that the absolute Lorenz curve is invariant under location changes, so the location parameters γ_1 and γ_2 can be set null, without loss of generality.

($\Rightarrow$) From Theorem 3 (ii) it follows that if $X_1 \leq_{la} X_2$, then

$$E[\Phi(X_1-\mu_1)] \leq E[\Phi(X_2-\mu_2)]$$

for all convex functions Φ, provided these expectations exist. In particular, by taking $\Phi(x) = x^2$, we obtain that $V[X_1] \leq V[X_2]$. From (11) we have that

$$V[X_i] = \beta_i^2\frac{\Gamma(p)\Gamma\left(p+\frac{2}{a}\right) - \left[\Gamma\left(p+\frac{1}{a}\right)\right]^2}{[\Gamma(p)]^2}, \quad (i=1,2).$$

Therefore $\beta_1 \leq \beta_2$.

($\Leftarrow$) Suppose $\beta_1 < \beta_2$ (the case $\beta_1 = \beta_2$ is trivial). Let f_1 and f_2 be the density functions of the random variables X_1 and X_2, respectively. From (11), it follows that

$$E[X_i] = \frac{\beta_i \Gamma\left(p+\frac{1}{a}\right)}{\Gamma(p)} = \beta_i t, \quad (i=1,2),$$

where $t = \Gamma\left(p+\frac{1}{a}\right) / \Gamma(p)$. Since $\beta_1 < \beta_2$, by Corollary 1, we will show that the ratio $f_1(x+\beta_1 t)/f_2(x+\beta_2 t)$ is unimodal for $x \geq -\beta_2 t$. It is clear that $f_1(x+\beta_1 t)/f_2(x+\beta_2 t) = 0$ for $-\beta_2 t \leq x \leq -\beta_1 t$. Now suppose $x > -\beta_1 t$ and denote

$$h(x) = \frac{f_1(x+\beta_1 t)}{f_2(x+\beta_2 t)} = \left(\frac{\beta_2}{\beta_1}\right)^{ap} \cdot \left(\frac{x+\beta_1 t}{x+\beta_2 t}\right)^{ap-1} \cdot \exp\left\{\left(\frac{x+\beta_2 t}{\beta_2}\right)^a - \left(\frac{x+\beta_1 t}{\beta_1}\right)^a\right\}.$$

Since $h(x) > 0$ for all $x > -\beta_1 t$ and $\lim_{x \to -\beta_1 t} h(x) = 0$, in order to prove the unimodality of $h(x)$, it is sufficient to show that $h'(x)$ has at most one real root on $(-\beta_1 t, \infty)$. Since

$$h'(x) = M(x) \cdot \left[\frac{(ap-1)(\beta_2-\beta_1)t}{(x+\beta_2 t)(x+\beta_1 t)} + \frac{a(x+\beta_2 t)^{a-1}}{\beta_2^a} - \frac{a(x+\beta_1 t)^{a-1}}{\beta_1^a}\right],$$

where

$$M(x) = \left(\frac{\beta_2}{\beta_1}\right)^{ap} \cdot \exp\left\{\left(\frac{x+\beta_2 t}{\beta_2}\right)^a - \left(\frac{x+\beta_1 t}{\beta_1}\right)^a\right\} \cdot \frac{(x+\beta_1 t)^{ap+a-2}}{(x+\beta_2 t)^{ap-1}} \geq 0$$

for $x \geq -\beta_1 t$, it follows that $h'(x) = 0$ if and only if

$$\frac{(ap-1)(\beta_2-\beta_1)t}{(x+\beta_2 t)(x+\beta_1 t)^a} + \frac{a}{\beta_2^a}\left(\frac{x+\beta_2 t}{x+\beta_1 t}\right)^{a-1} = \frac{a}{\beta_1^a}. \tag{12}$$

By defining

$$R(x) = \frac{(ap-1)(\beta_2-\beta_1)t}{(x+\beta_2 t)(x+\beta_1 t)^a}, \quad S(x) = \frac{a}{\beta_2^a}\left(\frac{x+\beta_2 t}{x+\beta_1 t}\right)^{a-1},$$

we have that $h'(x_0) = 0$ if and only if

$$R(x_0) + S(x_0) = \frac{a}{\beta_1^a}. \tag{13}$$

When $ap > 1$ and $a > 1$, both functions $R(x)$ and $S(x)$ are strictly decreasing, for $x \geq -\beta_1 t$. Hence, $R(x) + S(x)$ is strictly decreasing and (13) has at most one solution on $(-\beta_1 t, \infty)$. Similarly, if $ap < 1$ and $a < 1$, it follows that $R(x) + S(x)$ is strictly

increasing for $x \geq -\beta_1 t$, attaining the value a/β_1^a at most once. If $a = 1$ or $ap = 1$, it is easy to see that (13) has at most one solution on $(-\beta_1 t, \infty)$. □

Salem and Mount (1974) used ordinary gamma distributions for fitting empirical income data. The ordinary gamma distribution $G(p, \beta, \gamma)$ is obtained by setting $a = 1$ in (10). The next result characterizes, as a particular case of Corollary 2, the parameter space of this family, in terms of absolute Lorenz ordering when the shape parameter p is fixed.

Corollary 3. *Let* $X_1 \sim G(p, \beta_1, \gamma_1)$ *and* $X_2 \sim G(p, \beta_2, \gamma_2)$. *Then*

$$X_1 \leq_{la} X_2 \Longleftrightarrow \beta_1 \leq \beta_2.$$

Corollary 1 can also be used to prove that the ordinary gamma distribution can be ordered by the parameter p when the scale parameter β is fixed. The following result can be proven as Corollary 2.

Corollary 4. *Let* $X_1 \sim G(p_1, \beta, \gamma_1)$ *and* $X_2 \sim G(p_2, \beta, \gamma_2)$. *Then*

$$X_1 \leq_{la} X_2 \Longleftrightarrow p_1 \leq p_2.$$

5. REMARKS AND RELATED TOPICS

The generalized Lorenz order and the absolute Lorenz order are closely connected with the usual Lorenz ordering (see Arnold, 1987). In general, the rankings produced by these orderings do not coincide unless distributions have common means. In fact, if $\mu_X = \mu_Y$, Theorem 3.2 of Arnold (1987) is obtained as a particular case of Theorem 3 (i). The Lorenz and generalized Lorenz orderings within the gamma parametric family have been considered in Wilfling (1996) and Ramos *et al.* (2000), respectively.

The concave order is of interest in reliability theory. Stoyan (1983) states that X is smaller in mean used life than Y if $X \leq_{cv} Y$ holds. The origins of this ordering may be found in Marshall and Proschan (1970), where it appeared as a particular relationship useful for obtaining bounds for the mean lifetime of series systems in reliability theory. The concave ordering is a counterpart of the so-called convex ordering (Stoyan, 1983). The convex order is defined by requiring (2) to hold for all non-decreasing and convex functions Φ for which the expectations exist. In the case of equal means, the convex and the dilation orders are equivalent.

The relationships between some notions that are common to reliability theory and economics have been studied by several authors. Chandra and Singpurwalla (1981) proved that the Lorenz curve can be used to characterize DFR (Decreasing Failure Rate) random variables. Kochar (1989) showed that Lorenz ordering is the same as HNBUE-ordering (Harmonic New Better than Used in Expectation), which is a well-known concept for the comparison of the aging properties of two life distributions.

It is well known that the concave order is preserved under an increasing concave transformation (see Theorem 3.A.5 of Shaked and Shanthikumar, 1994). Now, from Theorem 3 (i), it follows that this property holds with $\leq_{gl}$ replacing $\leq_{cv}$. Thus, the following result generalizes the sufficient condition of Theorem 3.1, given by Moyes (1989) for a finite population.

Corollary 5. *Let X and Y be two non-negative random variables. If $X \leq_{gl} Y$ and g is any increasing and concave function, then $g(X) \leq_{gl} g(Y)$.*

In economics, the preservation of the ranking of distributions has obvious implications when one views the transformation g as a taxation scheme (see Moyes, 1988, for further discussion).

The relationship between the absolute Lorenz curve and the dilation order suggests a similar characterization between the so-called dispersion function and the dilation order. The *dispersion function* of a random variable X with a finite mean is given by (Muñoz-Pérez and Sánchez-Gómez, 1990)

$$D_X(u) = E[|X - u|] \quad \text{for all } u \in \mathbb{R}.$$

This function characterizes the distribution function and measures directly the dispersion of X about each point $u \in \mathbb{R}$. The dispersion function provides a partial ordering between random variables with respect to the dispersion in the following sense: we say that Y is at least as dispersed as X if $D_{X-\mu_X}(u) \leq D_{Y-\mu_Y}(u)$ for all $u \in \mathbb{R}$. Now, from Theorem 1 of Muñoz-Pérez and Sánchez-Gómez (1990) and Theorem 3 (ii), it follows that the ordering induced by such a function is equivalent to the ordering induced by the absolute Lorenz curve. This is stated formally in the following result.

Corollary 6. *Let X and Y be two non-negative random variables. Then*

$$X \leq_{la} Y \text{ if and only if } D_{X-\mu_X}(u) \leq D_{Y-\mu_Y}(u) \text{ for all } u \in \mathbb{R}.$$

This result suggests that the dispersion function and the absolute Lorenz curve play a similar role in statistics.

On the other hand, the functional form of inequality indices has been studied by several authors (see, for example, Kolm, 1976; Atkinson, 1970; Nygard and Sandström, 1981; Champernowne and Cowell, 1998). Moyes (1987) showed that the absolute Lorenz ordering is consistent and is implied by the unanimous partial order generated by the class of *absolute inequality indices*, introduced by Kolm (1976). Theorem 3 (ii) suggests that a reasonable summary measure of absolute inequality is provided by an index of the form $E[\Phi(X-\mu_X)]$ for any convex functions Φ. In particular, the choices $\Phi(x)=x^2$ and $\Phi(x)=|x|$ lead, respectively, to the variance and the absolute mean deviation of X. These results show that, in some way, the concepts of *absolute inequality* and *dispersion* could be considered as *allotropic forms* of the same primary concept.

Some other general results and examples relating dispersion and inequality can be found in Frosini (1984).

Finally, the «single-crossing property» has often been used to compare distributions under orderings related to the absolute Lorenz order. Theorem 6.4 of Arnold (1987) and Theorem 2.1 of Ramos *et al.* (2000) provide sufficient conditions for the Lorenz and the generalized Lorenz orders, respectively, based on this property. Theorem 1.5.1 of Stoyan (1983) provides a sufficient condition for the convex order based on the «cut criterion», which is a property very similar to the «single-crossing property».

ACKNOWLEDGMENTS

The authors are grateful to the referee for interesting suggestions on an earlier draft of this paper.

REFERENCES

Arnold, B. C. (1987). *Majorization and the Lorenz order: A brief Introduction.* Springer, New York.

Atkinson, A. B. (1970). «On the Measurement of Inequality». *Journal of Economic Theory*, 2, 244-263.

Chandra, M. and Singpurwalla, N. D. (1981). «Relationships between some notions which are common to reliability theory and economics». *Mathematics of Operations Research*, 6, 113-121.

Champernowne, D. G. and Cowell, F. A. (1998). *Economic inequality income distribution.* Cambridge University Press, United Kingdom.

Chong, K. M. (1974). «Some extensions of a theorem of Hardy, Littlewood and Polya and their applications». *Canadian Journal of Mathematics*, 26, 1321-1340.

Frosini, B. V. (1984). «Concentration, dispersion and spread: an insight into their relationship». *Statistica*, 3, 373-393.

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1994). *Continuous univariate distributions* (2.ª ed.). John Wiley & Sons, inc. New York.

Hardy, G. H., Littlewood, J. E. and Pólya, G. (1929). «Some simple inequalities satisfied by convex functions». *Messenger of Mathematics*, 58, 145-152.

Hickey, R. J. (1986). «Concepts of dispersion in distributions: a comparative note». *Journal of Applied Probability*, 23, 914-921.

Kochar, S. C. (1989). «On extensions of DMRL and related partial orderings of life distributions». *Communications in Statistics-Stochastic Models*, 5, 235-245.

Kolm, S.-C. (1976). «Unequal inequalities 1». *Journal of Economic Theory*, 12, 416-422.

Lambert, P. (1993). *The distribution and redistribution of income*. Manchester: Manchester University Press.

Lewis, T. and Thompson, J. W. (1981). «Dispersive distributions and the connection between dispersivity and strong unimodality». *Journal of Applied Probability*, 18, 76-90.

Marshall, A. W. and Proschan, F. (1970). «Mean life of series and parallel systems». *Journal of Applied Probability*, 7, 165-174.

McDonald, J. B. (1984). «Some generalized functions for the size distribution of income». *Econometrica*, 52, 647-663.

Moyes, P. (1987). «A new concept of Lorenz domination». *Economic Letters*, 23, 203-207.

Moyes, P. (1988). «A note on minimally progressive taxation and absolute income inequality». *Social Choice and Welfare*, 5, 227-234.

Moyes, P. (1989). «Some Classes of Functions that Preserve the Inequality and Welfare Orderings of Income Distributions». *Journal of Economic Theory*, 49, 347-359.

Muñoz-Pérez, J. and Sánchez-Gómez, A. (1990). «Dispersive ordering by dilation». *Journal of Applied Probability*, 27, 440-444.

Nygard, F. and Sandström, A. (1981). *Measuring income inequality*. Almqvist & Wiksell International, Stockholm.

Ramos, H. M., Ollero, J. and Sordo, M. A. (2000). «A sufficient condition for generalized Lorenz order». *Journal of Economic Theory*, 90, 286-292.

Ryff, J. V. (1963). «On the representation of doubly stochastic operators». *Pacific Journal of Mathematics*, 13, 1379-1386.

Salem, A. B. Z. and Mount, T. D. (1974). «A convenient descriptive model of income distribution». *Econometrica*, 42, 1115-1127.

Shaked, M. and Shanthikumar, J. G. (1994). *Stochastic orders and their applications.* Academic Press, London.

Shorrocks, A. F. (1983). «Ranking Income Distributions». *Economica*, 50, 3-17.

Stoyan, D. (1983). *Comparison Methods for Queues and Other Stochastic Models.* John Wiley, New York.

Wilfling, B. (1996). «A sufficient condition for Lorenz ordering». *Sankhyã*, Series B 58, 62-69.

Yitzhaki, S. (1999). «Neccesary and sufficient conditions for dominace using generalized Lorenz curves». In: *Advances in Econometrics, Income distribution and methodology of science: essays in honor of Camilo Dagum.* Springer.

QÜESTIIÓ, vol. 26, 1-2, p. 29-59, 2002

STATISTICAL PROCEDURES FOR SPATIAL POINT PATTERN RECOGNITION

J. MATEU*

Spatial structures in the form of point patterns arise in many different contexts, and in most of them the key goal concerns the detection and recognition of the underlying spatial pattern. Particularly interesting is the case of pattern analysis with replicated data in two or more experimental groups. This paper compares design-based and model-based approaches to the analysis of this kind of spatial data. Basic questions about pattern detection concern estimating the properties of the underlying spatial point process within each experimental group, and comparing the properties between groups. The paper discusses how either approach can be implemented in the specific context of a single-factor replicated experiment and uses simulations to show how the model-based approach can be more efficient when the underlying model assumptions hold, but potentially misleading otherwise.

Keywords: Expected significance levels, Replicated patterns, Spatial point structures

AMS Classification (MSC 2000): 62M05, 60G55

* Department of Mathematics, Universitat Jaume I, E-12071, Castellón, Spain.
Correspondence to: Jorge Mateu, Department of Mathematics, Universitat Jaume I, E-12071, Castellón, Spain. Email:mateu@mat.uji.es.

– Received July 2000.
– Accepted January 2001.

1. INTRODUCTION

For the single replicate point patterns which dominate the spatial point pattern literature (Cressie (1993), Diggle (1983) and Ripley (1981)), there has been a strong emphasis on fitting models to detect and recognize the spatial structures.

Given the wealth of opportunities for collecting replicated spatial point pattern data, for example in clinical neuroanatomy, or in materials science, where replicate images can easily be obtained using standard microscopical equipment, it is surprising that few appropriate methods of analysis have been suggested. Design-based methods have been suggested by Diggle *et al.* (1991), Wilson (1998), Wilson *et al.* (1998) and Baddeley *et al.* (1985, 1993). The first of these presents a method, motivated by analysis of variance, for the assessment of replicated spatial point patterns, allowing for a single replicate from each individual of interest. The authors illustrate their method with an example of two-dimensional replicated data from clinical neuroanatomy. The second extends the methodology developed by Diggle *et al.* (1991), to allow for multiple replicates per individual, and to incorporate more complex experimental set-ups. Again, the example data is from clinical neuro-anatomy, but the data is three-dimensional. Baddeley *et al.* (1993) present an alternative method, which is again motivated by analysis of variance, but uses a ratio regression approach, and again allows for replication within individuals as well as between individuals. In this paper, we review in detail the methods developed by Wilson (1998).

In many applications such as biological or neuroanatomical applications, the points of interest are the centres of cells, and it is a key point the modelling of cell centre positions. Figure 1 represents one such example, where each one of the 12 plots shows the spatial positions (in form of a point in the square window) of pyramidal neurons of the Cingulate Cortex of a particular selected individual. This data set was analyzed by Diggle *et al.* (1991) and further information on the data can be found there. At one extreme, the cell centres might be thought of as a hard-core process, with the hard-core parameter being equal to the cell diameter. A pairwise interaction process (Ripley, 1977; Diggle *et al.*, 1994) might be a more appropriate model for cell centre data; this class of point process is described at length in Section 2. Broadly speaking, cells are separated by a particular distance δ with a given probability, where small distances may be assigned small probability (a hard-core process of hard-core radius ρ is a special case of this; the separation probability is 0 for distances less than ρ and 1 for distances greater than ρ). Through this framework, the rigid regularity of the hard-core process can be relaxed, presenting a more realistic description of cell centre behaviour. There are various ways in which we could choose to fit models to the replicated spatial point pattern data. Appropriate single replicate methods are reviewed by Diggle *et al.* (1994). Our chosen approach is an extension of the maximum pseudo-likelihood method.

We have thus highlighted two contrasting approaches to the assessment of replicated spatial point patterns, the design-based ANOVA approach and the model-based maximum pseudo-likelihood approach. Both methodologies are competent and useful at detecting pattern structure though their practical performance is little known.

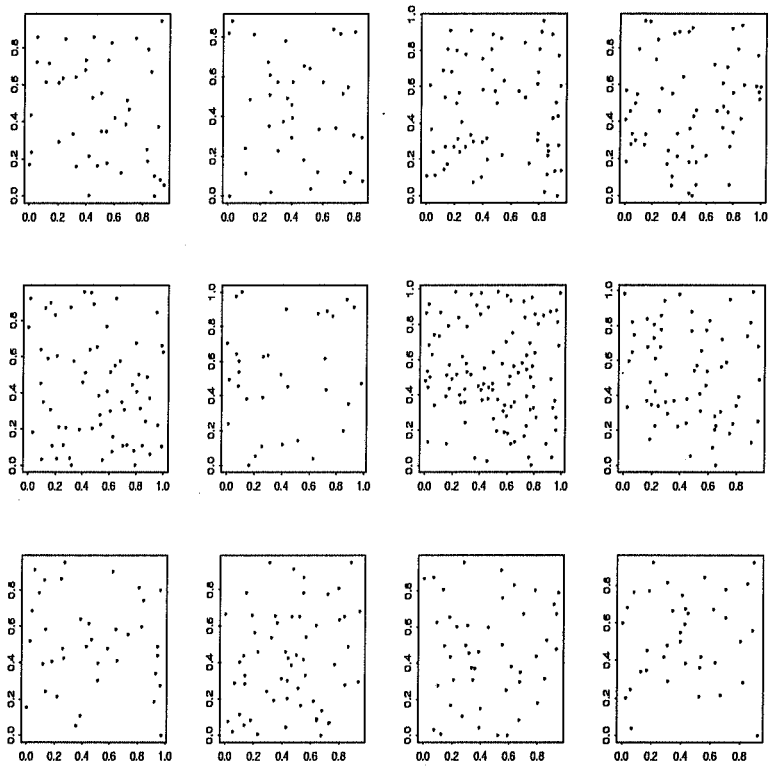


Figure 1. Digitized pyramidal neuron positions.

In this paper, we compare and contrast the two approaches under a variety of experimental conditions. As part of the study, we investigate the effect of fitting various pairwise interaction models to different kinds of replicated spatial point pattern data. There are two questions which we seek in particular to answer:

- When the model specification is correct, is the model-based approach more powerful than the design-based approach for a given set of data?
- When the model specification is incorrect, does the model-based procedure break down, to give misleading results?

In Section 2, we introduce formally the pairwise interaction process, and highlight the classes of this process upon which we base this study. In Section 3, we review the design-based ANOVA approaches of Diggle *et al.* (1991) and Wilson (1998). In Section 4, we review the maximum pseudo-likelihood method for replicated spatial data. Section 5 defines the concept of Expected Significance Level used throught the simulation study. Finally, in Section 6, we give the results of a simulation study which investigates the above two questions.

2. PAIRWISE INTERACTION POINT PROCESSES

2.1. Theory setup

Assume we have observed a pattern of points $X = \{x_i \in A : i = 1,\ldots,n\}$ in a planar region A. The x_i are called *events* to distinguish them from generic points in $x \in A$. Each point process on A can be defined by a sequence of intensity functions $\lambda_n(x_1,x_2,\ldots,x_n)$ which are measurable functions that do not depend on the order of the events. Then, the point process will contain n events with probability p_n,

$$p_n = \begin{cases} \exp(-\nu(A)) & \text{if } n=0 \\ \frac{\exp(-\nu(A))}{n!}\int_{A^n}\lambda_n(x_1,x_2,\ldots,x_n)dx_1\cdots dx_n & \text{if } n\geq 1 \end{cases} \tag{2.1}$$

where $\nu(A)$ represents the Lebesgue measure of A and $\sum_{n=0}^{\infty} p_n = 1$ (Kelly & Ripley, 1976).

The joint density function for X,

$$f[(x_1,x_2,\ldots,x_n),n] = \frac{\exp(-\nu(A))}{n!}\lambda_n(x_1,x_2,\ldots,x_n) \tag{2.2}$$

can be factored uniquely for $n = 1,2,\ldots$ as (Daley & Vere-Jones, 1988)

$$\begin{aligned} f[(x_1,x_2,\ldots,x_n),n] &= \frac{\exp(-\nu(A))}{\alpha n!}\exp\left\{\sum_{i=1}^{n} g_1(x_i) + \sum_{i=1}^{n}\sum_{j>i} g_{12}(x_i,x_j) + \right. \\ &\left. \cdots + g_{12\cdots n}(x_1,x_2,\ldots,x_n)\right\} \end{aligned} \tag{2.3}$$

where α is the normalising constant which usually does not have a closed-form expression.

The likelihood (Janossy density) is given by

$$l(x_1,x_2,\ldots,x_n) = n!f[(x_1,x_2,\ldots,x_n),n] \tag{2.4}$$

The point process defined by (2.3) and (2.4) is called *Gibbs process*. Adding a nearest-neighbour condition to a Gibbs process we have a Markov process. The nearest-neighbour

condition is usually defined in terms of the Euclidean distance. Two events x_i and x_j are said to be neighbours if $\|x_i - x_j\| \leq r$, for some $r > 0$. A *clique* is defined to be a single event or a set of events, all of which are neighbours of each other. By the Hammersley-Clifford Theorem, $g_{1,2,\ldots,k}(x_{i_1}, x_{i_2}, \ldots, x_{i_k}) = 0$ unless the events $x_{i_1}, x_{i_2}, \ldots, x_{i_k}$ form a clique. Then, the point process defined by (2.3) is Markov of range r. Markov point processes were introduced by Ripley & Kelly (1977), whereas the concept of Gibbs distributions has been used in statistical physics for a longer time (Preston, 1977). Since the introduction of Markov point processes in spatial statistics attention has focussed on the special case of *pairwise interaction models* in which each configuration of events interacts only via pairs of points from this configuration. These provide a large variety of complex patterns starting from simple potential functions which are interpretable as attractive and/or repulsive forces acting among events. Pairwise interaction models are simple exponential families and are widely used as they are very amenable to simulation and iterative statistical methods.

For a *pairwise interaction point process*, X, of n events in a bounded region A, the joint density (2.3) is of the form

$$f(X) = C^{-1} \frac{\beta^n}{n!} \exp\left\{ -\sum_{i=1}^{n} \sum_{j>i} \Phi\left(\|x_i - x_j\|; \theta\right) \right\} \tag{2.5}$$

where C is a normalising constant depending on β and Φ, $\|.\|$ denotes the Euclidean distance, Φ is a *potential function* depending on a set of parameters θ and β is a parameter which determines the intensity of the process. Note that (2.5) is (2.3) with $g_1(x_i) = \log\beta$, $g_{12}(x_i, x_j) = -\Phi(.)$, $g_{12\ldots k}(.) = 0$ for $k \geq 3$ and $C = \alpha / \exp(-\nu(A))$. The function

$$U_n(X) = \sum_{i=1}^{n} \sum_{j>i} \Phi\left(\|x_i - x_j\|; \theta\right) \tag{2.6}$$

is usually called *total potential energy*. Often, an *interaction function* is used instead of the potential function,

$$e(t) = \exp(-\Phi(t)) \tag{2.7}$$

Using (2.7), then (2.5) is usually represented as (Baddeley & Møller, 1989)

$$f(X) = \alpha \prod_i b(x_i) \prod_{i<j} e(x_i, x_j) \tag{2.8}$$

where α stands for the normalising constant and b, e (the interaction function) are non-negative measurable functions.

Note that restrictions on the form of the potential function are needed to ensure that the normalising constant is finite. A sufficient condition is that $\Phi(t) \geq 0$ for all t and decreases as t increases.

The resulting processes are called inhibitory and generate patterns with varying degrees of spatial regularity according to the specific specification of the potential/interaction function. They do not seem to be able to produce clustered patterns in sufficient variety. The original clustering model of Strauss (Strauss, 1975) turned out (Kelly & Ripley, 1976) to be non-integrable for those parameter values such that $\Phi(t) < 0$ for $t > h$, where $h > 0$ is called the *hard-core distance*. Gates & Westcott (1986) showed that partly-attractive potentials may violate a stability condition, implying that they produce extremely clustered patterns with high probability. Also, simulations by Møller (1993) suggest that the behaviour of the Strauss model with fixed n undergoes an abrupt transition from «Poisson-like» patterns to tightly clustered patterns rather than exhibiting intermediate, moderately clustered patterns.

It is known that the distribution given by (2.5) coincides with the stationary distribution of a spatial birth-death process on A which provides one method for simulating realisations of the point process so defined.

2.2. Models

Attention in this paper is focussed on three different parametric families of models, each having a specific interaction function depending on a scalar parameter.

Our first model has interaction function

$$e(t) = \begin{cases} 1-(1-t^2/\theta^2)^2 & \text{if } t \le \theta \\ 1 & \text{if } t > \theta. \end{cases} \tag{2.9}$$

The interaction function is continuously differentiable with respect to t and θ defines the range of interaction. This model was used by Diggle (1986) in a comparison between approximate maximum likelihood and maximum pseudo-likelihood estimators for θ. It was also used by Diggle *et al.* (1994) to compare several parametric estimation techniques. Small values of θ are indicative of weak interaction whereas larger values indicate strong interaction. This model will be called *Diggle* model throughout the paper.

The second model has interaction function

$$e(t) = 1 - \exp\left(-\frac{t^2}{\theta^2}\right). \tag{2.10}$$

It is called *Very-Soft-Core* (VSC) model and was used in Ogata & Tanemura (1984) and Diggle *et al.* (1994). The range of interaction is infinite, i.e., $e(t) < 1$ for all finite t. Again, small values of θ are indicative of weak interaction whereas larger values indicate strong interaction.

Finally, we have used a third model which is qualitatively different from the other first two. It has interaction function

$$e(t) = \begin{cases} \frac{t}{\theta} & \text{if } t \leq \theta \\ 1 & \text{if } t > \theta \end{cases} \tag{2.11}$$

and due to its lineal form will be called *linear* model. As with the Diggle model, the interaction function is continuously differentiable with respect to t and θ defines the range of interaction. Small values of θ are indicative of weak interaction whereas larger values indicate strong interaction.

2.3. Simulation method

The spatial birth-and-death process provides the framework under which Ripley (1977, 1981) proposes to simulate a Markov point process on the bounded Borel set $A \subset \Re^d$ with n fixed. The method is related to Markov processes used in statistical mechanics and surveyed originally by Hastings (1970). Other techniques such as Metropolis-Hastings algorithms (Geyer & Møller, 1994; Møller, 1992) are clearly possible.

Consider a set of particles interacting according to a certain potential function, on a square A with periodic boundary, i.e., A is identified with a torus. The algorithm is as follows.

First, select n events from a uniform distribution on A and call this initial point pattern $X^n(0)$. At step $(t+1)$, delete systematically in turn one of the n events of $X^n(t) = \{x_1, x_2, \ldots, x_n\}$, say event x_i, and let $X^n(t) \setminus \{x_i\}$ denote the point pattern formed by removing x_i from $X^n(t)$. Let

$$p(u; X^n(t) \setminus \{x_i\}) = \frac{f(X^n(t) \setminus \{x_i\}, u)}{f(X^n(t) \setminus \{x_i\})} \tag{2.12}$$

denote the conditional intensity at $u \in A$ given $X^n(t) \setminus \{x_i\}$. Define

$$M = \sup_{u \in A} p(u; X^n(t) \setminus \{x_i\}).$$

Finally, select an event u from a uniform distribution on A and set $X^n(t+1) = \{X^n(t) \setminus \{x_i\}, u\}$ with probability $p(u; X^n(t) \setminus \{x_i\})/M$; otherwise, selection is repeated until a qualifying u is found. This method ensures that samples taken every n steps have no points in common. Ultimately, convergence to a Markov point process with likelihood $f(.)$ will occur. This algorithm is analogous to the Gibbs sampler on the spatial lattice (Geman & Geman, 1984).

In the simulation study we simulate on a larger polygon $(-0.5, 1.5) \times (-0.5, 1.5)$ and use a toroidal shift to make our point process compatible with the MPL method discussed in Section 4. We then extract the central unit square from the polygon, and use this part of the data for further analysis.

3. A DESIGN-BASED APPROACH

A descriptor which makes use of the K-function (the reduced second moment measure (Ripley, 1977, 1988) is suggested. For a stationary, isotropic, orderly point process, the K-function may be defined as

$$K(t) = 2\pi\lambda^{-2}\int_0^t \lambda_2(r)dr \tag{3.1}$$

where λ is the point process intensity (the expected number of events per unit area in two dimensions, volume in three dimensions and so on), and $\lambda_2(r)$ is the corresponding second order intensity function. Heuristically, the K-function describes the expected number of cells within a distance t of an arbitrarily chosen point of the process, and, as such, provides a good indicator of regularity or clustering in a given point pattern.

Diggle *et al.* (1991) develop a method for investigating whether schizophrenic patients suffer from some structural defect of the cerebral cortex. Specifically, the authors set out to establish whether the brain of the schizophrenic demonstrates unusual arrangements of neurons. A single, two dimensional neuronal cell pattern is obtained from each of several schizophrenic, schizoaffective and control individuals.

The statistic defined by Diggle *et al.* (1991) assumes the simplest case scenario, namely that there is a one-way experimental set-up, with g groups of replicated spatial data, and within each of those g groups, each individual provides a single point pattern realisation. Their proposed statistic is given by the following expression

$$D_g = \sum_{i=1}^{g}\int_0^{t_0}\left[\bar{K}_i^{1/2}(t) - \bar{K}^{1/2}(t)\right]^2 dt, \tag{3.2}$$

where

$$\bar{K}_i(t) = \frac{\sum_{j=1}^{m_i} n_{ij}\hat{K}_{ij}(t)}{\sum_{j=1}^{m_i} n_{ij}} \tag{3.3}$$

and

$$\bar{K}(t) = \frac{1}{n}\sum_{i=1}^{g} n_i\bar{K}_i(t). \tag{3.4}$$

In addition, n_{ij} is the number of points in the sampling window from individual j in group i, $n_i = \sum_{j=1}^{m_i} n_{ij}$ and $n = \sum_{i=1}^{g} n_i$. We see that $\bar{K}_i(t)$ is the weighted mean K-function in group i, and $\bar{K}(t)$ is the weighted mean K-function across all groups.

The statistic D_g is similar to a between-treatment sum of squares in an ordinary one way analysis of variance, in that it compares group mean K-functions with the overall

mean K-function, in an attempt to quantify between group differences. This test statistic in combination with a Monte-Carlo procedure is used to assess the significance of between-group differences in the spatial patterns.

3.1. A new test statistic

In a further step, Wilson (1998) suggested a new test statistic for comparing groups of spatial data, which is a modification of the Diggle *et al.* (1991) test statistic D_g.

To motivate the new test statistic, first recall a result from the analysis of variance literature. Suppose that we have a simple, one-way experiment, with observations $y_{i1}, \ldots, y_{im_i}$ in g groups (so that group i contains m_i observations). Then under the assumption that

$$\mathrm{Var}(y_{ij}) = \frac{\sigma^2}{n_{ij}},$$

it can be shown that the best estimate of the between-treatment sum of squares is given by expression (3.5)

$$\mathrm{BTSS} = \left[\sum_{i=1}^{g} n_i (\bar{y}_i - \bar{y})^2\right], \tag{3.5}$$

where

$$\bar{y}_i = \frac{\sum_{j=1}^{m_i} n_{ij} y_{ij}}{n_i} \tag{3.6}$$

and

$$\bar{y} = \frac{\sum_{i=1}^{g} n_i \bar{y}_i}{n}, \tag{3.7}$$

with $n = \sum_{i=1}^{g} n_i$. Returning to spatial point patterns, consider now an estimated K-function from individual j in one of the groups i, $\hat{K}_{ij}(t)$, say. Fix t at a particular distance t_d, say. Then $\hat{K}_{ij}(t_d)$ represents a single realised value. Furthermore, assume that $\mathrm{Var}(\hat{K}_{ij}(t_d))$ is proportional to $1/n_{ij}$, to reflect the fact that a K-function estimated from a large number of points has a small variance, and vice-versa. The variance structure here is identical to that defined above, and so we use this analysis of variance result as the basis for a new statistic.

We define

$$\bar{K}_i(t_d) = \frac{\sum_{j=1}^{m_i} n_{ij} \hat{K}_{ij}(t_d)}{n_i}$$

and extend this to cover the range of values for t_d

$$\bar{K}_i(t) = \frac{\sum_{j=1}^{m_i} n_{ij}\hat{K}_{ij}(t)}{n_i}. \tag{3.8}$$

In a similar way, we define

$$\bar{K}(t) = \frac{\sum_{i=1}^{g} n_i\bar{K}_i(t)}{n}, \tag{3.9}$$

where n_i and n are defined as before.

This new test statistic is defined as follows

$$D_{g2} = \sum_{i=1}^{g}\int_0^{t_0} w(t)n_i\left[\bar{K}_i(t) - \bar{K}(t)\right]^2 dt. \tag{3.10}$$

Note that, in addition to the new factor n_i, the statistic does not have the square root transformation, suggested by Diggle *et al.* (1991) as a variance-stabilising mechanism. This has been replaced with a weighting function $w(t)$, which down-weights the variance of the K-function estimates at large t. We select an appropriate weighting function by considering a spatially random patterns; the estimated K-function for each pattern behaves like a Poisson random variable. This follows because the K-function looks at count data within discs (in two dimensions). As the K-function itself is Poisson, then by the definition of a Poisson distribution, the variance of $\hat{K}$ is proportional to K.

Furthermore, for a spatially random process, $K(t) = \pi t^2$, i.e. $K(t) \propto t^2$. We therefore select the reciprocal of t^2 as an appropriate weighting function (for one-dimensional data we would use $w(t) = 1/t$, and for three-dimensional data we would use $w(t) = 1/t^3$, by the same argument).

3.2. A Monte-Carlo procedure

We adopt the Monte-Carlo procedure suggested by Diggle *et al.* (1991) as a means of assigning significance to observed values of the new test statistic. The procedure is reproduced from Diggle *et al.* (1991).

Suppose we have a set of point process data, divided into three groups. Suppose also that there are r_i individuals in group i. We maintain the definitions of n_{ij} and n_i given earlier.

- We begin by defining the *residual K-function*. This may be written as

$$\hat{R}_{ij}(t) = n_{ij}^{\frac{1}{2}}\left[\hat{K}_{ij}(t) - \bar{K}_i(t)\right].$$

- Using this definition, we can further obtain new functions $\hat{K}^*(t)$, which we define as

$$\hat{K}^*_{ij}(t) = \bar{K}(t) + n_{ij}^{-\frac{1}{2}} \hat{R}^*_{ij}(t),$$

  where the $\hat{R}^*_{ij}(t)$ are obtained by permuting at random, across groups, the residual K-functions, so that $\hat{R}^*_{ij}(t)$ is one of the residual K-functions, drawn at random and without replacement.

- Recompute $\bar{K}^*_i(t)$ and $\bar{K}^*(t)$ from this permutation sample, and compute the statistic D^*_{g2}.

- Repeat this procedure n_s times, and each time obtain a new value for D^*_{g2}.

- Rank each statistic from the original data amongst the n_s statistics obtained from the permutation test, and from the rank compute a p-value for each statistic. A large statistic corresponds to a small p-value and *vice-versa.*

We thus have a test statistic and a method for assigning significance to observed values from data sets. Several simulation studies in the thesis of Wilson (1998) demonstrate the improved performance of D_{g2} in comparison with D_g. Particularly, D_{g2} detects differences between groups of replicated spatial data which are not obvious from visual inspection alone.

4. A MODEL-BASED APPROACH

4.1. Parameter estimation

In this paper we focus on the edge-corrected *pseudo-likelihood* estimation method as it can be used routinely in applications and does not place artificial restrictions on the parametric form of the potential function. Moreover, this method can be easily implemented for the case where replicated spatial point pattern data is available. Other general methods such as approximations to maximum likelihood require numerical or Monte Carlo approximations to the normalising constant (Ogata & Tanemura, 1981, 1984, 1989; Penttinen, 1984) or recursive approximation methods (Moyeed & Baddeley, 1991). The Takacs-Fiksel method (Takacs, 1986; Fiksel, 1984, 1988) has been proposed in literature but will not be explored further in this paper. For a more detailed review see (Diggle *et al.*, 1994; Ripley, 1988; Geyer & Thompson, 1992).

The pseudo-likelihood for Markov processes is defined (Besag, 1977; Jensen & Møller, 1991) by

$$PL(\theta, \beta; X) = \exp\left\{ -\int_A \lambda(u; X) du \right\} \prod_{i=1}^{n} \lambda(x_i; X^i) \tag{4.1}$$

where $X^i = X - \{x_i\}$ and $\lambda(u;X)$ represents the *conditional intensity* (Papangelou conditional intensity, see Daley & Vere-Jones, 1988) of an event u, given the pattern X and is defined by $\lambda(u;X) = \frac{f(X\cup\{u\})}{f(X)}$. Usually, (4.1) is re-cast in terms of its logarithm. Maximisation of (4.1) with respect to parameters θ, β yields the maximum pseudo-likelihood estimators. The estimating equation for parameter β is (Besag, 1977)

$$n = \beta \int_A \lambda_0(u;X)du \tag{4.2}$$

where $\lambda_0(u;X)$ is derived as (Diggle *et al.*, 1994)

$$\lambda(u;X) = \beta \exp\left\{ -\sum_{j=1}^{n} \Phi\left(\|u - x_j\|;\theta\right)\right\} = \beta\lambda_0(u;X). \tag{4.3}$$

Then, the estimating function for parameter θ becomes

$$PL(\theta) = \sum_{i=1}^{n} \log\left\{\lambda_0(x_i;X^i)\right\} - n\log\left\{\int_A \lambda_0(u;X)du\right\}. \tag{4.4}$$

For inhibitory pairwise interaction models it is known that pseudo-likelihood estimation is a special case of the Takacs-Fiksel method when the interaction radius is fixed (Ripley, 1988; Diggle *et al.*, 1994). The pseudo-likelihood equation (4.1) can be interpreted as the limit case of pseudo-likelihood for lattice processes (Besag, 1977; Besag *et al.*, 1982). For Markov processes of finite range, maximum pseudo-likelihood estimators are consistent (Jensen & Møller, 1991). Asymptotic normality is considered in Jensen (1993).

The extension to the replicated case is straightforward. Let $PL_i(\theta,\beta;X_i)$ denote the pseudo-likelihood equation (4.1) for the i-th pattern, $i = 1,\ldots,k$, k denoting the number of replicates. Then, the pooled pseudo-likelihood function for the replicated patterns is given by

$$PL^P(\theta,\beta;X^P) = \prod_{i=1}^{k} PL_i(\theta,\beta;X_i), \tag{4.5}$$

where X^P denotes the whole set of point patterns and X_i stands for the i-th pattern. If we use $log\ PL^P$, then we sum up the k terms instead of multiplying them.

Usually, in applications, the region A is a sampled sub-region of a much larger region within which the phenomenon of interest operates, and some form of edge-correction is vital. We use here an adaptation of Ripley's edge-correction (Ripley, 1977, 1988). See also Diggle *et al.* (1994). The idea is to replace those summations of the form

$$\sum_{j>i} \Phi\left(\|x_i - x_j\|;\theta\right)$$

appearing, for example, in the density function (2.5), by

$$\frac{1}{2}\sum_{j\neq i} w_{ij}^{-1}\Phi\left(\|x_i - x_j\|;\theta\right)$$

where w_{ij} is the proportion of the circumference of the circle with centre x_i and radius $\|x_i - x_j\|$ which is contained within A. This edge-correction compensates for the omission of contributions to this total potential from unobserved events outside A.

4.2. A formal test: Testing for group differences

Let θ denote the interaction parameter in a pairwise interaction model. Then to test for group differences for replicated data via the model-based or parametric method, we use the following procedure:

- Fit a different θ to each one of the g groups and get the maximised value of the log-pseudo-likelihood, PL^1, where $PL^1 = \sum_{j=1}^{g} PL_j^p$ and PL_j^p equals the maximised value of the log-pseudo-likelihood of the j-th group.
- Pool all the data, ignoring groupings, and fit a common θ' to all the data. Obtain the maximised value of the log-pseudo-likelihood, PL^0.
- Define a test statistic $T = PL^1 - PL^0$.

To assess between-group spatial differences, we use the following Monte Carlo procedure:

- Condition on the number of groups g, number of replicates per group r_i and the number of events per replicate n_{ij}. Treat this latter as the expected number of events per replicate. Simulate a new data set using these parameters and the estimate of the pooled θ' from above as the interaction parameter for all replicates.
- Compute the test statistic using the method outlined above, and get a value for this new data set.
- Repeat this procedure 99 times.
- Rank the test statistic from original data amongst these simulated test statistics, to get a p-value. A high rank of the test statistic from original data is evidence against the hypothesis of a common value of θ in all the groups.

5. ASSIGNING SIGNIFICANCE TO OBSERVED TEST STATISTICS: EXPECTED SIGNIFICANCE LEVEL (ESL)

Rather than carrying out a conventional power study to compare the performance of the two approaches, we assess the performance of each approach using the concept of Expected Significance Level (ESL), which was first introduced by Dempster and Schatzoff (1965). The theory of power is based upon decision rules, that is, in some sense arbitrary choices of cut-offs; for a test of size α, where α represents the probability of a type I error, we select a value of power to be $(1-\beta)$, where β is the probability of a type II error (so that the power represents the probability of detecting an effect given that it exists).

The idea of ESL is more flexible than this; suppose that we have a test statistic T, and in a set of n simulations under the alternative hypothesis H_A, we observe p-values $\alpha_1, \alpha_2, \ldots, \alpha_n$. In a classical power framework, we would estimate the power of the statistic for a given level α' to be

$$(1-\beta) = 1 - \frac{(\text{number of } \alpha_i : \alpha_i < \alpha')}{n}.$$

This requires the (somewhat arbitrary) specification of α'. Instead, we could take the ESL approach, and estimate the ESL by

$$\bar{\alpha} = \frac{\sum_{i=1}^{n} \alpha_i}{n}.$$

Immediately we are preserving information about the observed significance levels which is lost in conventional power summaries; in addition, there is scope to look at the range and distribution of α_i, and we make use of this in this paper. Formally, the ESL is equivalent to a uniform weighting of power over $\{\alpha : 0 \leq \alpha \leq 1\}$.

6. COMPARISON BETWEEN THE STATISTICAL PROCEDURES

We develop a simulation study to quantify and compare the parametric model-based and the non-parametric design-based approaches under different experimental circumstances. Consider several scenarios:

- Suppose we assume a certain pairwise interaction model for the data which is correct. Then we would expect the model-based approach to be better at detecting group differences than the design-based approach, as a consequence of the fact that we allow ourselves to make more assumptions about the model which generates the data. We must verify that the model-based approach gives smaller estimated Expected Significance Levels than the design-based approach under comparable experimental circumstances.

- Suppose instead that we assume a certain pairwise interaction model for the data which is incorrect. Then we might expect the model-based procedure to give misleading results; specifically,
  - significance levels obtained from the model-based procedure may or may not be wrong; to test whether this is the case, we must check that the distribution of the observed p-values under the null hypothesis of no difference between groups is Uniform on the range $(0, 1)$.
  - the parametric approach may or may not be better than the non-parametric approach in this situation. Again, we assess this by looking at Expected Significance Levels.

We must address these two questions separately, as the former requires simulations of situations where the null hypothesis is false, namely where there are differences between the groups, and the latter requires simulation of data from the null, where no between-group differences exist.

6.1. Description of simulation procedures

The simulation study proceeds in the following way:

- Simulate some replicated spatial data from model **A**, where there are r_1 replicates in group 1, r_2 replicates in group 2, and n_1 is the expected number of points per replicate in group 1, n_2 is the expected number of points per replicate in group 2. When we address the first question of the model being correct, we set $\theta_1 \neq \theta_2$, to allow us to see whether the parametric model-based procedure is superior when a difference exists between groups. When we address the issue of incorrect model specification, we set $\theta_1 = \theta_2$, to allow us to investigate the behaviour of the parametric procedure under the null hypothesis. (The non-parametric design-based procedure must perform correctly under the null, as no modelling assumptions are required by definition, and so we do not strictly need to test this).
- Perform the spatial ANOVA on the data. Obtain a p-value for the data.
- Propose a model **B** to fit the data, where **B** may or may not be the same as **A**, depending on the set-up of interest. When we address the question of the model being correct, the model specification matches that which generates the data; when we address the incorrect model specification question, the opposite is true.
- Fit the model to the data. Obtain a test statistic via the MPL approach outlined above.
- Now simulate a new data set from model **B**, and fit model **B** to the data realization. Obtain a test statistic. Repeat this 99 times. This constitutes the Monte Carlo procedure.

- Obtain a second p-value for the data.
- Repeat all of these steps many times to obtain two sets of p-values, one from the design-based approach, and one from the model-based approach.

6.2. Fundamentals for the simulation study

We fix the number of groups as 2, the number of replicates per group as 10, and the expected number of events per replicate, $E(n)$, as 30. Each replicate is fixed on a unit square. We then simulate the observed number of events to be

$$n_{ij} \sim Poisson(30). \tag{6.1}$$

We select these numbers as a compromise between an ideal simulation set-up and feasibility. If it were possible, we would simulate large numbers of replicates per group, and would look at wide varieties of combinations of r, n and g. However, computing time imposes certain constraints, and so we focus upon the above specific case. A simulation study carried out in Wilson (1998) demonstrated that ten replicates per group in the two group case is adecuate for the non-parametric procedure to give reliable results.

6.3. Case 1: Model is correct

We specify here the combinations of models **A** and **B** which are used in this case, in combination with the values of r, n and g given in the previous Section:

Table 1. (a) Data model and (b) fitted model combinations when model is correct.

(a)	(b)
Diggle	Diggle
VSC	VSC
Linear	Linear

The parameter combinations for the three models we look at are given in Table 2. Note at this stage that a parameter θ equal to 0.03, say, for the Diggle interaction process does not have the same effect as a parameter θ equal to 0.03 for the very soft core interaction process. We consider the same range of parameters principally for consistency across models.

We seek here to answer the following questions:

- When the discrepancy between θ_1 and θ_2 is large, in some sense, do both the model-based and design-based procedure perform well?
- When the discrepancy between θ_1 and θ_2 is more subtle, is the model-based procedure better at detecting the group difference?

Table 2. Data model and fitted model parameter combinations. * indicates that combination is considered, - indicates that combination is not considered.

		θ_2				
		0.03	0.05	0.07	0.08	0.09
θ_1	0.03	-	-	-	-	-
	0.05	*	-	-	-	-
	0.07	*	*	-	-	-
	0.08	*	*	*	-	-
	0.09	*	*	*	*	-

6.3.1. Results: Data = Diggle, Model = Diggle

We simulate the set-up described in Sections 6.1 and 6.2, for each pair (θ_1, θ_2) in turn. For each set-up, we estimate (a) the expected significance level for the parametric model-based procedure, (b) the expected significance level for the non-parametric design-based procedure, (c) the difference in expected significance levels for the two procedures, (d) the standard error of this difference and (e) upper and lower 95% confidence limits for the difference. The results for the Diggle data, Diggle model set-up are summarised in Table 7.

There are several cases where the parametric procedure outperforms the non-parametric (starred in the table). Moreover, in each of these cases, the absolute ESL for the parametric procedure is much smaller than that for the non-parametric. Examining these cases more closely, we see that four out of the five estimated ESLs for the parametric procedure achieve significance at the 10% level, whereas only one in the non-parametric case crosses the 10% boundary.

The initial suggestion from this, therefore, is two-fold:

- the model-based procedure is able to detect differences which are too subtle for the design-based procedure to pick up;
- when the Diggle interaction model is correctly specified, the model-based procedure performs appreciably better than the equivalent design-based procedure.

Table 7. Data model (Diggle) and fitted model (Diggle) parameter combinations. L and U indicate lower and upper 95% confidence limits respectively, on the difference in p-values from the model-based and design-based procedures. $*$ indicates a significant result in favour of the model-based procedure. Each mean is obtained from 12 simulations.

θ_1	θ_2	$E(p)$	$E(np)$	$E(p-np)$	$SE(\overline{p}-\overline{np})$	L	U
0.03	0.05	0.1292	0.3633	−0.2342	0.0911	−0.4126	−0.0557*
0.03	0.07	0.0967	0.2425	−0.1458	0.0498	−0.2434	−0.0483*
0.03	0.08	0.0925	0.2150	−0.1225	0.0591	−0.2384	−0.0066*
0.03	0.09	0.0100	0.0892	−0.0792	0.0278	−0.1336	−0.0247*
0.05	0.07	0.3892	0.2641	0.1250	0.1147	−0.0998	0.3498
0.05	0.08	0.2350	0.3112	−0.0767	0.0865	−0.2462	0.0929
0.05	0.09	0.0667	0.1550	−0.0883	0.0398	−0.1664	−0.0102*
0.07	0.08	0.3675	0.5383	−0.1708	0.1173	−0.4008	0.0591
0.07	0.09	0.2500	0.2833	−0.0333	0.0569	−0.1448	0.0782
0.08	0.09	0.4533	0.5067	−0.0533	0.0863	−0.2224	0.1158

6.3.2. Results: Data = very soft core, Model = very soft core

We look to other pairwise interaction process models, to confirm that the results observed in Section 6.3.1 are not an artefact of the Diggle process. First, we consider the very soft core model described in Section 2.2. Equivalent simulations are performed, and the results displayed in Table 8.

Table 8. Data model (VSC) and fitted model (VSC) parameter combinations. L and U indicate lower and upper 95% confidence limits respectively, on the difference in p-values from the model-based and design-based procedures. $*$ indicates a significant result in favour of the model-based procedure. Each mean is obtained from 12 simulations.

θ_1	θ_2	$E(p)$	$E(np)$	$E(p-np)$	$S.E.(\overline{p}-\overline{np})$	L	U
0.03	0.05	0.1208	0.1942	−0.0733	0.0536	−0.1783	0.0317
0.03	0.07	0.0117	0.0200	−0.0083	0.0055	−0.0191	0.0024
0.03	0.08	0.0100	0.0117	−0.0017	0.0017	−0.0049	0.0016
0.03	0.09	0.0100	0.0100	0.0000	0.0000	0.0000	0.0000
0.05	0.07	0.1700	0.2342	−0.0642	0.0675	−0.1966	0.0682
0.05	0.08	0.1442	0.2633	−0.1192	0.0492	−0.2155	−0.0228*
0.05	0.09	0.0217	0.0533	−0.0316	0.0168	−0.0646	0.0012
0.07	0.08	0.3325	0.4575	−0.1250	0.0795	−0.2808	0.0308
0.07	0.09	0.2475	0.3650	−0.1175	0.0706	−0.2558	0.0208
0.08	0.09	0.3150	0.3292	−0.0147	0.0722	−0.1561	0.1268

The absolute ESLs estimated for the design-based procedure are in general smaller than they were for the Diggle model in the previous section. We suggest that this is because the very soft core process has infinite range, whereas the Diggle interaction function is 1 beyond θ for any given θ.

6.3.3. Results: Data = Linear, Model = Linear

Again, we repeat the simulations, this time setting the data generating model and the fitted model to have the linear interaction function defined in Section 2.2. Broadly similar results are observed (see Table 9), namely that the parametric model-based procedure gives lower ESLs than the non-parametric design-based, and the parametric procedure detects differences at a reasonable level of significance when the non-parametric fails to do so.

Table 9. Data model (Linear) and fitted model (Linear) parameter combinations. L and U indicate lower and upper 95% confidence limits respectively, on the difference in p-values from the model-based and design-based procedures. $*$ indicates a significant result in favour of the model-based procedure. Each mean is obtained from 12 simulations.

θ_1	θ_2	$E(p)$	$E(np)$	$E(p-np)$	$S.E.(\overline{p}-\overline{np})$	L	U
0.03	0.05	0.3625	0.4208	−0.0583	0.1477	−0.3479	0.2312
0.03	0.07	0.1192	0.3225	−0.2033	0.0668	−0.3344	−0.0723*
0.03	0.08	0.0350	0.1067	−0.0717	0.0354	−0.1411	−0.0022*
0.03	0.09	0.0450	0.0700	−0.0250	0.0126	−0.0498	−0.0002*
0.05	0.07	0.2108	0.3892	−0.1783	0.0640	−0.3038	−0.0528*
0.05	0.08	0.0992	0.2083	−0.1092	0.0332	−0.1742	−0.0441*
0.05	0.09	0.1783	0.1983	−0.0200	0.0599	−0.1375	0.0975
0.07	0.08	0.4950	0.5342	−0.0392	0.0713	−0.1789	0.1006
0.07	0.09	0.5600	0.4800	0.0800	0.0873	−0.0911	0.2512
0.08	0.09	0.4120	0.4120	0.0000	0.0781	−0.1531	0.1531

6.3.4. Conclusions

We have thus shown that, under correct model specification, the parametric model-based procedure outperforms the non-parametric ANOVA, for a variety of classes of pairwise interaction process. There are situations where both procedures do equivalently well; however, when this is the case, the difference between the θ parameters in both groups tends to be non-trivial. Similarly, as we might expect, there are also situations where neither procedure detects a difference which exists; this is generally in cases where the difference between θs is very small.

6.4. Case 2: Model is incorrect

The alternative question of interest concerns the relative performance of the two procedures when the underlying model is mis-specified; specifically, does the model-based procedure give misleading results under this case scenario?

We specify again the combinations of models **A** and **B** which are used in this case in Table 3.

Table 3. (a) Data model and (b) fitted model combinations when model is incorrect.

(a)	(b)
Diggle	VSC
Linear	VSC
VSC	Linear

The parameter combinations we look at are given in Tables 4, 5 and 6.

6.4.1. Data Diggle, fitted model very soft core

We begin by looking at the results from the case where the data is simulated from a Diggle model, and the very soft core model is fitted. The examined parameter values are given in Table 4.

Table 4. Data model (Diggle) and fitted model (VSC) parameter combinations.

θ_1	0.03	0.05	0.07	0.09	0.11
θ_2	0.03	0.05	0.07	0.09	0.11

The plots in Figures 2.1 to 2.5 demonstrate no departure from Uniformity over the required $(0,1)$ range, for either the parametric or non-parametric procedure. This indicates that the correct result holds under the null hypothesis, when a very soft core model is fitted to data with the Diggle potential function, with parameter 0.03, 0.05 or 0.07. We suggest that the very soft core approximation is good as a consequence of the fact that the interaction functions are somehow *similar*, in that they are both smooth, and one can follow the shape of the other quite closely (*viz.* Figures 3.1 to 3.5).

We note, however, that for the case of $\theta = 0.09$, the parametric model-based procedure yields a marginally significant departure from Uniformity over the required range. However, the expected uniformity is once again observed for the case where $\theta = 0.11$, and so the departure in the $\theta = 0.09$ may be spurious.

Figure (2.1) $\theta_1 = \theta_2 = 0.03$. For parametric, K-S = 0.135 ($p = 0.4223$); for non-parametric, K-S = 0.165 ($p = 0.2024$).

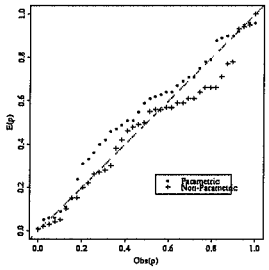


Figure (2.2) $\theta_1 = \theta_2 = 0.05$. For parametric, K-S = 0.100 ($p = 0.6623$); for non-parametric, K-S = 0.140 ($p = 0.5614$).

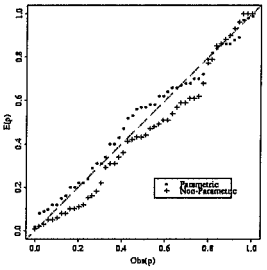


Figure (2.3) $\theta_1 = \theta_2 = 0.07$. For parametric, K-S = 0.090 ($p = 0.7793$); for non-parametric, K-S = 0.120 ($p = 0.4338$).

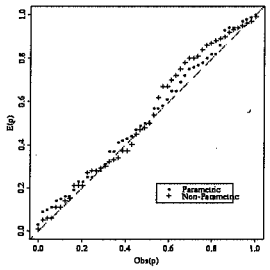


Figure (2.4) $\theta_1 = \theta_2 = 0.09$. For parametric, K-S = 0.180 ($p = 0.0688$); for non-parametric, K-S = 0.140 ($p = 0.2561$).

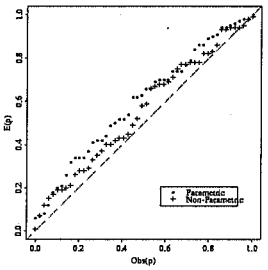


Figure (2.5) $\theta_1 = \theta_2 = 0.11$. For parametric, K-S = 0.100 ($p = 0.6623$); for non-parametric, K-S = 0.180 ($p = 0.0688$).

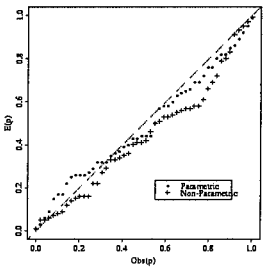


Figure 2. Observed versus Expected p-values under H_o, in the case where the data are simulated from the Diggle model, and a very soft core model is fitted, with the parameters indicated. In each case, «K-S» indicates the Kolmogorov-Smirnov test statistic (associated p-value in parentheses). (a) is the value in the model-based case, (b) in the design-based.

Figure (3.1) $\theta = 0.03$.

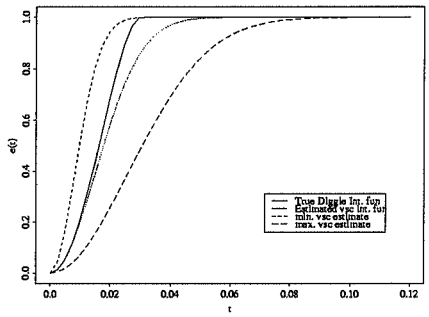


Figure (3.2) $\theta = 0.05$.

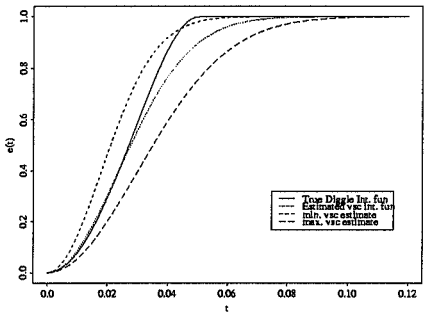


Figure (3.3) $\theta = 0.07$.

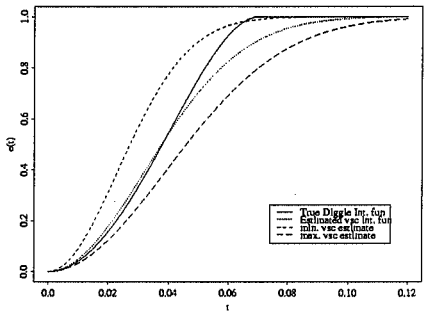


Figure (3.4) $\theta = 0.09$.

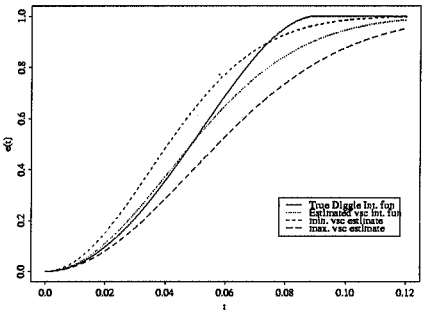


Figure (3.5) $\theta = 0.11$.

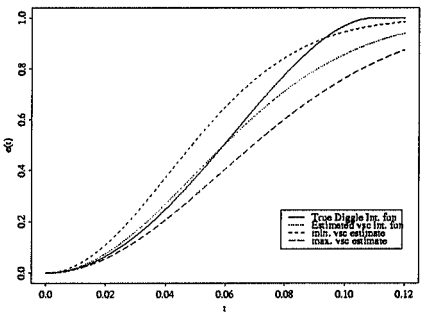


Figure 3. Interaction functions for true underlying process (Diggle potential with indicated parameter) and mean and range of fitted very soft core parameters, in each case from repeated simulations.

We conclude from this part of the study that the very soft core process provides a good model for data which has an underlying Diggle potential function.

6.4.2. Data Linear, fitted model very soft core

To provide a contrast, we now look at how well the very soft core process can model data for which the underlying interaction function is linear. We hypothesise that the model-based procedure is more likely to break down here than in the Diggle/very soft core case, as a consequence of the fact that the interaction functions are more dissimilar.

The parameter values we examine are given in Table 5.

Table 5. Data model (Linear) and fitted model (VSC) parameter combinations.

θ_1	0.03	0.05	0.06	0.08	0.10	0.11	0.12	0.13	0.14
θ_2	0.03	0.05	0.06	0.08	0.10	0.11	0.12	0.13	0.14

Again, we fix the n_{ij} and r_i. From repeated simulations of the experimental set-ups described above, we obtain the sets of p-values and associated Kolmogorov-Smirnov statistics illustrated in Figures 4.1 to 4.9.

The situation here is different in that, as suspected, the model-based procedure breaks down for certain parameter values. For the smaller parameters, clear departures from uniformity are demonstrated. As the parameters become larger, the p-values tend towards uniformity, however; we suggest that this is because as the parameter gets larger, the linear potential function becomes smoother in some sense, and so the very soft core model is better equipped to approximate the linear in this case (*viz.* Figures 5.1 to 5.9).

6.4.3. Data Very Soft Core, fitted model Linear

We look at the opposite situation, namely where the data is from a very soft core model and we attempt to fit a linear interaction function to it. The parameter values examined are displayed in Table 6.

Table 6. Data model (VSC) and fitted model (Linear) parameter combinations.

θ_1	0.10	0.11	0.12	0.13	0.14
θ_2	0.10	0.11	0.12	0.13	0.14

We repeat the procedure, and obtain the p-values and associated Kolmogorov-Smirnov statistics shown in Figures 6.1 to 6.5.

Figure (4.1) $\theta_1 = \theta_2 =$ 0.03. (a) K-S = 0.303 ($p = 0.0024$); (b) K-S = 0.094 ($p = 0.8860$).

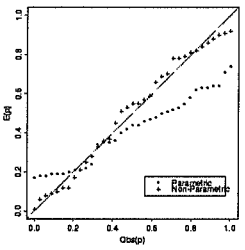


Figure (4.2) $\theta_1 = \theta_2 =$ 0.05. (a) K-S = 0.370 ($p = 0.0238$); (b) K-S = 0.223 ($p = 0.3854$).

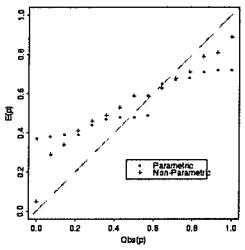


Figure (4.3) $\theta_1 = \theta_2 =$ 0.06. (a) K-S = 0.480 ($p = 0.0122$); (b) K-S = 0.290 ($p = 0.3067$).

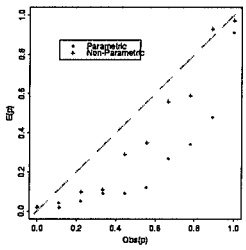


Figure (4.4) $\theta_1 = \theta_2 =$ 0.08. (a) K-S = 0.273 ($p = 0.0248$); (b) K-S = 0.204 ($p = 0.1678$).

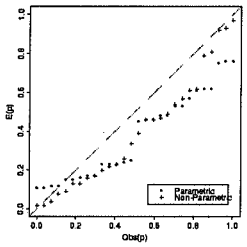


Figure (4.5) $\theta_1 = \theta_2 =$ 0.10. (a) K-S = 0.270 ($p = 0.0129$); (b) K-S = 0.1964 ($p = 0.1369$).

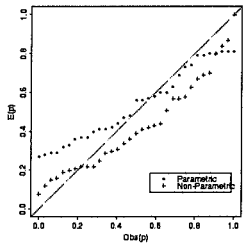


Figure (4.6) $\theta_1 = \theta_2 =$ 0.11. (a) K-S = 0.260 ($p = 0.0281$); (b) K-S = 0.090 ($p = 0.9502$).

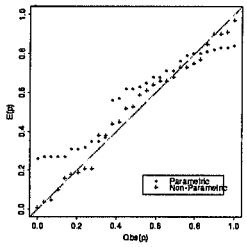


Figure (4.7) $\theta_1 = \theta_2 =$ 0.12. (a) K-S = 0.120 ($p = 0.4895$); (b) K-S = 0.0948 ($p = 0.7676$).

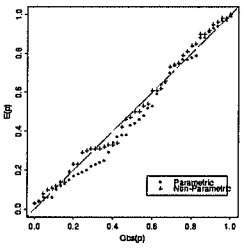


Figure (4.8) $\theta_1 = \theta_2 =$ 0.13. (a) K-S = 0.125 ($p = 0.3394$); (b) K-S = 0.147 ($p = 0.1721$).

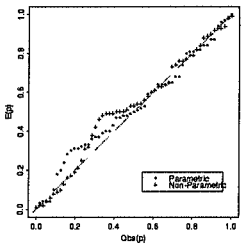


Figure (4.9) $\theta_1 = \theta_2 =$ 0.14. (a) K-S = 0.170 ($p = 0.1262$); (b) K-S = 0.092 ($p = 0.7955$).

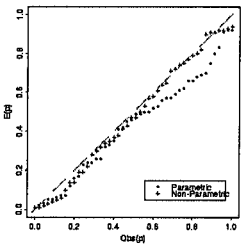


Figure 4. Observed versus Expected p-values under H_o, in the case where the data are simulated from the Linear model, and a very soft core model is fitted, with the parameters indicated. In each case, «K-S» indicates the Kolmogorov-Smirnov test statistic (associated p-value in parentheses). (a) is the value in the model-based case, (b) in the design-based.

Figure(5.1) $\theta = 0.03$.

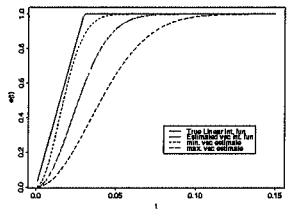


Figure (5.2) $\theta = 0.05$.

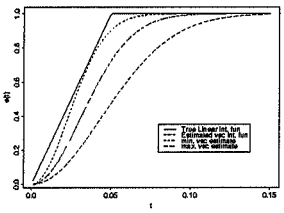


Figure (5.3) $\theta = 0.06$.

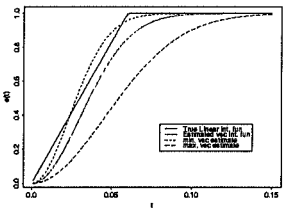


Figure (5.4) $\theta = 0.08$.

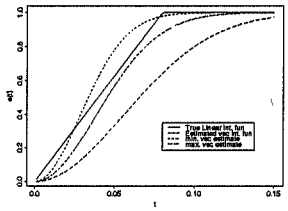


Figure (5.5) $\theta = 0.10$.

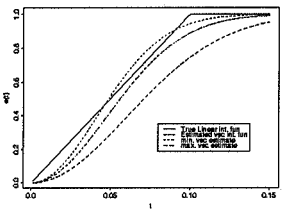


Figure (5.6) $\theta = 0.11$.

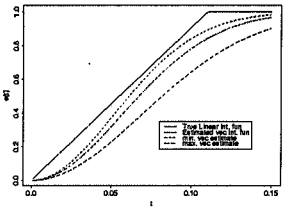


Figure (5.7) $\theta = 0.12$.

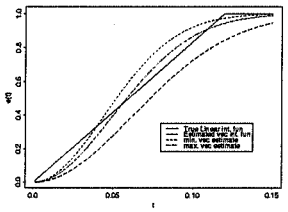


Figure (5.8) $\theta = 0.13$.

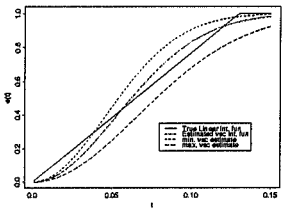


Figure (5.9) $\theta = 0.14$.

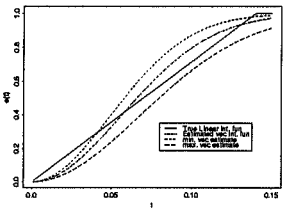


Figure 5. Interaction functions for true underlying process (Linear potential with indicated parameter) and mean and range of fitted very soft core parameters, in each case from repeated simulations.

Figure (6.1) $\theta_1 = \theta_2 = 0.10$. (a) K-S = 0.333 ($p = 0.1087$); (b) K-S = 0.343 ($p = 0.0912$).

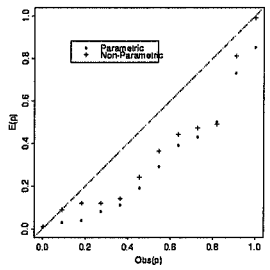


Figure (6.2) $\theta_1 = \theta_2 = 0.11$. (a) K-S = 0.1900 ($p = 0.1741$); (b) K-S = 0.141 ($p = 0.5013$).

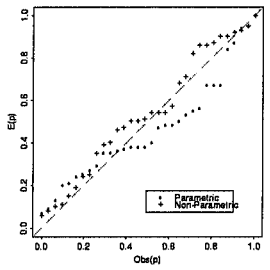


Figure (6.3) $\theta_1 = \theta_2 = 0.12$. (a) K-S = 0.340 ($p = 0.0145$); (b) K-S = 0.150 ($p = 0.7045$).

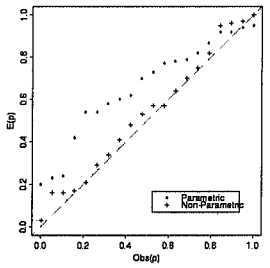


Figure (6.4) $\theta_1 = \theta_2 = 0.13$. (a) K-S = 0.237 ($p = 0.0294$); (b) K-S = 0.090 ($p = 0.9073$).

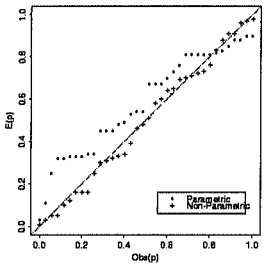


Figure (6.5) $\theta_1 = \theta_2 = 0.14$. (a) K-S = 0.302 ($p = 0.0366$); (b) K-S = 0.1995 ($p = 0.3284$).

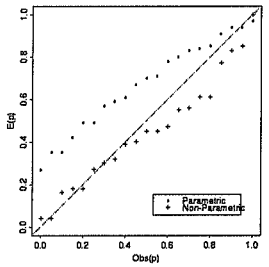


Figure 6. Observed versus Expected p-values under H_o, in the case where the data are simulated from the very soft core model, and a linear model is fitted, with the parameters indicated. In each case, «K-S» indicates the Kolmogorov-Smirnov test statistic (associated p-value in parentheses). (a) is the value in the model-based case, (b) in the design-based.

Figure (7.1) $\theta = 0.10$.

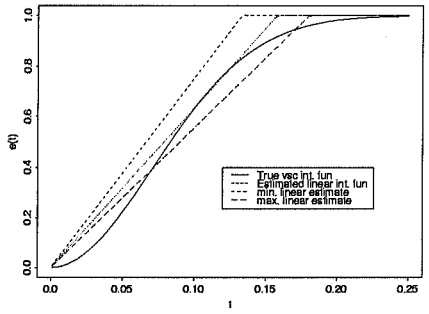


Figure (7.2) $\theta = 0.11$.

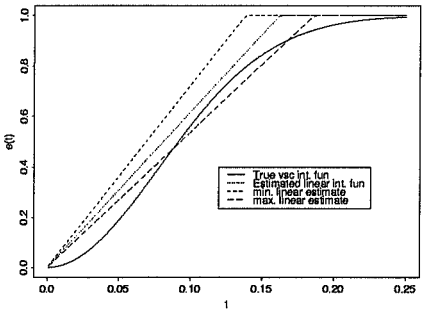


Figure (7.3) $\theta = 0.12$.

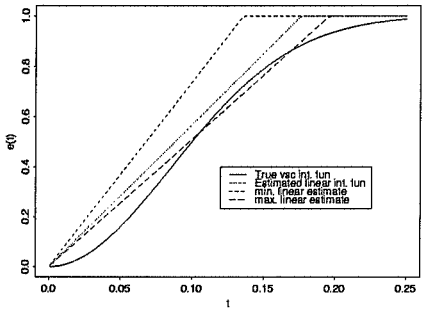


Figure (7.4) $\theta = 0.13$.

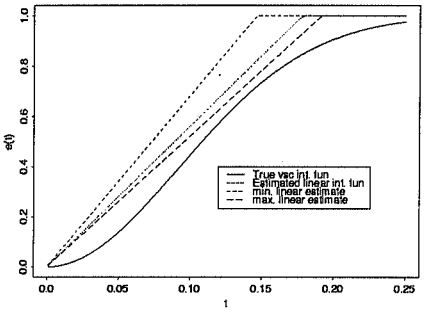


Figure (7.5) $\theta = 0.14$.

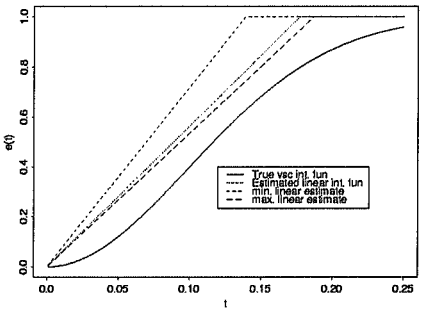


Figure 7. Interaction functions for true underlying process (Very soft core potential with indicated parameter) and mean and range of fitted linear parameters, in each case from repeated simulations.

The results can be explained once again by the relationship between the interaction functions in the two types of process. Now the parametric model-based procedure breaks down for the larger parameters, for which clear departures from uniformity are again demonstrated. Note again that as the parameter gets larger, the linear potential function is no longer equipped to fit the very soft core model (*viz.* Figures 7.1 to 7.5).

6.4.4. Conclusions

When both the true and the assumed model were either VSC or Diggle model, the parametric test led to empirical distributions of p-values which were well fitted by the uniform distribution on $(0,1)$; note that both models have differentiable interaction functions. Significant departures from uniformity arose when the non-differentiable linear model was used to simulate the data but VSC model was assumed, and *vice versa*. Note that the number of simulations contributing to Case 2 situation varied between 20 and 50 because of limitations on the available computing time; in general, simulations of pairwise interaction point processes become time-consuming as the strength of the interaction between points increases.

7. DISCUSSION

Pairwise interaction point processes are widely used as descriptive models for spatial pattern recognition. Approximate likelihood-based methods of inference are available for these models, but rely on computer-intensive Monte Carlo methods of inference. Whether this matters in practice depends on practical constraints which vary between applications, but it is a potentially limiting factor for data consisting of large numbers of replicate patterns. There are many interesting areas, such as medicine, biology or neuroanatomy, in which replicated point pattern data frequently arise. In this context, individual patterns often consist of spatial point coordinates, where it is straightforward to obtain large numbers of such patterns by repeated sectioning, either within or between experimental subjects.

In this study, we have confirmed our suspicion that the parametric model-based approach is more efficient at recognizing the spatial structure than an equivalent non-parametric design-based approach under correct model specification. In addition, we have ascertained that the performance of the parametric procedure under model misspecification depends upon the type of model being considered. Specifically, we have determined that the pairwise interaction process with Diggle potential and that with very soft core potential can generally model each other quite well. In contrast, neither of these two models can be applied successfully to data with linear potential function, and the reverse applies. In summary, when the model driving the data is unable to be approximated, in some sense, by the model being fitted, the procedure appears to break down.

ACKNOWLEDGEMENTS

This paper has been written after several visits of the author to Lancaster University, and it is a consequence of the fruitful discussions with Prof. Peter Diggle and Dr. Helen Wilson. The referees are also acknowledged for their suggestions on an earlier draft.

REFERENCES

Baddeley, A. J., Reid, S., Howard, V. & Boyde, A. (1985). «Unbiased estimation of particle density in the tandem scanning reflected light microscope». *Journal of Microscopy*, 138, 203-212.

Baddeley, A. J. & Møller, J. (1989). «Nearest-neighbour Markov point processes and random sets». *International Statistical Review*, 57, 89-121.

Baddeley, A. J., Moyeed, R., Howard, C. & Boyde, A. (1993). «Analysis of a three-dimensional point pattern with replication». *Journal of Applied Statistics*, 42, 641-646.

Besag, J. (1977). «Some methods of statistical analysis for spatial data». *Bulletin of the International Statistical Institute*, 47, 77-92.

Besag, J., Milne, R. & Zachary, S. (1982). «Point process limits of lattice processes». *Journal of Applied Probability*, 19, 210-216.

Cressie, N. A. (1993). *Statistics for spatial data*, Wiley, Revised edition.

Daley, D. J. & Vere-Jones, D. (1988). *Introduction to the theory of point processes.* New York: Springer.

Dempster, A. P. & Schatzoff, M. (1965). «Expected Significance Level as a sensitivity index for test statistics». *Journal of the American Statistical Association*, 60, 420-436.

Diggle, P. J. (1983). *Statistical analysis of spatial point patterns.* Academic Press, London.

Diggle, P. J. (1986). «Parametric and non-parametric estimation for pairwise interaction point processes». *Proceedings of the 1st World Congress of the Bernoulli Society.*

Diggle, P. J., Lange, N. & Beněs, F. (1991). «Analysis of variance for replicated spatial point patterns in clinical neuroanatomy». *Journal of the American Statistical Association*, 86, 415, 618-625.

Diggle, P. J., Fiksel, T., Grabarnik, P., Ogata, Y., Stoyan, D. & Tanemura, M. (1994). «On parameter estimation for pairwise interaction point processes». *International Statistical Review*, 62, 99-117.

Fiksel, T. (1984). «Estimation of parametrized pair potentials of marked and non-marked Gibbsian point processes». *Elektron. Inform. Kybernet.*, 20, 270-278.

Fiksel, T. (1988). obrir Estimation of interaction potentials of Gibbsian point processes». *Math. Operationsf. Statist. Ser. Statist.*, 19, 77-86.

Gates, D. J. & Westcott, M. (1986). «Clustering estimates for spatial point distributions with unstable potentials». *Ann. Inst. Statist. Math.*, 38, 123-135.

Geman, S. & Geman, D. (1984). «Stochastic relaxation, Gibbs distributions and the bayesian restoration of images». *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721-741.

Geyer, C. J. & Thompson, E. A. (1992). «Constrained Monte Carlo maximum likelihood for dependent data (with discussion)». *Journal of the Royal Statistical Society*, B 54, 657-699.

Geyer, C. J. & Møller, J. (1994). «Simulation and likelihood inference for spatial point processes». *Scandinavian Journal of Statistics*, 21, 359-373.

Hastings, W. K. (1970). «Monte Carlo sampling methods using Markov chains and their applications». *Biometrika*, 57, 97-109.

Jensen, J. L. (1993). «Asymptotic normality of estimates in spatial point processes». *Scandinavian Journal of Statistics*, 20, 97-109.

Jensen, J. L. & Møller, J. (1991). «Pseudolikelihood for exponential family models of spatial point processes». *Annals of Applied Probability*, 1, 445-461.

Kelly, F. P. & Ripley, B. D. (1976). «On Strauss' model for clustering». *Biometrika*, 63, 357-60.

Møller, J. (1992). «Discussion contribution». *Journal of the Royal Statistical Society*, B 54, 692-693.

Møller, J. (1993). «Spatial point processes and Markov chain Monte Carlo methods». Lecture at *Conference on Stochastic Processes and their Applications*, Amsterdam.

Moyeed, R. A. & Baddeley, A. J. (1991). «Stochastic approximation for the MLE of a spatial point process». *Scandinavian Journal of Statistics*, 18, 39-50.

Ogata, Y. & Tanemura, M. (1981). «Estimation of interaction potentials of spatial point patterns through the maximum likelihood procedure». *Annals of the Institute of Statistical Mathematics*, 33 B, 315-338.

Ogata, Y. & Tanemura, M. (1984). «Likelihood analysis of spatial point patterns». *Journal of the Royal Statistical Society*, B 46, 496-518.

Ogata, Y. & Tanemura, M. (1989). «Likelihood estimation of soft-core interaction potentials for Gibbsian point patterns». *Annals of the Institute of Statistical Mathematics*, 41 B, 583-600.

Penttinen, A. (1984). «Modelling interaction in spatial point patterns: parameter estimation by the maximum likelihood method». Number 7 in *Jyvaskyla Studies in Computer Science, Economics and Statistics.*

Preston, C. J. (1977). «Spatial birth-and-death processes». *Bull. Inst. Intern. Statist.*, 46, 371-391.

Ripley, B. D. (1977). «Modelling spatial patterns (with Discussion)». *Journal of Royal Statistical Society*, B 39, 172-212.

Ripley, B. D. (1981). *Spatial Statistics.* Wiley, New York.

Ripley, B. D. (1988). *Statistical inference for spatial processes.* Cambridge, Cambridge University Press.

Ripley, B. D. & Kelly, F. P. (1977). «Markov point processes». *J. London Math. Soc.*, 15, 188-192.

Strauss, D. J. (1975). «A model for clustering». *Biometrika*, 63, 467-475.

Takacs, R. (1986). «Estimator for the pair-potential of a Gibbsian point process». *Math. Operationsf. Statist. Ser. Statist.*, 17, 429-433.

Wilson, H. E. (1998). «Statistical analysis of replicated spatial point patterns». *Unpublished PhD. thesis*, University of Lancaster, U.K.

Wilson, H. E., Diggle, P. J. and Howard, C. V. (1998). «Methods for the analysis of replicated spatial point patterns in clinical neuroanatomy». *Advances in Applied Probability*, 30(2). Abstract in proceedings of 9th International Workshop in Stereology, Stochastic Geometry and Image Analysis, Comillas, October 1997.

QÜESTIIÓ, vol. 26, 1-2, p. 61-85, 2002

DETECCIÓN DE RASGOS EN IMÁGENES BINARIAS MEDIANTE PROCESOS PUNTUALES ESPACIALES MARCADOS

J. MATEU*
G. LORENZO**
Universitat Jaume I*

En este trabajo consideramos el problema de la detección de rasgos bajo la presencia de ruido en imágenes que tras un cierto tratamiento se reducen a binarias, por la presencia de dos tipos de elementos. Podemos encontrar ejemplos de este problema en la detección de minas por medio de imágenes de avión o satélite, en la búsqueda de rasgos en imágenes microscópicas de células, o en la caracterización de fallas en zonas de terremotos.
En primer lugar revisamos algunos métodos de detección jerárquicos basados en modelos probabilísticos, en los que los rasgos proceden de distribuciones normales multivariantes y el ruido surge según un proceso de Poisson espacial. Posteriormente, presentamos una nueva solución al problema mediante el uso de procesos puntuales espaciales marcados. Definimos un proceso puntual marcado en el que a cada localización se le asigna un par de marcas: las distancias al K-ésimo vecino más cercano y una variable dicotómica diferenciadora del rasgo frente al ruido. Estas distancias se modelizan como una mixtura de distribuciones cuyos parámetros se determinan mediante el algoritmo EM.
Finalmente, la nueva metodología es evaluada y contrastada sobre simulaciones y casos reales.

Detection of features in binary images by means of spatial marked point processes

Palabras clave: Algoritmo EM, mixturas de distribuciones, proceso de Poisson, procesos puntuales espaciales

Clasificación AMS (MSC 2000): 62M30, 60G55

* Autor para la correspondencia. Email: mateu@mat.uji.es. Fax: 964.728429
** Email: valentin@mat.uji.es
* Departamento de Matemáticas, Universitat Jaume I. Campus Riu Sec, E-12071 Castellón, Spain.
- Recibido en diciembre de 2000.
- Aceptado en febrero de 2002.

1. INTRODUCCIÓN

El análisis de imágenes difusas en las que bajo un cierto ruido se desarrolla algún fenómeno de interés es necesario en muchos contextos. Un ejemplo es el problema de detectar campos de minas (rasgos en general) en una superficie sobre la base de una imagen tomada por un avión de reconocimiento (Dasgupta & Raftery, 1998; Kaufman & Rousseeuw, 1990). Después de su procesamiento, tal imagen se reduce a un conjunto de objetos, algunos de los cuales serán minas y otros ruido (como por ejemplo, rocas u objetos metálicos). Los objetos identificados son pequeños y pueden ser representados por puntos sin pérdida de información relevante. La tarea del investigador será determinar si hay o no minas y dónde están localizadas (Cox, 1975). En este sentido partimos de la base de que las imágenes a analizar han sido procesadas de forma que nuestras unidades de trabajo consisten en localizaciones puntuales en una región del plano. Este problema puede verse también como el problema de buscar subregiones de mayor densidad dentro de una región donde se desarrolla un proceso puntual (Byers & Raftery, 1996, 1998). Un ejemplo típico se muestra en la Figura 1.

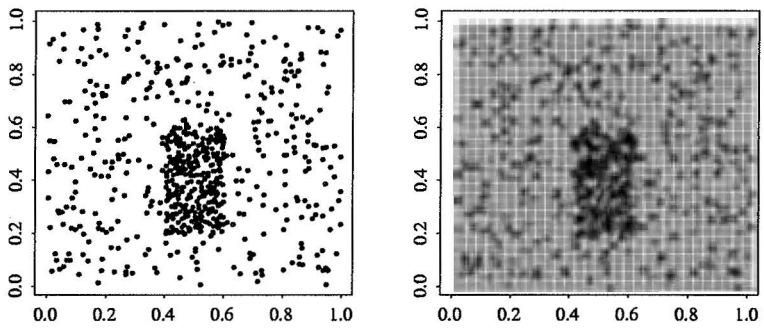


Figura 1. Datos simulados (ruido y rasgo) y retículo regular 30×30 sobre los datos.

Por lo dicho anteriormente, la teoría de procesos puntuales espaciales proporcionará el marco adecuado para la resolución de este problema. Un proceso puntual espacial es una colección de puntos distribuidos espacialmente en una región plana del espacio (Diggle, 1983). Los procesos puntuales constituyen una rama de la metodología estadística que es capaz de analizar dependencias espaciales entre las observaciones, que se suponen han sido generadas por algún mecanismo aleatorio y desconocido. Datos en la forma de localizaciones espaciales aparecen en muy diversos contextos en donde su análisis es esencial: localizaciones de árboles en bosques, centros de núcleos de células, epicentros de terremotos, localizaciones de ciudades, etc. Además, disponemos de un proceso, el proceso de Poisson, que sirve de marco con el que comparar cualquier tipo de estructura puntual. Este proceso se caracteriza por la ausencia de dependencia espa-

cial entre las observaciones y constituye el enlace entre la teoría de patrones espaciales y las técnicas usuales estadísticas.

En los últimos años, el problema de la detección de rasgos subyacentes en imágenes con ruido se ha centrado en el uso de métodos cluster basados en modelos probabilísticos (Banfield & Raftery, 1993; Dasgupta & Raftery, 1998; Fraley & Raftery, 1999) adaptando al caso de patrones puntuales espaciales algunas ideas surgidas en el contexto del análisis cluster. Aunque estos métodos funcionan adecuadamente en la práctica, muestran algunos inconvenientes: *a*) suelen ser matemáticamente bastante complicados; *b*) exigen suposiciones de Gaussianidad sobre los datos que no siempre son adecuadas; *c*) para su correcta aplicación necesitan la especificación de la forma y tamaño de los clusters.

Los inconvenientes antes mencionados motivan la propuesta de modelización que presentamos en este trabajo. Dado que las imágenes binarias pueden ser consideradas como un proceso puntual espacial bivariante, proponemos la definición de un proceso puntual marcado con dos tipos de marcas asociadas a cada posición espacial: una primera marca viene definida por las bien conocidas distancias al K-ésimo punto más cercano (Diggle, 1983; Collins, 1995) y una segunda marca que habrá que calcular y que identificará o clasificará la correspondiente localización espacial en rasgo o ruido. La novedad de esta propuesta no reside en el uso en sí de las distancias sino en la metodología que proponemos para la clasificación. Este método es totalmente original y lo más importante no hace ninguna suposición sobre la forma, número de rasgos o modelo probabilístico de los mismos.

El plan de este trabajo es el siguiente. En la sección 2 se definen los procesos puntuales espaciales. La sección 3 se centra en el método probabilístico, mientras que la nueva metodología propuesta viene en la sección 4. Finalmente, la sección 5 analiza datos simulados y reales.

2. PROCESOS PUNTUALES ESPACIALES

Un proceso puntual es un modelo estocástico dado por un conjunto de puntos $\{x_i\}_{i=1}^n$ en algún subconjunto $X \in \mathbb{R}^d$ (normalmente $d = 2$). Si las localizaciones contienen medidas o etiquetas, el proceso puntual será un proceso puntual marcado (Diggle, 1983; Penttinen, Stoyan & Henttonen, 1992; Cressie, 1993). Por ejemplo, sucesos pueden ser árboles en un bosque, ciudades en una región geográfica, o epicentros de terremotos. Las marcas correspondientes pueden ser, respectivamente, especies, tipos o diámetros de árboles, tamaños de ciudades o magnitudes de terremotos.

En general asumiremos que las localizaciones de los sucesos $\{x_i\}_{i=1}^n$ y sus marcas $\{Z(x_i)\}_{i=1}^n$ son realizaciones de algún proceso estocástico de la forma $\{Z(\mathbf{x}) : \mathbf{x} \in D\}$,

donde $\mathbf{x} = \{x_1, \ldots, x_n\}$ y tanto $Z(.)$ como D son aleatorios (Diggle, 1983; Cressie, 1993).

Un proceso puntual marcado se define de la siguiente forma. Sea (Ω, Λ, ρ) un espacio de medida, tal que $\rho(\Omega) < \infty$. Tomamos ahora $\Omega = X \times F$ con $X \subset \mathbb{R}^d$ y $F \subset \mathbb{R}$, $\Lambda = \chi \times \xi$, donde χ es la σ-álgebra de Borel de X y ξ son los conjuntos de Borel de F y, finalmente, $\rho = \nu \times \pi$, donde ν es una medida finita, no atómica en X (típicamente la medida Lebesgue) y π es una medida en F correspondiendo a la distribución de marcas. El espacio exponencial (Cressie, 1993) se denota por Ω_e y se equipa con la correspondiente σ-álgebra Λ_e. Un elemento $(\mathbf{x}, Z(\mathbf{x})) \in \Omega_e$ se toma como un patrón puntual espacial. Los procesos puntuales marcados se pueden también definir a través de medidas aleatorias (Cressie, 1993).

Dos conceptos importantes de los procesos puntuales son la estacionariedad y la isotropía. Un proceso puntual es *estacionario* si es estadísticamente invariante bajo traslaciones, es decir, $N = [\mathbf{x}, Z(\mathbf{x})]$ y $N_t = \{[\mathbf{x}+t; Z(\mathbf{x}+t)]\}$ tienen la misma distribución para todo t. Un proceso puntual es *isotrópico* si su distribución es invariante bajo rotaciones, es decir, N y $N_r = \{[r\mathbf{x}; Z(r\mathbf{x})]\}$ tienen la misma distribución para cada rotación r con respecto al origen. Si el proceso es a la vez estacionario e isotrópico se conoce como invariante frente al movimiento.

Un proceso puntual espacial de Poisson es aquel proceso en el que los puntos son independientes y localizados aleatoriamente en la región de estudio. Este proceso constituye el patrón básico de comparación (Cressie, 1993).

3. MÉTODOS BASADOS EN MODELOS PROBABILÍSTICOS

3.1. Introducción

Las técnicas de detección basadas en modelos probabilísticos se basan en la suposición de que los datos vienen generados por mixturas de distribuciones de probabilidad, distinguiendo, cuando sea posible, entre las distribuciones de los rasgos frente a las del ruido (McLachlan & Basford, 1988; Banfield & Raftery, 1993; Bensmail, Celeux, Raftery & Robert, 1994).

Estos métodos consideran a priori una información importante para el desarrollo del modelo puesto que determinará de un modo decisivo elementos como la cantidad, forma, tamaño y orientación de los clusters. Se trata de una colección de técnicas paramétricas en las que suponemos que los datos proceden de una distribución Normal Multivariante (NMV) o modelo Gaussiano. De los dos parámetros a considerar en dicho modelo (vector de medias y matriz de varianzas-covarianzas), tiene especial relevancia la descomposición espectral de la matriz de varianzas-covarianzas Σ. Esta matriz toma

diferentes expresiones dependiendo del cluster o grupo donde se defina. De esta forma, una reparametrización de esta matriz en el grupo o cluster k-ésimo toma la siguiente forma

$$\Sigma_k = \lambda_k D_k A_k D_k', \tag{1}$$

donde

1. λ_k denota el primer valor propio de la descomposición espectral de Σ_k, y controla el *volumen* de los rasgos (clusters),
2. D_k es la matriz de vectores propios de Σ_k, y controlará la *orientación* de los rasgos,
3. A_k es la matriz diagonal cuyos elementos son proporcionales a los valores propios, y controla la *forma* de los rasgos en el siguiente sentido: si $A_k = \text{diag}\{\alpha_{1k}, \cdots, \alpha_{pk}\}$ con $1 = \alpha_{1k} \geq \alpha_{2k} \geq \cdots \geq \alpha_{pk} > 0$, entonces si los α_{jk} son de magnitud similar, el k-ésimo rasgo tenderá a ser esférico, mientras que si $\alpha_{2k} \ll 1$, estará concentrado sobre una línea; si $\alpha_{2k} \approx 1$ y $\alpha_{3k} \ll 1$, estará concentrado sobre un plano bidimensional en un espacio p-dimensional y así sucesivamente.

Las características de forma, volumen y orientación de las distribuciones son estimadas a partir de los datos y pueden variar o mantenerse fijas entre los rasgos. Así, por ejemplo, si $\Sigma_k = \sigma^2 I = \lambda I$ para todo k, siendo I la matriz identidad, el método lleva a detectar rasgos hiperesféricos con la misma varianza. Si $\Sigma_k = \Sigma$, para todo k, obtenemos el método de varianza constante (Wolfe, 1970; Fraley & Raftery, 1998).

3.2. Adaptación al caso de procesos puntuales espaciales

En este caso, el modelo considerado será un modelo Gaussiano para los rasgos y un modelo de Poisson para el ruido (Gower, 1967; Marriott, 1975; Banfield & Raftery, 1993). La detección de los rasgos se realiza en dos etapas:

1. En la primera etapa, mediante un modelo de cluster jerárquico se proporciona una primera estimación de los rasgos.

   Para la primera etapa, partimos de la ya conocida descomposición espectral de la matriz Σ de varianzas-covarianzas, $\Sigma_k = \lambda_k D_k A_k D_k'$, pero adaptándola al caso que nos ocupa, es decir, con patrones puntuales espaciales (McLachlan & Basford, 1988). En este caso, la forma de los rasgos es la misma, pero su volumen y orientación son diferentes, es decir, $A_k = \text{diag}\{1, \alpha\}$, $k = \{1, \ldots, G\}$ donde $\alpha < 1$ (ya que en un proceso puntual espacial hay dos dimensiones, $d = 2$) y G indica el número de rasgos a detectar. Además, $\alpha = \frac{\lambda_2}{\lambda_1}$, $(\lambda_1 > \lambda_2)$. Cuando α es mucho más pequeño que uno, los rasgos que resultan tenderán a ser largos y estrechos, mientras que si se acerca a uno, los rasgos tenderán a ser de una forma más circular. Por esta razón α se llama el *parámetro forma*.

2. En la segunda etapa estos clusters-rasgos se refinan utilizando el algoritmo E. M. (Dempster *et al.*, 1977) como método para la obtención de la estimación máximo verosímil en presencia de datos faltantes. Supongamos una población con distribución una mixtura de densidades de la forma

$$f(X;\theta) = \sum_{k=1}^{G} \pi_k f_k(X;\theta) \tag{2}$$

donde:

(*a*) $f_k(X;\theta)$ se distribuye como una $NMV(\mu_k, \Sigma_k)$

(*b*) $\sum_{k=1}^{G} \pi_k = 1$

Así, para n observaciones de esta mixtura de distribuciones, definimos un proceso puntual espacial marcado de la forma $\text{pem}_i = (x_i, z_i)$, donde $x_i = (x_{i1}, x_{i2})$ son las localizaciones espaciales originales y $z_i = (z_{i1}, \ldots, z_{iG})$, con

$$z_{ik} = \begin{cases} 1 & \text{si la } i\text{-ésima observación está en el } k\text{-ésimo rasgo} \\ 0 & \text{en otro caso} \end{cases}$$

las marcas añadidas.

El vector de z_i sigue una distribución multinomial con parámetros $(1;\pi_1, \ldots, \pi_G)$. Esto permite calcular la función de verosimilitud completa

$$L(\text{pem};\theta) = \sum_{i=1}^{n} \sum_{k=1}^{G} z_{ik} \{\log \pi_k + \log f_k(x_i;\theta)\} \tag{3}$$

En el primer paso del algoritmo EM (paso E) hay que calcular el estimador para las etiquetas que hemos puesto a cada observación $\hat{z}_{ik} = p_{ik} = E(z_{ik} | x_1, \ldots, x_n; \theta)$, la cual es la probabilidad a posteriori de que x_i esté en el rasgo k-ésimo. En el paso siguiente (paso M) se obtienen los estimadores máximo verosímiles de θ y π. Para ello, se maximiza la versión estimada de (3) dada por

$$L^*(\text{pem};\theta) = \sum_{i=1}^{n} \sum_{k=1}^{G} \hat{z}_{ik} \{\log \pi_k + \log f_k(x_i;\theta)\} \tag{4}$$

El parámetro de forma α se estimará por máxima verosimilitud a partir del modelo de mixturas utilizado por el algoritmo EM. Supongamos que tenemos una mixtura de distribuciones Gaussianas, satisfaciendo $\Sigma_k = \lambda_k D_k A_k D_k'$, con $A_k = A = \text{diag}\{1, \alpha, \ldots, \alpha\}$ $(k = 1, \ldots, G)$, y un ruido que se distribuye como un proceso de Poisson homogéneo.

Sea n_0 el número de puntos del ruido, d la dimensión de los datos (usualmente $d = 2$) y $\widehat{\Sigma}_k$ la matriz de varianzas-covarianzas muestral para el rasgo k-ésimo y cuya descomposición espectral es $\widehat{\Sigma}_k = L_k \Omega_k L_k'$. Si comenzamos a partir de la función de verosimilitud maximizada, asumiendo que α es conocida,

$$2\log L = -(n-n_0)(d\log(2\pi) + d(1-\log(d)) + \log(|A|)) - d\sum_{k=1}^{G} n_k \log\left(\frac{tr(\Omega_k A^{-1})}{n_k}\right) - 2n_0\log(V) \tag{5}$$

donde V es el volumen ocupado por los datos en R^d. Una buena definición de V viene dada por el volumen del rectángulo más pequeño cuyos lados son paralelos a los ejes coordenados, que contienen a los datos

$$V = \prod_{j=1}^{d} \left(\max_{i=1,\ldots,n} \{x_{ij}\} - \min_{i=1,\ldots,n} \{x_{ij}\} \right) \tag{6}$$

Finalmente, maximizando la probabilidad (5), con respecto a α, obtenemos

$$\sum_{k=1}^{G} \left[\frac{n_k}{n-n_0}\right] \frac{w_{k2} + \cdots + w_{kd}}{\alpha w_{k1} + \cdots + w_{kd}} = 1 - \frac{1}{d} \tag{7}$$

donde los w_{kj} representan los elementos de la matriz Ω_k.

La ecuación (7) es un polinomio en α de orden G con G raíces distintas como máximo. Sin embargo, sólo nos interesan las que se encuentran entre $[0, 1]$. En el caso particular en el que sólo hay un grupo ($G = 1$),

$$\widehat{\alpha} = \frac{w_2 + \cdots + w_d}{(d-1)w_1}$$

Por lo general, (7) se resuelve mediante un sistema de búsqueda de posibles valores de α en entornos de retículos.

Finalmente, se puede determinar el número óptimo de rasgos representando cada elección como un modelo estadístico y comparándolos mediante el factor Bayes. Hay una serie de razones que lo avalan (contempla la distribución asintótica en modelos mixtos, considera modelos que pueden ser comparados por métodos frecuencialistas, utiliza los p-valores) (Sokal & Michener, 1958). Además, gracias al algoritmo EM, podemos encontrar la función de verosimilitud maximizada de las mixturas y utilizar un método BIC (Bayesian Information Criterion) (Smith & Spiegelhalter, 1980; Kass & Raftery, 1995).

Como puede verse esta metodología, aunque se ha demostrado eficaz (Dasgupta & Raftery, 1998), no deja de tener un cierto fundamento matemáticamente complicado

con fuertes suposiciones sobre los modelos de probabilidad de los datos y no resulta sencilla llevarla a la práctica. Cierto es que con la ayuda del software MCLUST (Fraley & Raftery, 1999) la aplicación se hace más llevadera. Obsérvese que una colección de localizaciones espaciales en el plano equivale a disponer de dos variables sobre las que es susceptible la aplicación de las técnicas tradicionales de cluster jerárquico. Y esto es lo que se hace en una primera etapa. Posteriormente se refina el resultado obtenido mediante la reparametrización de la matriz de varianzas-covarianzas en términos del parámetro de forma α. Además, se aplica el algoritmo EM para mejorar la clasificación, la cual se lleva a cabo obteniendo los valores de la variable clasificadora z_{ik}.

4. MÉTODO DE DETECCIÓN BASADO EN DISTANCIAS AL *K*-ÉSIMO VECINO MÁS PRÓXIMO

En esta sección proponemos un nuevo método de detección de rasgos en imágenes con ruido basado en la construcción de procesos puntuales marcados donde las marcas provienen a partir del cálculo de cierto tipo de distancias entre los sucesos. Se trata de un método no paramétrico en que no hace falta establecer ninguna suposición sobre la Gaussianidad de los datos, forma o número de rasgos. En este sentido consideramos que se trata de un método de mayor aplicabilidad y sencillez.

La originalidad de esta propuesta no se basa en el uso de las distancias al K-ésimo vecino más próximo pues bien es sabido que ya fueron definidas por Diggle (1983) y posteriormente reanalizadas por Collins (1995), sino en el método de clasificación. La idea subyacente es formar un proceso puntual espacial marcado donde una de las marcas viene definida por este tipo de distancias (donde haremos uso de las propiedades que de ellas se derivan) y la otra marca actuará de datos faltantes a ser estimados para proveer la clasificación. Por tanto las dos secciones siguientes tratan estos dos puntos básicos.

4.1. Construcción del proceso puntual marcado

Asumiremos que el ruido está distribuido como un *proceso puntual de Poisson homogéneo* y el rasgo también está distribuido como un *proceso de Poisson* restringido a la imagen de definición del mismo, y que puede tener cualquier forma.

Construimos el siguiente proceso puntual marcado. Sea $N = \{x_i, y_i, d_i, \delta_i\}_{i=1}^{n}$ con (x_i, y_i) las localizaciones originales, d_i las distancias de cada localización a su K-ésimo vecino más cercano (Collins, 1995) y $\delta_i \in \{0, 1\}$. Además, $\delta_i = 1$, si el i-ésimo punto pertenece a un rasgo y $\delta_i = 0$ si se encuentra en el ruido.

Veamos algunas propiedades de este tipo de distancias. Como disponemos de un proceso de Poisson homogéneo, podemos encontrar la función de distribución de la distancia

al K-ésimo vecino más próximo, D_K, de un punto aleatorio en el proceso

$$P(D_K \geq d) = \sum_{k=0}^{K-1} \frac{e^{-\lambda\pi d^2}\left(-\lambda\pi d^2\right)^k}{k!} = 1 - F_{D_K}(d) \tag{8}$$

siendo λ la intensidad del proceso o número de puntos por unidad de área.

Esta expresión generaliza la obtenida por Diggle (1983) para distancias al primer vecino más próximo. Si D_K es mayor que d entonces debe ser uno de los $0, 1, \ldots, K-1$ puntos en el círculo. Pero además, sabiendo que la función de densidad $f_{D_K}(d)$ viene dada por la derivada de F_{D_K}, podemos desarrollarla para obtener

$$\begin{aligned}
f_{D_K}(d) &= \frac{\partial F_{D_K}(d)}{\partial d} = \frac{d\left(1 - \frac{e^{-\lambda\pi d^2}\left(-\lambda\pi d^2\right)^k}{k!}\right)}{dd} \\
&= -\sum_{k=0}^{K-1}\left[\frac{e^{-\lambda\pi d^2}\left(-2\lambda\pi d\right)\left(\lambda\pi d^2\right)^k}{k!} + \frac{e^{-\lambda\pi d^2} k\left(\lambda\pi d^2\right)^{k-1}\left(2\lambda\pi d\right)}{k!}\right] \\
&= -\sum_{k=0}^{K-1} \frac{e^{-\lambda\pi d^2} 2\lambda^k \pi^k d^{2k-1}\left(-\lambda\pi d^2 + k\right)}{k!} \\
&= e^{-\lambda\pi d^2} 2 \sum_{k=0}^{K-1} \frac{(\lambda\pi)^k d^{2k-1}\left(\lambda\pi d^2 - k\right)}{k!} \\
&= e^{-\lambda\pi d^2} 2\left[\sum_{k=0}^{K-1} \frac{(\lambda\pi)^{k+1} d^{2k+1}}{k!} - \sum_{k=0}^{K-1} \frac{(\lambda\pi)^k d^{2k-1} k}{k!}\right] \\
&= e^{-\lambda\pi d^2} 2\left[\sum_{k=0}^{K-1} \frac{\left(\lambda\pi d^2\right)^k \lambda\pi d}{k!} - \sum_{k=0}^{K-1} \frac{\left(\lambda\pi d^2\right)^k d^{-1} k}{k!}\right] \\
&= e^{-\lambda\pi d^2} 2\left[\lambda\pi d \sum_{k=0}^{K-1} \frac{\left(\lambda\pi d^2\right)^k}{k!} - \left(\lambda\pi d^2\right) \sum_{k=0}^{K-1} \frac{\left(\lambda\pi d^2\right)^{k-1}}{(k-1)!}\right] \\
&= \frac{e^{-\lambda\pi d^2} 2\left(\lambda\pi\right)^K d\left(d^2\right)^{K-1}}{(K-1)!} = \frac{e^{-\lambda\pi d^2} 2\left(\lambda\pi\right)^K d^{2K-1}}{(K-1)!} \qquad (9)
\end{aligned}$$

La función de densidad anterior (9) corresponde a la función de densidad de una variable aleatoria con una distribución Gamma transformada, $Y \sim \Gamma(K, \lambda\pi)$ donde $Y = (D_K)^2$. Concretamente es un ejemplo de una *Distribución Gamma Generalizada* (Stacy, 1962).

Además, dada esta estructura, encontrar el estimador máximo verosímil de la razón λ es relativamente sencillo. Para el conjunto de observaciones de distancias d_i, la función de verosimilitud toma la forma

$$L(d_i;\lambda) = \prod_{i=1}^{n} f_{D_K}(d_i;\lambda) = \frac{e^{-\lambda\pi\sum_{i=1}^{n} d_i^2}\left[2(\lambda\pi)^K\right]^n \prod_{i=1}^{n} d_i^{2K-1}}{(K-1)!}$$

Tomando logaritmos y derivando obtenemos

$$\begin{aligned}\ln L(d_i;\lambda) = & -\lambda\pi\sum_{i=1}^{n} d_i^2 + n[\ln 2 + K[\ln\lambda + \ln\pi]] + \ln\left[\prod_{i=1}^{n} d_i^{2K-1}\right] \\ & - \ln((K-1)!)\end{aligned}$$

$$\frac{d\ln L(d_i;\lambda)}{d\lambda} = -\pi\sum_{i=1}^{n} d_i^2 + \frac{nK}{\lambda},$$

de donde podemos obtener el estimador máximo verosímil para la intensidad λ, por medio de la expresión

$$\widehat{\lambda} = \frac{nK}{\pi\sum_{i=1}^{n} d_i^2}. \tag{10}$$

4.2. Determinación de los parámetros de clasificación. Algoritmo EM

Teniendo en cuenta que el modelo más simple está formado por un rasgo y por un ruido (el resto de observaciones), ambos distribuidos como dos procesos de Poisson homogéneos espaciales, podemos decir que la distribución de D_K es una mixtura de las dos distribuciones Gamma Generalizadas anteriormente comentadas

$$D_K \simeq p\Gamma^{(1/2)}(K,\lambda_1\pi) + (1-p)\Gamma^{(1/2)}(K,\lambda_2\pi) \tag{11}$$

siendo $\Gamma^{(1/2)}(a,b)$ la distribución de la Gamma Generalizada dada en (9).

El objetivo ahora es la obtención de los valores de las marcas de clasificación δ_i, $\forall i \in \{1,\ldots,n\}$ y también de los parámetros λ_1 ,λ_2 y p. Para ello haremos uso del algoritmo EM (Dempster *et al.*, 1977; Celeux & Govaert, 1992; Hathaway, 1986). Los dos pasos de este algoritmo son:

1. **Paso E.** En este paso obtendremos el valor esperado de las variables δ_i, $E(\widehat{\delta}_i(t+1))$, a partir de los valores estimados de $p(t),\lambda_1(t),\lambda_2(t)$ en la iteración anterior. Se

trata de encontrar la función $Q(p,\lambda_1,\lambda_2|X,d) = E(\ln(L(p,\lambda_1,\lambda_2|d_i,\delta_i))$, que para la localización i-ésima y la intensidad λ_1 toma la forma

$$Q_i(p,\lambda_1,\lambda_2|d_i,\delta_i) = \frac{p(d_i,\delta_i|\lambda_1)}{p(d_i,\delta_i|\lambda_1)p(\lambda_1)+p(d_i,\delta_i|\lambda_2)p(\lambda_2)}$$

$$= \frac{\widehat{p}(t)\,f_{D_k}(d_i;\widehat{\lambda}_1(t))}{\widehat{p}(t)\,f_{D_k}(d_i;\widehat{\lambda}_1(t))+(1-\widehat{p}(t))\,f_{D_k}(d_i;\widehat{\lambda}_2(t))} \tag{12}$$

2. **Paso M.** En este paso se obtienen los estimadores para λ_1 y λ_2 que en la iteración posterior se utilizarán para realizar el paso E. La probabilidad p se puede estimar por

$$\widehat{p}(t+1) = \sum_{i=1}^{n} \frac{\widehat{\delta}_i(t+1)}{n}. \tag{13}$$

Teniendo en cuenta que la función de verosimilitud viene dada por

$$L = \prod_{i=1}^{n} f_{D_K}(d_i;\lambda),$$

para λ_1 tenemos que

$$L = \prod_{i=1,\delta_i=1}^{n} f_{D_K}(d_i;\lambda) = \prod_{i=1,\delta_i=1}^{n} \frac{e^{-\lambda_1\pi d_i^2}2\,(\lambda_1\pi)^K d_i^{2K-1}}{(K-1)!}$$

$$= \frac{e^{\sum_{i=1,\delta_i=1}^{n}\lambda_1\pi d_i^2\delta_i}2^{\sum_{i=1,\delta_i=1}^{n}\delta_i}(\lambda_1\pi)^{\sum_{i=1,\delta_i=1}^{n}\lambda_1\pi d_i^2\delta_i}\prod_{i=1,\delta_i=1}^{n} d_i^{2K-1}\delta_i}{[(K-1)!]^{\sum_{i=1,\delta_i=1}^{n}\delta_i}},$$

con logaritmo

$$\ln L = \sum_{i=1}^{n}\lambda_1\pi d_i^2\delta_i + \sum_{i=1}^{n}\delta_i(\ln 2) + k\left(\sum_{i=1}^{n}\delta_i\right)\ln(\lambda_1\pi)$$

$$+\sum_{i=1}^{n}\ln\left(d_i^{2K-1}\delta_i\right) - \sum_{i=1}^{n}\delta_i\ln(K-1)!$$

Finalmente, derivando obtenemos

$$\frac{\partial \ln L}{\partial\lambda_1} = \frac{\partial}{\partial\lambda_1}\left[-\lambda_1\pi\sum_{i=1}^{n} d_i^2\delta_i + K\left(\sum_{i=1}^{n}\delta_i\right)[\ln\lambda_1+\ln\pi]\right]$$

$$= -\pi\sum_{i=1}^{n} d_i^2\delta_i + K\left(\sum_{i=1}^{n}\delta_i\right)\frac{1}{\lambda_1},$$

de donde se obtiene el estimador para λ_1 (escrito ya en notación iterativa)

$$\widehat{\lambda}_1(t+1) = \frac{K\sum_{i=1}^{n}\widehat{\delta}_i(t+1)}{\pi\sum_{i=1}^{n}d_i^2\widehat{\delta}_i(t+1)}. \tag{14}$$

Analogamente, el estimador para λ_2 viene dado por

$$\widehat{\lambda}_2(t+1) = \frac{K\sum_{i=1}^{n}\left(1-\widehat{\delta}_i(t+1)\right)}{\pi\sum_{i=1}^{n}d_i^2\left(1-\widehat{\delta}_i(t+1)\right)}. \tag{15}$$

Finalmente, el método de clasificación consistirá en clasificar una observación como rasgo si su δ_i final se encuentra más cerca del 1 y como ruido en otro caso.

5. SIMULACIONES Y APLICACIONES

En esta sección ilustramos la metodología presentada en este trabajo sobre unos datos simulados y reales. El objetivo es mostrar que la propuesta de clasificación para la detección de rasgos frente a ruido basada en la construcción de procesos marcados con marcas definidas a partir de un cierto tipo de distancias funciona correctamente en la práctica. Por tanto constituirá una buena alternativa al uso de las técnicas de cluster jerárquico. Quede claro que, como bien mostramos en primer lugar en esta sección, las técnicas de cluster jerárquico funcionan correctamente en la práctica (ver Figura 2) a costa de la fuerte exigencia sobre la distribución de los datos y no se trata tanto de comparar ambas metodologías como de mostrar que nuestra alternativa produce al menos los mismos resultados bajo un menor costo de exigencia probabilística y también computacional.

El primer conjunto de datos que se utiliza (Figura 1) consta de 700 puntos simulados, 300 de los cuales son rasgo y corresponderían a un proceso aleatorio en la región $[0.4, 0.6] \times [0.2, 0.6]$, mientras que los 400 puntos restantes son ruido, también definidos como un proceso aleatorio, pero esta vez en toda la región $[0,1] \times [0,1]$. En esta misma figura, en la segunda gráfica aparece un retículo regular de 900 celdas (30×30). Esta malla colocada sobre los datos nos permitirá calcular posteriores distancias.

Antes de analizar el comportamiento de nuestra metodología, vamos a clasificar y separar el rasgo del ruido por medio de técnicas jerárquicas basadas en modelos probabilísticos. Para ello hacemos uso del software MCLUST (Fraley & Raftery, 1999). Los resultados de una primera aplicación de este software vienen representados en la Figura 2. Observamos que los puntos representados por círculos se encuentran aglutinados en la zona donde se encuentra el rasgo; el ruido viene representado por las localizaciones en forma de cuadrados. En general, la clasificación obtenida por este método se ajusta

adecuadamente a los datos simulados. Además, un análisis de los valores BIC confirmó que el número óptimo de rasgos a detectar era 1. Tal y como hemos comentado en la sección 3, este procedimiento funciona bien pero tiene fuertes exigencias sobre la distribución de probabilidad de los datos y su uso es complicado en la práctica.

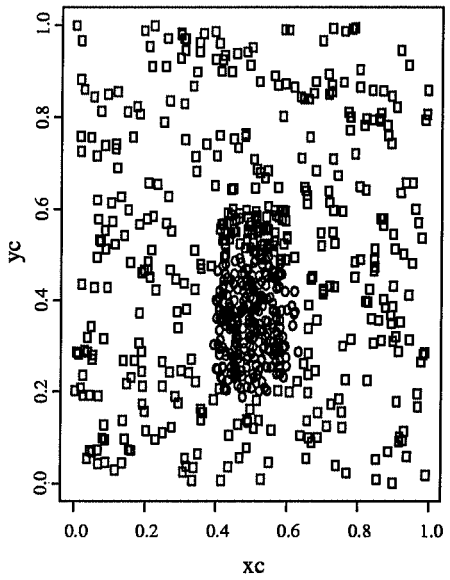


Figura 2. Método jerárquico probabilístico sobre los datos simulados.

En segundo lugar, y como motivación para el uso de nuestra metodología, aplicamos sobre los datos simulados diferentes tipos de distancias, las definidas al K-ésimo vecino más cercano (D_K) y otras definidas desde un retículo regular impuesto sobre los datos (ver Figura 1) al K-ésimo vecino de entre las observaciones reales más cercano (DG_K) (Diggle, 1983; Collins, 1995). En este caso nuestro único objetivo es ver si estas distancias pueden detectar de forma intuitiva diferencias entre rasgo y ruido.

En la Figura 3 se muestran histogramas de frecuencias absolutas para las distancias D_K para distintos valores de K. Se observa una clara bimodalidad al aumentar el valor de K (es decir, al aumentar el número del vecino más próximo que se toma de referencia para calcular la distancia). Las distancias alrededor de la primera moda (distancias más pequeñas), corresponden a los valores de las distancias de los puntos del rasgo, mientras que la segunda moda corresponde a los valores de las distancias de los puntos del ruido. De esta forma las distancias la K-ésimo vecino más cercano pueden ser utilizadas para discriminar entre rasgo y ruido.

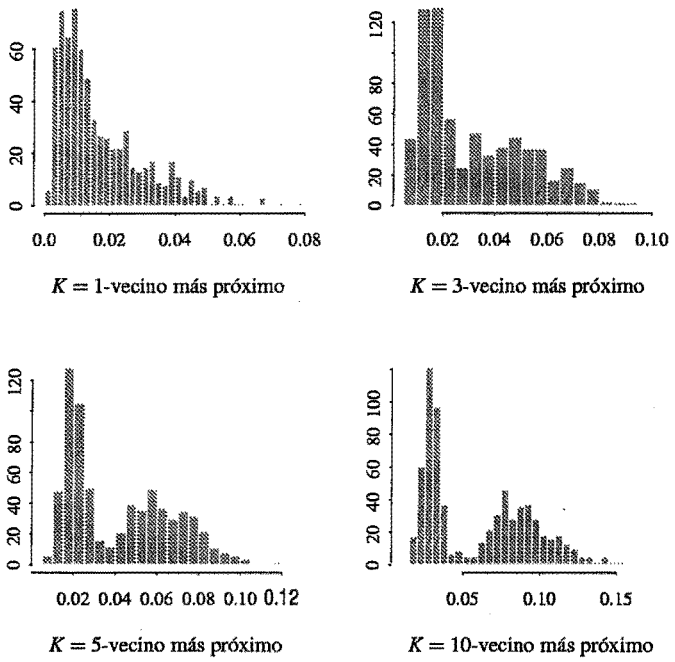


Figura 3. Histogramas para las distancias D_K con $K = 1, 3, 5, 10$ para los datos simulados.

Si evaluamos la distancia DG_K desde cada punto de un retículo o malla regular al K-ésimo punto más cercano de entre los originales encontramos de nuevo claras diferencias entre las distancias de punto de retículo a localización de ruido o rasgo (ver Tablas 1 y 2 para diferentes valores de K).

Tabla 1. Medidas descriptivas de las distancias DG_K para el rasgo de los datos simulados.

DG_K (Rasgo)	Min	Q_{25}	Mediana	Media	Q_{25}	Max
$K = 1$	0.001071	0.005459	0.008011	0.008237	0.010858	0.01823
$K = 3$	0.005813	0.01134	0.01338	0.01434	0.01675	0.02709
$K = 5$	0.008987	0.01613	0.01923	0.01908	0.02138	0.03088
$K = 10$	0.01917	0.02605	0.0281	0.02824	0.03102	0.04037

Tabla 2. Medidas descriptivas de las distancias DG_K para el ruido de los datos simulados.

DG_K (Ruido)	Min	Q_{25}	Mediana	Media	Q_{25}	Max
$K=1$	0.0008557	0.01575	0.0234	0.02528	0.03301	0.07655
$K=3$	0.008291	0.03625	0.04668	0.04744	0.05698	0.104
$K=5$	0.02069	0.05135	0.06182	0.06286	0.07294	0.1329
$K=10$	0.03391	0.07633	0.08909	0.0903	0.1019	0.1731

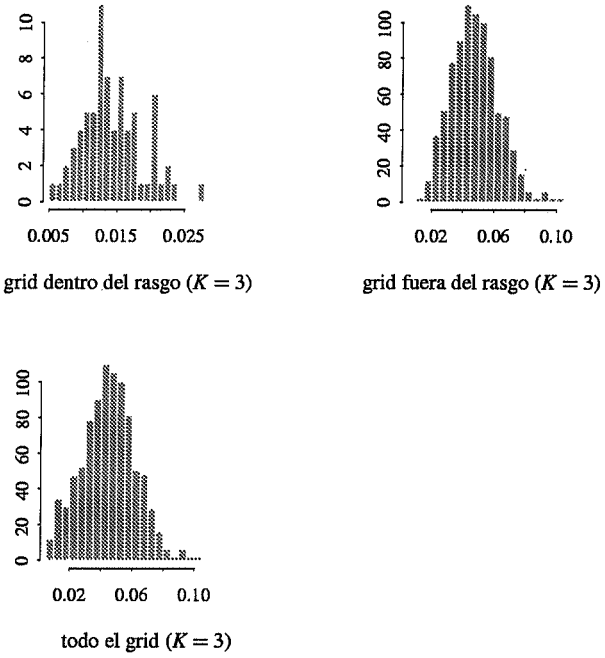


Figura 4. Histogramas para las distancias DG_K con $K=3$ para los datos simulados.

La Figura 4 representa histogramas referentes a las distancias DG_K con $K=3$. Notar que el tercer histograma es el resumen conjunto de los datos anteriores y para el caso $K=3$ no se aprecia tanto la bimodalidad (que representaría las diferencias entre rasgo y ruido). Sin embargo, analizándolos por separado si vemos que el primer histograma (rasgo) muestra distancias 0.005 a 0.025, mientras que en el segundo el rango de distancias es de 0.02 a 0.1. En la Figura 5 repetimos el proceso anterior pero para $K=10$, y en este caso, en la tercera gráfica sí aparece una bimodalidad clara. Por tanto, mayores valores de K proporcionan diferencias más claras entre rasgo y ruido.

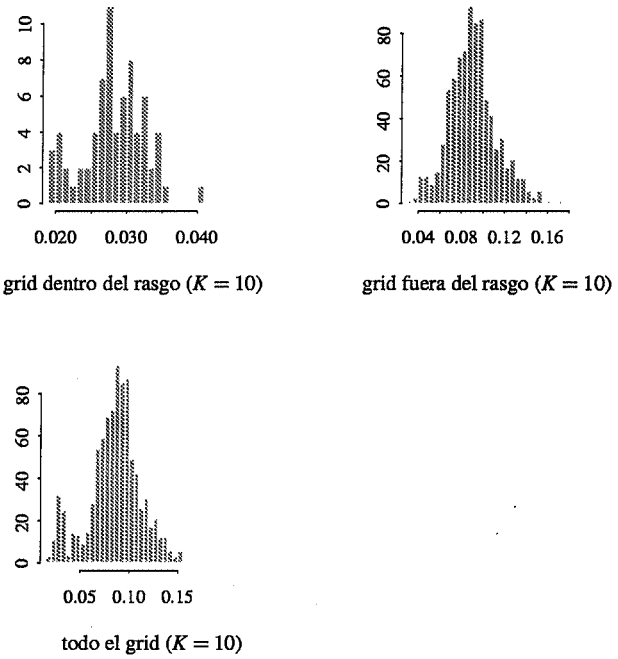


Figura 5. Histogramas para las distancias DG_K con $K = 10$ para los datos simulados.

Visto pues que el uso de cierto tipo de distancias puede discriminar entre rasgo y ruido, analizamos finalmente nuestros datos simulados con nuestra propuesta basada en la definición de un proceso puntual espacial marcado con unas marcas basadas en las distancias al K-ésimo vecino más cercano y otras marcas utilizadas como criterio de clasificación y obtenidas por el algoritmo EM. Para ello utilizamos diferentes valores de K, $K = 5, 10, 15$, y los resultados de clasificación se muestran en las Figuras 6, 7 y 8. Es claro que un aumento del valor de K va asociado con una mejor clasificación. En cada gráfica se presentan los histogramas de las distancias D_K, así como un mapa de probabilidad de pertenencia al rasgo. Vemos que los resultados demuestran la utilidad de nuestra metodología en el objetivo de la detección de rasgos.

Finalmente, para acabar con nuestro estudio de detección de rasgos, analizamos unos datos reales consistentes en la detección de zonas cancerígenas (zonas de una mayor acumulación de cuerpos) en una célula de un tejido biológico. El objetivo consiste en detectar y separar los dos rasgos del ruido existente en la imagen (ver Figura 9). Para ello hemos aplicado de nuevo nuestra técnica. Los resultados de clasificación para $K = 5, 10, 15$ se muestran en las Figuras 9, 10 y 11. De nuevo, es claro que un aumento del valor de K va asociado con una mejor clasificación.

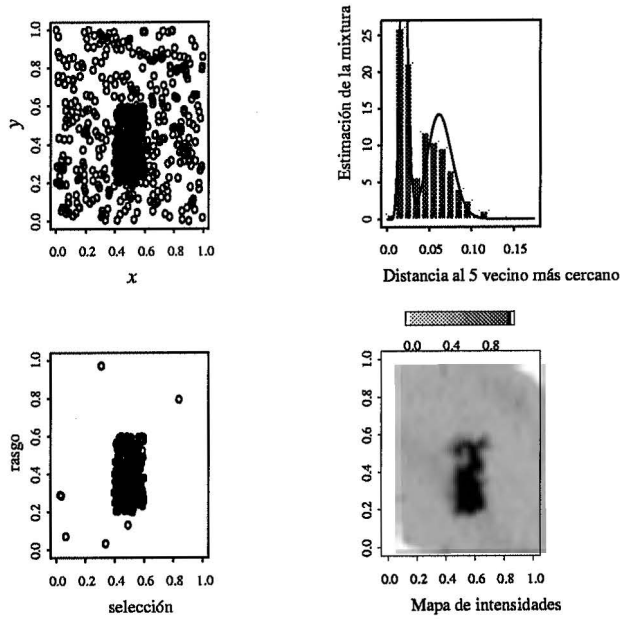


Figura 6. Detección de un rasgo con distancias al 5^o vecino más próximo.

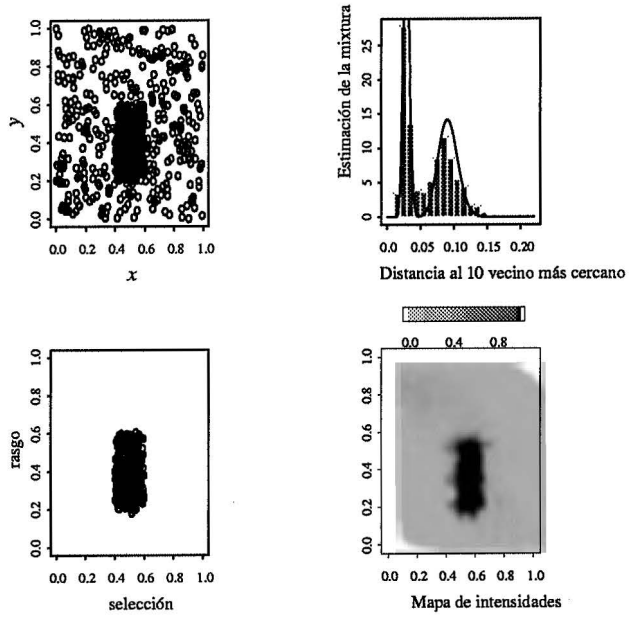


Figura 7. Detección de un rasgo con distancias al 10^o vecino más próximo.

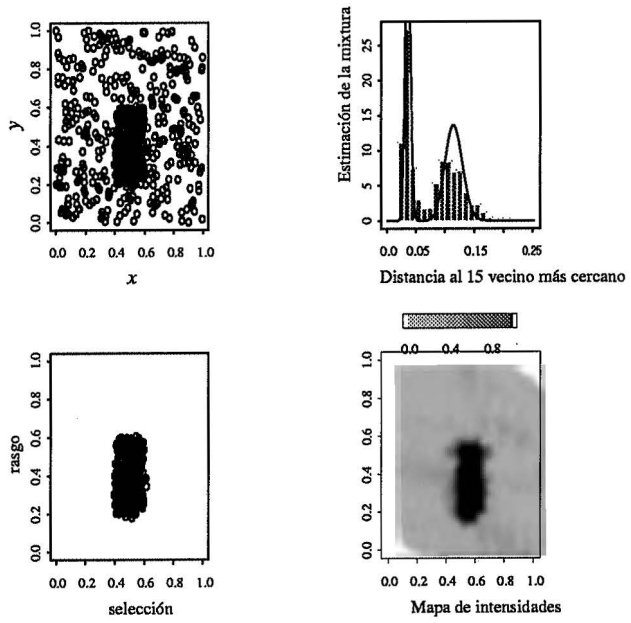


Figura 8. Detección de un rasgo con distancias al 15^o vecino más próximo.

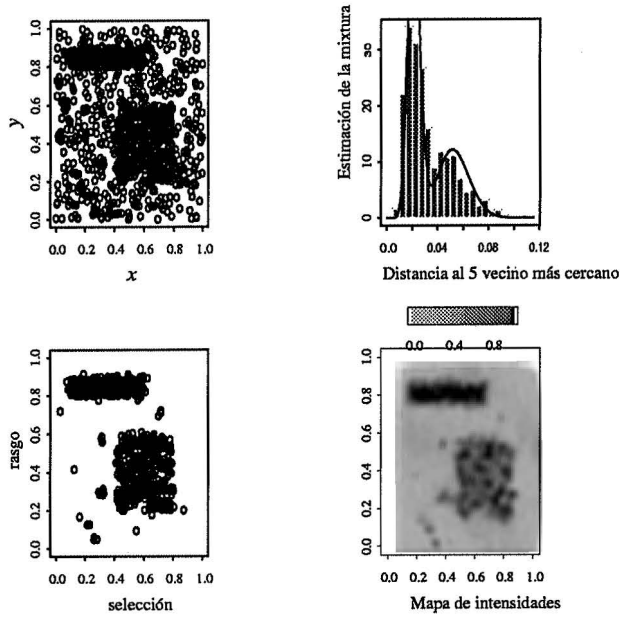


Figura 9. Detección de dos rasgos en células enfermas con distancias al 5^o vecino más próximo.

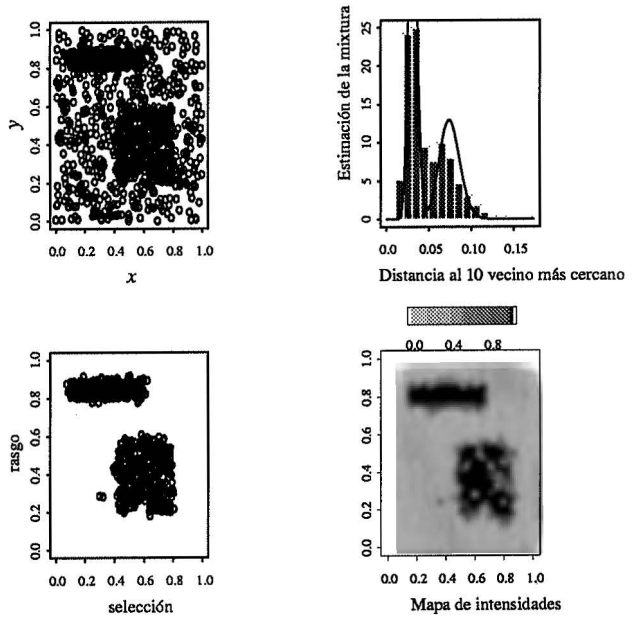


Figura 10. Detección de dos rasgos en células enfermas con distancias al 10^o vecino más próximo.

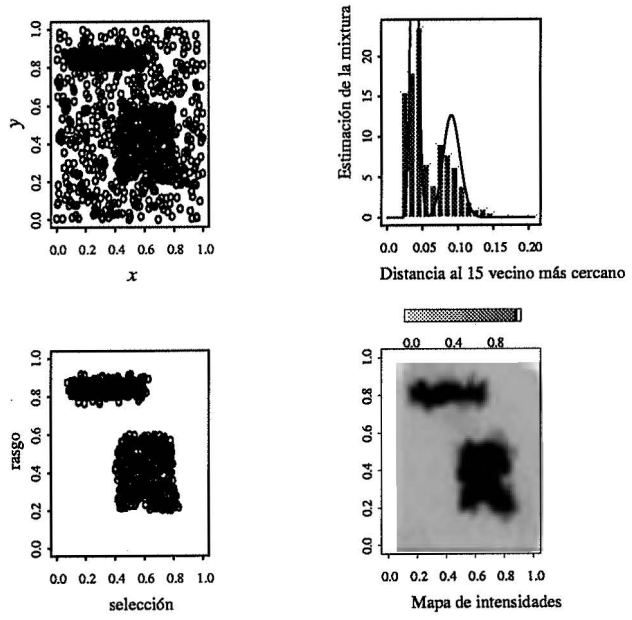


Figura 11. Detección de dos rasgos en células enfermas con distancias al 15^o vecino más próximo.

6. CONCLUSIONES Y DISCUSIÓN

En este trabajo se han presentado dos métodos complementarios sobre detección de rasgos subyacentes a ruido en imágenes usualmente digitalizadas. El primero de ellos, conocido y presentado en los ultimos años, consiste en una adaptación al caso espacial de técnicas de cluster jerárquico. El segundo explota las propiedades de cierto tipo de distancias entre sucesos en el contexto de los procesos espaciales marcados para obtener una clasificación a través del algoritmo EM. Ambos métodos discriminan correctamente entre rasgo y ruido, si bien el segundo de ellos es más sencillo de llevar a la práctica y más robusto en el sentido de que no hace ningún tipo de suposición sobre los modelos probabilísticos de los datos. Además, el método de distancias clasifica, en general para valores altos de K, mejor que los modelos jerárquicos.

Por otra parte, las expresiones utilizadas en el algoritmo EM se restringen a sólo dos intensidades diferentes (λ_1, λ_2). Una buena generalización de este método sería el incorporar la posibilidad de detección de todo un rango de posibles intensidades. En este caso la distribución de las distancias D_K sería la siguiente:

$$D_K \backsim \sum_{j=1}^{s} p_i \Gamma^{(\frac{1}{2})}(K, \lambda_j \pi), \tag{16}$$

donde

$$\sum_{j=1}^{s} p_j = 1$$

En este caso, el paso E del algoritmo EM vendría dado por

$$E\left(\widehat{\delta}_{ij}(t+1)\right) = \frac{\widehat{p}_j(t) f_{D_s}\left(d_i; \widehat{\lambda}_j(t)\right)}{\sum_{j=1}^{s} \widehat{p}_j(t) f_{D_s}\left(d_i; \widehat{\lambda}_j(t)\right)} \quad , \quad \sum_{j=1}^{s} \widehat{p}_j = 1 \quad i = 1, \ldots n. \tag{17}$$

Y el paso M por

$$\widehat{\lambda}_j(t+1) = \frac{s \sum_{i=1}^{n} \widehat{\delta}_{ij}(t+1)}{\pi \sum_{i=1}^{n} d_i^2 \, \widehat{\delta}_{ij}(t+1)} \quad j = 1, \ldots, s \tag{18}$$

con

$$\widehat{p}_j(t+1) = \frac{\sum_{i=1}^{n} \delta_{ij}(t+1)}{n}, \tag{19}$$

siendo

$$\sum_{j=1}^{s} \widehat{\delta}_{ij}(t+1) = 1.$$

Actualmente los autores están trabajando sobre el uso de funciones LISA, funciones basadas en densidades producto de procesos puntuales, para proporcionar nuevos métodos de clasificación.

AGRADECIMIENTOS

Los autores quieren agradecer a los referees las sugerencias recibidas sobre una versión anterior del trabajo.

REFERENCIAS

Banfield, J. D. & Raftery, A. E. (1993). «Model-based Gaussian and non-Gaussian clustering». *Biometrics*, 49, 803-849.

Bensmail, H., Celeux, G., Raftery, A. E. & Robert, C. P. (1997). «Inference in model-based cluster analysis». *Statistics and Computing*, 7, 1-10.

Byers, S. & Raftery, A. E. (1997). *Bayesian estimation and segmentation of spatial point processes using Voronoï tilings*, Technical Report, University of Washington, 32.

Byers, S. & Raftery, A. E. (1998). «Nearest neighbor clutter removal for estimating features in spatial point processes». *Journal of the American Statistical Association*, 93, 577-584.

Celeux, G. & Govaert, G. (1992). «A classification EM algorithm for clustering and two stochastic versions». *Computational Statistics and Data Analysis*, 14, 315-332.

Collins, L. B. (1995). *Inter-event distancie methods for the statistical analysis of spatial point processes*. Ph. D. Dissertation, Department of Statistics, University of Chicago, Chicago, Illinois.

Cox, D. R. (1975). «Partial likelihood». *Biometrika*, 62, 269-276.

Cressie, N. A. C. (1993). *Statistics for spatial data*, Wiley-Interscience.

Dasgupta, A. & Raftery, A. E. (1998). «Detecting features in spatial point processes with clutter via model-based clustering». *Journal of the American Statistical Association*, 93, 294-302.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). «Maximum likelihood for incomplete data via the EM algorithm (with discussion)». *Journal of the Royal Statistical Society*, B, 39, 1-38.

Diggle, P. J. (1983). *Statistical analysis of spatial point patterns*, Academic Press.

Fraley, C. & Raftery, A. E. (1998). «How many clusters? Which clustering method?-Answers via model-based cluster analysis». *The Computer Journal*, 41, 578-588.

Fraley, C. & Raftery, A. E. (1999). «MCLUST: software for model-based cluster and discriminant analysis». *Technical report*, 342, Department of Statistics, University of Washington.

Gower, J. C. (1967). «A comparison of some methods of cluster analysis». *Biometrics*, 23, 623-637.

Hathaway, R. J. (1986). «Another interpretation of the EM algorithm for mixture distributions». *Journal of Statistics & Probability Letters*, 4, 53-56.

Kass, R. E. & Raftery, A. E. (1995). «Bayes Factors». *Journal of the American Statistical Association*, 90, 773-795.

Kaufman, L. & Rousseeuw, P. J. (1990). *Finding groups in data. An introduction to cluster analysis*. John Wiley & Sons.

Marriott, F. H. C. (1975). «Separating mixtures of normal distributions». *Biometrics*, 31, 767-769.

McLachlan, G. J. & Basford, K. E. (1988). *Mixture models: inference and applications to clustering*. Marcel Dekker.

Penttinen, A. K., Stoyan, D. & Henttonen, H. M. (1992). «Marked point processes in forest statistics». *Forest Science*, 38, 806-824.

Smith, A. & Spiegelhalter, D. (1980). «Bayes factors and choice criteria for linear models». *Journal of the Royal Statistical Society*, A, 42, 213-220.

Stacy, E. W. (1962). «A generalization of the gamma distribution». *Annals of Mathematical Statistics*, 33, 1187-1192.

Wolfe, J. H. (1970). «Pattern clustering by multivariate mixture analysis». *Multivariate Behavioral Research*, 5, 329-350.

ENGLISH SUMMARY

DETECTION OF FEATURES IN BINARY IMAGES BY MEANS OF SPATIAL MARKED POINT PROCESSES

J. MATEU*
G. LORENZO**
Universitat Jaume I*

In this paper the features detection in binary images under the presence of clutter is considered. Practical examples of this problem can be found in the mine detection through satellite images, in the detection of features in biological cell images or in the detection of faults in earthquake problems. We first revise detection methods based on probabilistic models, in which features are normally distributed, and the clutter comes from a spatial Poisson process. Then, we show a new solution to the problem in hand by means of spatial marked point processes. The marked processes are defined in terms of two variables: distances to the kth-nearest neighbour and a Bernoulli variable to define feature and clutter. The distances are modelled through a mixture of probability distributions whose parameters are obtained by the EM algorithm.
Finally, the new methodology is evaluated and tested over simulations and real-case studies.

Keywords: EM algorithm, mixtures, Poisson process, spatial point processes

AMS Classification (MSC 2000): 62M30, 60G55

* Autor para la correspondencia. Email: mateu@mat.uji.es. Fax: 964.728429
** Email: valentin@mat.uji.es
* Departamento de Matemáticas, Universitat Jaume I. Campus Riu Sec, E-12071 Castellón, Spain.
– Received December 2000.
– Accepted February 2002.

The analysis of noisy images in which features are present under substantive clutter is a key tool in many different contexts. After being processed, the image reduces to a set of objects (points), some of them will identify features and others will stand for clutter. So the final image consists of a collection of planar points located in a region of interest. Consequently, the theory of spatial point processes can handle this kind of problem. A spatial point process is a set of points spatially distributed over the plane that show certain types of interaction. A basic and benchmark process is the Poisson process, for which the points are randomly distributed.

In the last years, the problem of feature detection in presence of clutter has focused on probability-based cluster methods. However, these methods are rather complicated, they need the data to be Normally distributed and need the shape and size parameters be specified. Consequently, in this paper we propose a different approach based on spatial marked processes. Each location is marked with two variables, the K-th nearest-neighbour distances and a binary variable identifying the type of point (feature or clutter). This method puts no hypothesis in the probability distribution of the data and has no size or shape parameters to define.

Methods based on probabilistic models

These techniques are parametric and are based on the hypothesis that the data comes from a mixture of probability distributions. They consider a priori information determining parameters such as shape, size and orientation of the clusters. If the data is Gaussian, then the variance-covariance matrix can be decomposed using the spectral decomposition. This decomposition defines the basic tool to proceed as gives light to elements that characterize the cluster.

Method based on Kth nearest-neighbour distances

Now, we assume that both, features and clutter come from a Poisson spatial process and define the following spatial process. Let $N = \{x_i, y_i, d_i, \delta_i\}_{i=1}^{n}$ with (x_i, y_i) the original spatial locations, d_i the Kth nearest-neighbour distances and $\delta_i \in \{0, 1\}$. Moreover, $\delta_i = 1$, if the ith point belongs to a feature and $\delta_i = 0$ if it is clutter. Now, the probability distribution function of the Kth nearest-neighbour distance, D_K, is given by

$$P(D_K \geq d) = \sum_{k=0}^{K-1} \frac{e^{-\lambda\pi d^2} \left(-\lambda\pi d^2\right)^k}{k!} = 1 - F_{D_K}(d) \tag{1}$$

where λ is the intensity of the process.

The corresponding probability density function, $f_{D_K}(d)$, corresponds to a Generalized Gamma distribution where $Y = (D_K)^2$ and $Y \sim \Gamma(K, \lambda\pi)$. Then, the likelihood function takes the form

$$L(d_i;\lambda) = \prod_{i=1}^{n} f_{D_K}(d_i;\lambda) = \frac{e^{-\lambda\pi\sum_{i=1}^{n} d_i^2}\left[2(\lambda\pi)^K\right]^n \prod_{i=1}^{n} d_i^{2K-1}}{(K-1)!}$$

So, our basic starting point is the following:

$$D_K \simeq p\Gamma^{(1/2)}(K,\lambda_1\pi) + (1-p)\Gamma^{(1/2)}(K,\lambda_2\pi) \tag{2}$$

and the aim is to estimate the classification marks δ_i, $\forall i \in \{1,\ldots,n\}$ together with the intensity parameters λ_1 ,λ_2 and p. For this we use the EM algorithm.

QÜESTIIÓ, vol. 26, 1-2, p. 87-108, 2002

ESTIMACIÓN DE LA FUNCIÓN DE DISTRIBUCIÓN SOBRE POBLACIONES FINITAS MEDIANTE DISEÑOS MUESTRALES BIETÁPICOS APROPIADOS

J. A. MAYOR GALLEGO
M. MARTÍNEZ BLANES
Universidad de Sevilla*

Con el objeto de estimar la función de distribución de una variable de estudio sobre una población finita, se propone en este trabajo emplear el estimador de Horvitz-Thompson, lo que proporciona una estrategia muestral insesgada, siendo la varianza de dicho estimador una función real de variable real cuya minimización permite obtener diseños muestrales óptimos bajo diferentes criterios.
En este trabajo empleamos la norma $\|\cdot\|_1$ como criterio de optimización, minimizando la norma de la varianza, como función de la matriz del diseño muestral. De esta forma, suponiendo muestreo por conglomerados en dos etapas y considerando como dominio de búsqueda el conjunto de los diseños muestrales de tipo uniforme, en el sentido de ser iguales las probabilidades de inclusión de primer orden, se estudia la obtención de diseños muestrales adecuados para dicha estimación.

Estimating distributions functions using appropriate sampling designs in two stage cluster sampling

Palabras clave: Muestreo, poblaciones finitas, diseño muestral, función de distribución, conglomerados

Clasificación AMS (MSC 2000): 62D05

* Dpto. de Estadística e Investigación Operativa. Universidad de Sevilla. Facultad de Matemáticas. c/ Tarfia s/n, 41012 Sevilla, España. E-mail:mayor@cica.es

– Recibido en marzo de 2001.
– Aceptado en diciembre de 2001.

1. INTRODUCCIÓN

Si bien la teoría del muestreo en poblaciones finitas se ha centrado clásicamente en la estimación de parámetros poblacionales de tipo puntual como totales, medias, proporciones y varianzas, existen una serie de parámetros de tipo funcional que pueden proporcionarnos información relevante acerca del comportamiento global de la población.

En este trabajo consideramos el problema de la estimación de un parámetro de este tipo, en concreto, la función de distribución poblacional asociada a una variable numérica definida sobre la población. Este problema es importante por el interés intrínseco del parámetro funcional mencionado y también por su relación con otros parámetros de tipo no funcional como la mediana, los cuantiles o el índice de Gini, habiendo sido tratado con diferentes enfoques por varios autores.

Así, en relación a la estimación de la mediana y los cuantiles, es obligado citar el trabajo inicial de Wooddruff (1952), en el que se construye un intervalo de confianza para la estimación de la mediana poblacional y otras medidas de posición, empleando el muestreo aleatorio simple. Por otra parte, Sendransk y Meyer (1978) estudian este problema bajo un enfoque puramente probabilístico de distribución de estadísticos ordenados, para muestreo aleatorio simple y estratificado. Hill (1968) emplea un enfoque bayesiano y Kuk y Mak (1989), técnicas de información auxiliar proporcionada por otras variables.

Para el problema de la estimación de la función de distribución poblacional propiamente dicha, también encontramos diferentes enfoques en la bibliografía. Por ejemplo Chambers y Dunstan (1986) emplean un modelo de superpoblación para desarrollar un procedimiento de estimación. Kuk (1988) estudia y compara varios estimadores de la función de distribución poblacional empleando muestreo con probabilidades variables y Rao, Kovar y Mantel (1990) desarrollan estimadores que emplean información auxiliar. Citemos también los trabajos de Chambers *et al.* (1992) y Rao (1994).

A continuación vamos a considerar este problema con un enfoque distinto de los anteriores, y que se basa en el estudio de la función varianzá del estimador de Horvitz-Thompson con el fin de buscar diseños muestrales óptimos, en una clase especial de diseños muestrales. Para ello, vamos a considerar una población finita, $U = \{1,2 \ldots N\}$, y sea Y una variable de estudio numérica, cuyos valores sobre U, son $(Y_1, Y_2 \ldots Y_N)$, que, sin pérdida de generalidad, supondremos ordenados de menor a mayor, esto es, $Y_1 \leq Y_2 \leq \cdots \leq Y_N$. Nuestro objetivo es la estimación de la función de distribución de la variable Y,

$$F(t) = \frac{1}{N} \text{CARD}(\{i \in U \,|\, Y_i \leq t\})$$

Si empleamos las funciones indicadoras de los intervalos de la forma $[Y_i, +\infty)$, denotándolas $I_{[Y_i,+\infty)}(t)$, podemos expresar $F(t)$ como,

$$F(t) = \frac{1}{N} \sum_{i \in U} I_{[Y_i,+\infty)}(t)$$

es decir, una expresión lineal, que puede ser estimada mediante el estimador de Horvitz-Thompson.

Sea pues m una muestra obtenida de la población U mediante un diseño muestral, $d = (\mathcal{M}, P(\cdot))$, sin reemplazamiento y con matriz de diseño que denotaremos como $\Pi = \{\pi_{ij}\}_{1 \leq i,j \leq N}$, con $\pi_{ii} = \pi_i$. El mencionado estimador de $F(t)$ resulta ser,

$$\widehat{F}(t) = \frac{1}{N} \sum_{i \in m} \frac{I_{[Y_i,+\infty)}(t)}{\pi_i}$$

Como sabemos, este estimador es insesgado y su varianza puede ser expresada mediante la clásica fórmula,

$$V[\widehat{F}(t)] = \frac{1}{N^2} \sum_{i,j \in U} (\pi_{ij} - \pi_i \pi_j) \frac{I_{[Y_i,+\infty)}(t)}{\pi_i} \frac{I_{[Y_j,+\infty)}(t)}{\pi_j}$$

y así, para cada valor de $t \in \mathbb{R}$, podemos emplear dicha expresión como medida de la bondad de la estimación.

2. MUESTREO POR CONGLOMERADOS EN DOS ETAPAS

Al ser la varianza una función, no tiene sentido hablar de varianza mínima en el sentido usual, por ello, vamos a definir un criterio apropiado que nos permita obtener propiedades de los diseños muestrales más adecuados para la estimación que estamos realizando. El criterio que definimos viene determinado por la siguiente distancia, basada en la norma funcional $\|\cdot\|_1$,

$$\mathrm{d}(V[\widehat{F}(t)], 0) = \|V[\widehat{F}(t)]\|_1 = \int_{Y_1}^{Y_N} |V[\widehat{F}(t)]| dt = \int_{Y_1}^{Y_N} V[\widehat{F}(t)] dt$$

Notemos que $V[\widehat{F}(t)]$ cumple los requerimientos analíticos para que esta distancia esté bien definida. En particular, se verifica que $\|V[\widehat{F}(t)]\|_1 = 0$ si y solamente si $V[\widehat{F}(t)] =$

$0, \forall t \in [Y_1, Y_N]$. Notemos también que es posible emplear otro tipo de normas. La elección de la norma $\|\cdot\|_1$ en este trabajo viene justificada por razones de simplicidad en los desarrollos.

Supondremos que se realiza un muestreo por conglomerados en una etapa, así la población sobre la que se realiza el muestreo es una población de conglomerados, $U_c = \{C_1, \ldots, C_i, \ldots, C_M\}$, cada uno con tamaños respectivos $N_1, \ldots, N_i, \ldots, N_M$. Sobre esta población se emplea un diseño muestral $d_c = (\mathcal{M}_c, P_c)$, con matriz de diseño $\{\pi^c_{ij}\}$, siendo m_c la muestra de conglomerados obtenida. En una segunda etapa supondremos que en el conglomerado C_i, $i \in m_c$ se emplea un diseño muestral $d_i = (\mathcal{M}_i, P_i)$, con matriz de diseño $\{\pi^i_{kl}\}$, para obtener una muestra m_i con n_i unidades finales.

Las probabilidades de inclusión de dichas unidades finales, se construirán pues a partir de las probabilidades de inclusión de los diseños muestrales involucrados. Dichas probabilidades vienen dadas por,

$$\begin{aligned}
&\pi_k = \pi^c_i \pi^i_k && \text{si } k \in C_i && \forall\, k \in U \\
&\pi_{kl} = \pi^c_i \pi^i_{kl} && \text{si } k, l \in C_i,\ k \neq l && \forall\, k, l \in U \\
&\pi_{kl} = \pi^c_{ij} \pi^i_k \pi^j_l && \text{si } k \in C_i,\ l \in C_j,\ i \neq j && \forall\, k, l \in U
\end{aligned}$$

De esta forma, denotando por m la muestra de unidades finales provenientes de dichos conglomerado, el estimador de Horvitz-Thompson será,

$$\begin{aligned}
\widehat{F}(t) &= \sum_{k \in m} \frac{1}{N} \frac{I_{[Y_k, +\infty)}(t)}{\pi_k} = \sum_{i \in m_c} \sum_{k \in m_i} \frac{1}{N} \frac{I_{[Y_k, +\infty)}(t)}{\pi^c_i \pi^i_k} \\
&= \sum_{i \in m_c} \frac{1}{\pi^c_i} \sum_{k \in m_i} \frac{1}{N} \frac{I_{[Y_k, +\infty)}(t)}{\pi^i_k} = \sum_{i \in m_c} \frac{\widehat{F}_i(t)}{\pi^c_i}
\end{aligned}$$

donde $F_i(t)$ denota el parámetro funcional que se está estimando, valorado sobre el conglomerado C_i, es decir,

$$F_i(t) = \sum_{k \in C_i} \frac{1}{N} I_{[Y_k, +\infty)}(t)$$

y $\widehat{F}_i(t)$ denota su estimación de Horvitz-Thompson, esto es,

$$\widehat{F}_i(t) = \sum_{k \in m_i} \frac{1}{N} \frac{I_{[Y_k, +\infty)}(t)}{\pi^i_k}$$

Observemos que las funciones $F_i(t)$ aparecen como resultado de la descomposición usual del estimador $\widehat{F}(t)$, y no representan las funciones de distribución de los conglomerados.

La varianza de dicho estimador vendrá dada por la expresión usual para un muestreo bietápico, véase Fernández y Mayor (1995), y que adaptada a la estructura poblacional considerada, resulta ser,

$$\begin{aligned} V[\widehat{F}(t)] &= \sum_{i,j\in U_c}\sum (\pi^c_{ij} - \pi^c_i\pi^c_j)\frac{F_i(t)}{\pi^c_i}\frac{F_j(t)}{\pi^c_j} + \sum_{i\in U_c}\frac{1}{\pi^c_i}V[\widehat{F}_i(t)] \\ &= \sum_{i,j\in U_c}\sum (\pi^c_{ij} - \pi^c_i\pi^c_j)\frac{F_i(t)}{\pi^c_i}\frac{F_j(t)}{\pi^c_j} \\ &+ \frac{1}{N^2}\sum_{i\in U_c}\frac{1}{\pi^c_i}\sum_{k,l\in C_i}\sum (\pi^i_{kl} - \pi^i_k\pi^i_l)\frac{I_{[Y_k,+\infty)}(t)}{\pi^i_k}\frac{I_{[Y_l,+\infty)}(t)}{\pi^i_l} = V_1(t) + V_2(t) \end{aligned}$$

Observemos que la varianza se descompone en dos sumandos, $V_1(t)$ y $V_2(t)$, cada uno de los cuales representa el error de muestreo inherente a ambas etapas. Para aplicar nuestra metodología, vamos a estudiar la norma $\|\cdot\|_1$ de esta varianza, bajo ciertas hipótesis realizadas sobre los diseños muestrales. En primer lugar, notemos que al ser $V_1(t) \geq 0$ y $V_2(t) \geq 0$, $\forall t \in \mathbb{R}$, se verifica,

$$\|V[\widehat{F}(t)]\|_1 = \|V_1(t)\|_1 + \|V_2(t)\|_1 \quad \forall t \in \mathbb{R}$$

y en particular $\forall t \in [Y_1, Y_N]$. Seguidamente exponemos una serie de definiciones y resultados relacionados con el desarrollo de las cantidades anteriores.

Definición 1. *Dado un diseño muestral, $d = (\mathcal{M}, P)$, diremos que es uniforme si las probabilidades de inclusión de primer orden son iguales, es decir, $\pi_i = \alpha$, $\forall i$.*

Lema 1. *Si un diseño muestral, $d = (\mathcal{M}, P)$, definido sobre una población U con N unidades, es uniforme y sus muestras son de tamaño fijo, n, entonces se verifica $\pi_i = n/N, \forall i \in U$.*

Demostración

Dado $i \in U$, sea H_i una variable aleatoria con distribución de Bernoulli, definida sobre $\mathcal{M}$ como $H_i(m) = 1$ si $i \in m$ y $H_i(m) = 0$ si $i \neq m$, $\forall m \in \mathcal{M}$. La esperanza matemática de H_i es $E[H_i] = \pi_i$. Se tiene entonces,

$$\sum_{i\in U} H_i = n \quad \Rightarrow \quad \sum_{i\in U} \pi_i = \sum_{i\in U} E[H_i] = n$$

de donde se deduce el resultado. □

Observemos que el diseño muestral aleatorio simple, es decir, el formado por todas las posibles muestras de tamaño fijo n, con probabilidades iguales, es decir, $P(m) = 1/\binom{N}{n}$, es un diseño uniforme de tamaño fijo. Denotaremos este diseño muestral como $\mathrm{MAS}(N,n)$.

Lema 2. *Si los diseños muestrales d_i, $i \in U_c$ son uniformes con tamaños fijos respectivos n_i, y si denotamos,*

$$V^{(i)}(t) = \sum_{k,l\in C_i}\sum (\pi^i_{kl} - \pi^i_k\pi^i_l)\frac{I_{[Y_k,+\infty)}(t)}{\pi^i_k}\frac{I_{[Y_l,+\infty)}(t)}{\pi^i_l} \qquad i \in U_c$$

se verifica,

$$\|V^{(i)}(t)\|_1 = \frac{1}{2}\sum_{k,l\in C_i}\sum |Y_k - Y_l| - \frac{N_i^2}{2n_i^2}\sum_{k,l\in C_i}\sum \pi^i_{kl}|Y_k - Y_l|$$

Si además, d_i =MAS(N_i,n_i), $\forall i \in U_c$, entonces,

$$\|V^{(i)}(t)\|_1 = \frac{N_i - n_i}{2(N_i-1)n_i}\sum_{k,l\in C_i}\sum |Y_k - Y_l|$$

Demostración

Observemos que dados $k,l \in C_i$, se tiene,

$$\int_{Y_1}^{Y_N} I_{[Y_k,+\infty)}(t) I_{[Y_l,+\infty)}(t)\,dt = \int_{Y_1}^{Y_N} I_{[\max\{Y_k,Y_l\},+\infty)}(t)\,dt = Y_N - \max\{Y_k,Y_l\}$$

por consiguiente,

$$\|V^{(i)}(t)\|_1 \quad = \int_{Y_1}^{Y_N} V^{(i)}(t)\,dt = \sum_{k,l\in C_i}\sum \frac{\pi^i_{kl} - \pi^i_k\pi^i_l}{\pi^i_k\pi^i_l}\int_{Y_1}^{Y_N} I_{[Y_k,+\infty)}(t) I_{[Y_l,+\infty)}(t)\,dt$$

$$= \sum_{k,l\in C_i}\sum \frac{\pi^i_{kl} - \pi^i_k\pi^i_l}{\pi^i_k\pi^i_l}(Y_N - \max\{Y_k, Y_l\})$$

$$= \sum_{k,l\in C_i}\sum \frac{\pi^i_{kl}}{\pi^i_k\pi^i_l}(Y_N - \max\{Y_k, Y_l\}) - \sum_{k,l\in C_i}\sum (Y_N - \max\{Y_k, Y_l\})$$

Por ser d_i uniforme de tamaño fijo, tendremos por el Lema 1, que $\pi^c_l = n_i/N_i$, $\forall l \in C_i$, siendo pues,

$$\begin{aligned}
&\sum_{k,l\in C_i}\sum \frac{\pi^i_{kl}}{\pi^i_k\pi^i_l}(Y_N - \max\{Y_k, Y_l\}) \\
&\quad = \frac{N_i^2}{n_i^2}\sum_{k,l\in C_i}\sum \pi^i_{kl}(Y_N - \frac{1}{2}(Y_k + Y_l + |Y_k - Y_l|) \\
&\quad = \frac{N_i^2}{n_i^2}\left[\sum_{k,l\in C_i}\sum \pi^i_{kl}Y_N - \frac{1}{2}\sum_{k\in C_i} Y_k \sum_{l\in C_i} \pi^i_{kl} - \frac{1}{2}\sum_{l\in C_i} Y_l \sum_{k\in C_i} \pi^i_{kl} - \frac{1}{2}\sum_{k,l\in C_i}\sum \pi^i_{kl}|Y_k - Y_l|\right] \\
&\quad = \frac{N_i^2}{n_i^2}\left[n_i^2 Y_N - \frac{n_i^2}{N_i}T_i(Y) - \frac{1}{2}\sum_{k,l\in C_i}\sum \pi^i_{kl}|Y_k - Y_l|\right] \\
&\quad = N_i^2 Y_N - N_i T_i(Y) - \frac{N_i^2}{2n_i^2}\sum_{k,l\in C_i}\sum \pi^i_{kl}|Y_k - Y_l|
\end{aligned}$$

donde hemos denotado $T_i(Y) = \sum_{k\in C_i} Y_k$, y hemos tenido en cuenta que por ser el diseño muestral uniforme y de tamaño de muestra fijo, verifica,

$$\sum_{k\in C_i} \pi^i_{kl} = n_i\pi^i_l = \frac{n_i^2}{N_i} \qquad \text{y} \qquad \sum_{k,l\in C_i}\sum \pi^i_{kl} = n_i^2$$

Por otra parte, un cálculo similar al anterior proporciona,

$$\sum_{k,l\in C_i}\sum (Y_N - \max\{Y_k, Y_l\}) = N_i^2 Y_N - N_i T_i(Y) - \frac{1}{2}\sum_{k,l\in C_i}\sum |Y_k - Y_l|$$

Restando ambas se obtiene inmediatamente el primer resultado. En caso de ser $d_i = \text{MAS}(N_i, n_i)$, $i \in U_c$, se tendrá además para las probabilidades de inclusión de segundo

orden $\pi_{kl}^i = n_i(n_i-1)/N_i(N_i-1)$, $k \neq l$, y basta sustituir para obtener, mediante un cálculo directo, el segundo resultado. □

Lema 3. *Las funciones $F_i(t)$ definidas anteriormente, verifican,*

$$\int_{Y_1}^{Y_N} F_i(t)\,F_j(t)\,dt = \frac{1}{N^2}\sum_{k\in C_i}\sum_{l\in C_j}(Y_N - \max\{Y_k, Y_l\}) \qquad \forall i,j \in U_c$$

Demostración

En primer lugar, observemos que,

$$F_i(t)\,F_j(t) = \frac{1}{N^2}\sum_{k\in C_i}\sum_{l\in C_j} I_{[Y_k,+\infty)}(t) I_{[Y_l,+\infty)}(t)$$

luego,

$$\int_{Y_1}^{Y_N} F_i(t)\,F_j(t)\,dt = \frac{1}{N^2}\sum_{k\in C_i}\sum_{l\in C_j}\int_{Y_1}^{Y_N} I_{[Y_k,+\infty)}(t) I_{[Y_l,+\infty)}(t)\,dt$$

y basta sustituir la expresión obtenida en la demostración del Lema anterior para esta última integral. □

Aplicando los resultados anteriores, obtenemos el siguiente resultado acerca de la varianza de la estimación.

Teorema 4. *Si los diseños muestrales d_i, $i \in U_c$, son uniformes con tamaños fijos respectivos n_i, se verifica,*

$$\begin{aligned}\|V[\widehat{F}(t)]\|_1 = {} & \frac{1}{N^2}\sum_{i,j\in U_c}\frac{\pi_{ij}^c}{\pi_i^c\pi_j^c}\sum_{k\in C_i}\sum_{l\in C_j}(Y_N - \max\{Y_k,Y_l\}) \\ & -\frac{1}{N^2}\sum_{i,j\in U_c}\sum_{k\in C_i}\sum_{l\in C_j}(Y_N - \max\{Y_k,Y_l\}) \\ & +\frac{1}{N^2}\sum_{i\in U_c}\frac{1}{2\pi_i^c}\left(\sum_{k,l\in C_i}|Y_k - Y_l| - \frac{N_i^2}{n_i^2}\sum_{k,l\in C_i}\pi_{kl}^i|Y_k - Y_l|\right)\end{aligned}$$

Si además d_i =MAS(N_i, n_i), $i \in U_c$, el último sumando se puede expresar como,

$$\frac{1}{N^2} \sum_{i \in U_c} \frac{N_i}{\pi_i^c} \left(\frac{N_i}{n_i} - 1 \right) \frac{1}{2N_i(N_i - 1)} \sum_{k,l \in C_i} \sum |Y_k - Y_l|$$

□

3. REDUCCIÓN DE LA VARIANZA

Aunque la expresión obtenida permite considerar la búsqueda de diseños uniformes en cada uno de los conglomerados, que reduzcan la varianza, aquí restringiremos el planteamiento suponiendo d_i =MAS(N_i, n_i), $i \in U_c$, es decir, muestreo aleatorio en los conglomerados.

Como puede verse en el Teorema 4, bajo esta hipótesis, $\|V[\widehat{F}(t)]\|_1$ se descompone en tres términos, de los cuales el primero y el tercero dependen del diseño muestral empleado para obtener la muestra de conglomerados. Por otra parte, la cantidad que aparece en dicha expresión,

$$\frac{1}{2N_i(N_i - 1)} \sum_{k,l \in C_i} \sum |Y_k - Y_l|$$

puede considerarse como una medida de la dispersión de la variable de estudio en el conglomerado i–ésimo.

De esta forma, si suponemos afijación proporcional en los conglomerados, de manera que el tamaño de muestra en cada uno de ellos se realiza de forma proporcional al tamaño del mismo, y suponemos además que los conglomerado presentan una dispersión similar con respecto a la variable de estudio, las cantidades,

$$\left(\frac{N_i}{n_i} - 1 \right) \frac{1}{2N_i(N_i - 1)} \sum_{k,l \in C_i} \sum |Y_k - Y_l|$$

son muy homogéneas, luego una forma de disminuir la aportación del último término es tomar probabilidades de inclusión de primer orden proporcionales a los tamaños de los conglomerados, es decir, $\pi_i^c = nN_i/N$. Se obtiene entonces el siguiente resultado.

Teorema 5. *Si el muestreo en los conglomerados se realiza mediante diseños muestrales aleatorios simples, y si las probabilidades de inclusión de primer orden asociadas al muestreo de conglomerados vienen dadas por $\pi_i^c = nN_i/N$, $i \in U_c$, el primer sumando de la expresión de $\|V[\widehat{F}]\|_1$, obtenida en el Teorema 4. puede ser expresado como,*

$$\frac{1}{nN}\sum_{i\in U_c} N_i\left(Y_N-\overline{Y}_i-\frac{1}{2N_i^2}\sum\sum_{k,l\in C_i}|Y_k-Y_l|\right)$$

$$+\frac{1}{n^2}\left[n(n-1)(Y_N-\overline{Y})-\frac{1}{2}\sum\sum_{i,j\in U_c,i\neq j}\frac{\pi^c_{ij}}{N_iN_j}\sum\sum_{k\in C_i,l\in C_j}|Y_k-Y_l|\right]$$

donde $\overline{Y}$ denota la media poblacional de la variable de estudio, é $\overline{Y}_i$ la media sobre el conglomerado i−ésimo.

Demostración

En primer lugar observemos que dados $i,j\in U_c$, se tiene,

$$\begin{aligned}\sum\sum_{k\in C_i,l\in C_j}(Y_N-\max\{Y_k,Y_l\}) &= \sum\sum_{k\in C_i,l\in C_j}\left(Y_N-\frac{1}{2}Y_k-\frac{1}{2}Y_l-\frac{1}{2}|Y_k-Y_l|\right)\\ &= N_iN_jY_N-\frac{1}{2}N_jT_i(Y)-\frac{1}{2}N_iT_j(Y)-\frac{1}{2}\sum\sum_{k\in C_i,l\in C_j}|Y_k-Y_l|\\ &= N_iN_j\left(Y_N-\frac{1}{2}(\overline{Y}_i+\overline{Y}_j)-\frac{1}{2N_iN_j}\sum\sum_{k\in C_i,l\in C_j}|Y_k-Y_l|\right)\end{aligned}$$

así pues,

$$\begin{aligned}&\sum\sum_{i,j\in U_c}\frac{\pi^c_{ij}}{\pi^c_i\pi^c_j}\sum\sum_{k\in C_i,l\in C_j}(Y_N-\max\{Y_k,Y_l\})\\ &\quad=\frac{N}{n}\sum_{i\in U_c}N_i\left(Y_N-\overline{Y}_i-\frac{1}{2N_i^2}\sum\sum_{k,l\in C_i}|Y_k-Y_l|\right)\\ &\qquad+\frac{N^2}{n^2}\sum\sum_{i,j\in U_c,i\neq j}\pi^c_{ij}\left(Y_N-\frac{1}{2}(\overline{Y}_i+\overline{Y}_j)-\frac{1}{2N_iN_j}\sum\sum_{k\in C_i,l\in C_j}|Y_k-Y_l|\right)\\ &\quad=\frac{N}{n}\sum_{i\in U_c}N_i\left(Y_N-\overline{Y}_i-\frac{1}{2N_i^2}\sum\sum_{k,l\in C_i}|Y_k-Y_l|\right)\\ &\qquad+\frac{N^2}{n^2}\left[n(n-1)Y_N-n(n-1)\overline{Y}-\frac{1}{2}\sum\sum_{i,j\in U_c,i\neq j}\frac{\pi^c_{ij}}{N_iN_j}\sum\sum_{k\in C_i,l\in C_j}|Y_k-Y_l|\right]\end{aligned}$$

de donde se deduce inmediatamente el resultado. Nótese que hemos empleado la relación,

$$\sum_{j\in U_c, j\neq i} \pi^c_{ij} = (n-1)\pi^c_i = \frac{n(n-1)N_i}{N}$$

□

El resultado anterior nos dice que podemos reducir la varianza de la estimación actuando sobre el último término, es decir, eligiendo adecuadamente las probabilidades de inclusión de segundo orden, bajo el supuesto de que las de primer orden son proporcionales a los tamaños de los conglomerados. No obstante, bajo la hipótesis de que dichos tamaños son similares, podemos considerar diseño uniforme para el muestreo de los mismos. En este sentido se tiene el siguiente resultado, cuya demostración, similar a la del anterior, omitimos.

Teorema 6. *Si el muestreo en los conglomerados se realiza mediante diseños muestrales aleatorios simples, y si la selección de conglomerados se realiza mediante un diseño muestral uniforme, entonces el primer sumando de la expresión de* $\|V[\widehat{F}]\|_1$, *obtenida en el Teorema 4 puede expresarse como,*

$$\frac{M}{n}\sum_{i\in U_c} N_i^2\left(Y_N - \overline{Y}_i - \frac{1}{2N_i^2}\sum_{k,l\in C_i}\sum |Y_k - Y_l|\right)$$
$$+\frac{M^2}{n^2}\sum_{i,j\in U_c, i\neq j}\sum \pi^c_{ij}N_iN_j\left(Y_N - \frac{1}{2}(\overline{Y}_i + \overline{Y}_j) - \frac{1}{2N_iN_j}\sum_{k\in C_i, l\in C_j}\sum |Y_k - Y_l|\right)$$

donde $\overline{Y}_i$ *denota la media de la variable de estudio sobre el conglomerado* $i-$*ésimo.* □

Según este último resultado, y con las hipótesis consideradas, podemos reducir la varianza de la estimación minimizando la cantidad,

$$\sum_{i,j\in U_c, i\neq j}\sum \pi^c_{ij}N_iN_j\left(Y_N - \frac{1}{2}(\overline{Y}_i + \overline{Y}_j) - \frac{1}{2N_iN_j}\sum_{k\in C_i, l\in C_j}\sum |Y_k - Y_l|\right)$$
$$= \sum_{i,j\in U_c, i\neq j}\sum \pi^c_{ij}\sum_{k\in C_i, l\in C_j}\sum (Y_N - \max\{Y_k, Y_l\})$$

en las variables π^c_{ij}. Observemos que bajo la hipótesis de homogeneidad en los tamaños de los conglomerados, dicha minimización tenderá a asignar probabilidades de inclusión de segundo orden mayores para pares de conglomerados más distantes en el sentido de ser mayor la cantidad,

$$\frac{1}{2N_iN_j}\sum\sum_{k\in C_i, l\in C_j}|Y_k - Y_l|$$

ya que dicha cantidad se puede considerar como una medida de la disparidad entre los conglomerados C_i y C_j, en relación a los valores de la variable de estudio sobre ellos, así, las parejas de conglomerados más dispares deberían tener probabilidades de inclusión de segundo orden más elevadas.

Notemos finalmente que la variable Y es desconocida, por lo que el problema de minimización obtenido no puede ser considerado directamente, y por esta razón realizaremos algunas hipótesis sobre la estructura de la población. Así, vamos a estudiar la expresión a minimizar una estructura poblacional muy común, y que formalizaremos mediante un modelo de superpoblación adecuado.

4. MODELO DE REGRESIÓN LINEAL POR CONGLOMERADOS

$$Y_k = \alpha + \beta X_i + \varepsilon_k, \quad \beta > 0, \quad E_s[\varepsilon_k] = 0 \qquad \forall k \in C_i, \forall i \in U_c$$

Dicho modelo formaliza aquellas situaciones en las cuales existe una relación aproximada, de tipo lineal entre la variable de estudio y una variable auxiliar, completamente conocida, X, con el mismo valor en cada uno de los conglomerados.

Este tipo de relaciones suelen darse en situaciones reales en las cuales la variable de estudio presenta cierta homogeneidad en cada conglomerado pero estos tienen comportamientos distintos, aunque con un patrón de tipo lineal en relación a una variable conocida sobre U_c. Por ejemplo, en un estudio económico en el cual las provincias españolas son los conglomerados y los municipios las unidades finales, la renta percápita provincial en un determinado año estará en concomitancia con variables de estudio de tipo económico definidas sobre los municipios como renta percápita municipal en años diferentes, volumen económico de transacciones comerciales, etc.

Bajo este modelo, sustituimos el problema planteado por el de minimizar la expresión,

$$E_s\left[\sum\sum_{i,j\in U_c, i\neq j}\pi^c_{ij}\sum\sum_{k\in C_i, l\in C_j}(Y_N - \max\{Y_k, Y_l\})\right]$$

Denotando por X_{MAX} el valor máximo de la variable X, y por ν el índice del conglomerado al que pertenece la unidad poblacional N-ésima, y suponiendo que $k \in C_i$ y $l \in C_j$, se tiene,

$$\begin{aligned} E_s[(Y_N - \max\{Y_k, Y_l\})] &= E_s[Y_N] - E_s[\max\{Y_k, Y_l\}] \\ &\leq \alpha + \beta X_\nu - \max\{E_s[Y_k], E_s[Y_k]\} = \alpha + \beta X_\nu - \max\{\alpha + \beta X_i, \alpha + \beta X_j\} \\ &\leq \beta(X_{\mathrm{MAX}} - \max\{X_i, X_j\}) \end{aligned}$$

así pues,

$$\begin{aligned} E_s\left[\sum_{i,j\in U_c, i\neq j}\sum \pi^c_{ij} \sum_{k\in C_i, l\in C_j}\sum (Y_N - \max\{Y_k, Y_l\})\right] \\ \leq \beta \sum_{i,j\in U_c, i\neq j}\sum \pi^c_{ij} \sum_{k\in C_i, l\in C_j}\sum (X_{\mathrm{MAX}} - \max\{X_i, X_j\}) \\ = \beta \sum_{i,j\in U_c, i\neq j}\sum \pi^c_{ij} N_i N_j (X_{\mathrm{MAX}} - \max\{X_i, X_j\}) \end{aligned}$$

con lo cual, los diseños muestrales apropiados para la estimación que estamos realizando son aquellos que minimizan la expresión,

$$\sum_{i,j\in U_c, i\neq j}\sum \pi^c_{ij} N_i N_j (X_{\mathrm{MAX}} - \max\{X_i, X_j\})$$

4.1. Implementación bajo el modelo de regresión lineal por conglomerados

A continuación vamos a estudiar la viabilidad de la metodología descrita, en el sentido de su aplicabilidad a situaciones reales, y en el contexto del modelo de superpoblación considerado.

Como se ha visto, es factible utilizar probabilidades de inclusión de primer orden, π^c_i, proporcionales a los tamaños de los conglomerados, si estos presentaran grandes diferencias. No obstante, en este trabajo suponemos que ello no ocurre, por lo que empleamos un diseño uniforme. Con esta elección, tendremos necesariamente $\pi^c_i = n/M$, $\forall i \in U_c$, y la expresión a minimizar es,

$$\sum_{i,j\in U_c, i\neq j}\sum \pi_{ij}^c N_i N_j (X_{\mathrm{MAX}} - \max\{X_i, X_j\})$$

es decir, una función objetivo lineal en las variables π_{ij}^c, $i \neq j$.

Debido a la simetría de las cantidades π_{ij}^c, tendremos en total $M(M-1)/2$ variables, y teniendo en cuenta que el número de conglomerados no suele ser elevado, en comparación con el tamaño de la población, el problema resultante posee un tamaño razonable para abordar su resolución con los programas usuales. Por ejemplo, para $M = 100$ conglomerados, tendríamos un problema de programación lineal con 4950 variables, abordable por las rutinas de uso común para este tipo de problemas.

En cuanto a las restricciones del problema, tendremos en primer lugar $\pi_{ij}^c > 0$, con objeto de que el diseño permita estimar la varianza de la estimación. Por otra parte, de la relación entre probabilidades de inclusión de primer y segundo orden obtendremos las restricciones,

$$\sum_{j\in U_c, j\neq i} \pi_{ij}^c = (n-1)\pi_i^c = \frac{(n-1)n}{M} \qquad \forall i \in U_c$$

y finalmente, con objeto de poder construir estimaciones no negativas de la varianza de la estimación, introducimos las restricciones,

$$\pi_{ij}^c \leq \pi_i^c \pi_j^c = \frac{n^2}{M^2} \qquad \forall i, j \in U_c,\ i \neq j$$

En el apéndice de este trabajo se exponen algunos aspectos computacionales, incluyendo una rutina de implementación en el lenguaje de descripción de problemas AMPL, Fourer *et al.* (1993), así como un ejemplo numérico sobre varias poblaciones.

5. CONCLUSIONES

En primer lugar, es interesante observar que las metodologías clásicas estudiadas en el muestreo en poblaciones finitas, para estimar los parámetros usuales como medias y totales, no son las más apropiadas para estimar parámetros de tipo funcional como la función de distribución poblacional. En efecto, cuando se estiman parámetros puntuales como medias y totales poblacionales, la reducción de varianza se consigue a través de probabilidades de inclusión de primer orden proporcionales a una variable auxiliar relacionada con la de estudio. Por contra, el la estimación de la función de distribución, este planteamiento carece de sentido por la estructura especial de dicho parámetro.

La metodología introducida en este trabajo, basada en minimizar una determinada norma de la función $V[\widehat{F}(t)]$, en este caso la norma $\|\cdot\|_1$, se manifiesta como una alternativa prometedora pues minimizando dicha norma también disminuimos globalmente dicha función de varianza.

Para el caso de que exista relación entre la variable de estudio y una variable auxiliar, formalizada mediante un modelo de superpoblación del tipo de regresión lineal por conglomerados, la reducción de varianza se realiza a partir de la resolución de un problema de programación lineal que proporciona las probabilidades de inclusión de segundo orden. Los resultados numéricos obtenidos en el estudio empírico que se muestra en el apéndice indican que esta metodología puede producir de forma efectiva una reducción del error de la estimación.

6. APÉNDICE

A continuación, exponemos de forma esquemática una rutina en el lenguaje de descripción de problemas AMPL, para el tratamiento del problema considerado. Obsérvese que la restricción $\pi_{ij}^c > 0$, intratable de forma práctica, ha sido sustituida por $\pi_{ij}^c \geq \varepsilon$, siendo $\varepsilon > 0$ un número real muy próximo a cero.

```
param M;
set U:= 1..M;
set PAIRS:={i in U,j in U: i < j};
param X{U} >= 0;
var PI{PAIRS} >= epsilon <=1;
minimize FUNC: sum{(i,j) in PAIRS} (PI[i,j]*Ni*Nj
                                    *(XMAX-max(X[i],X[j])));
subject to RELATION {i in U} :  sum{j in U : j > i} PI[i,j]
                 + sum{j in U : j < i} PI[j,i] = (n-1)*n/M;
subject to BOUND {(i,j) in PAIRS}: PI[i,j] <= n*n/M*M;
```

Como aplicación, vamos a realizar un estudio empírico comparativo, empleando tres poblaciones, U_1, U_2 y U_3, cada una con $M = 20$ conglomerados, con tamaños iguales $N_i = 100$, $\forall i$. Estas poblaciones son artificiales pero indicativas de distintas situaciones que podemos encontrar en problemas reales, y han sido generadas a partir del modelo de superpoblación considerado.

Para la población U_1, hemos empleado los valores $X_i = 20 + i$, $i = 1, \ldots, 20$, siendo los valores poblacionales generados según el modelo,

$$Y_k = 10 + 2X_i + \varepsilon_k \qquad \forall k \in C_i,\ \forall i \in U_c$$

con $\varepsilon_k \sim N(0, 3^2)$, es decir, normales de esperanza $\mu = 0$ y varianza $\sigma^2 = 3^2$.

Para la población U_2, hemos empleado los valores $X_i = i^2$, $i = 1,\ldots,20$, siendo los valores poblacionales generados según el modelo,

$$Y_k = 20 + 2X_i + \varepsilon_k \qquad \forall k \in C_i,\ \forall i \in U_c$$

con $\varepsilon_k \sim N(0, 5^2)$.

Finalmente, para la población U_3, hemos tomado los valores X_i, $i = 1,\ldots,20$, dados por,

$$\{1,5,50,51,53,55,57,100,120,150,155,160,165,170,175,180,185,190,500,1000\}$$

con los valores poblacionales generados según el modelo,

$$Y_k = 20 + 2X_i + \varepsilon_k \qquad \forall k \in C_i,\ \forall i \in U_c$$

siendo $\varepsilon_k \sim N(0, 10^2)$.

Como puede verse, la primera población presenta valores de la variable de estudio con cierta homogeneidad. La segunda es menos homogénea y la tercera tiene una estructura muy dispar, con ciertos valores extremos muy distantes de la masa principal. En todos los caso, se supone muestreo aleatorio simple, MAS(100, 10), para el muestreo dentro de los conglomerados, es decir, en cada uno de los conglomerados muestreados en la primera etapa se extraen muestras aleatorias simples de 10 unidades finales.

Vamos a comparar la metodología considerada en este trabajo con el muestreo aleatorio simple de conglomerados, para lo cual utilizaremos como medida de eficiencia relativa la siguiente cantidad, expresada como porcentaje,

$$C = 100 \times \frac{\|V[\widehat{F}(t)]\|_1^{\mathrm{MAS}} - \|V[\widehat{F}(t)]\|_1^{\mathrm{MET}}}{\|V[\widehat{F}(t)]\|_1^{\mathrm{MAS}}} \%$$

siendo $\|V[\widehat{F}(t)]\|_1^{\mathrm{MAS}}$ y $\|V[\widehat{F}(t)]\|_1^{\mathrm{MET}}$ respectivamente las normas $\|\cdot\|_1$ de la varianza expuesta en el Teorema 4, para el muestreo aleatorio simple y para el método estudiado.

Observemos que en $\|V[\widehat{F}(t)]\|_1^{\mathrm{MET}}$ aparecen las probabilidades de inclusión de segundo orden obtenidas por minimización, mientras que en $\|V[\widehat{F}(t)]\|_1^{\mathrm{MAS}}$ aparecen las correspondientes al muestreo aleatorio simple de conglomerados, cuyos valores son $\pi_{ij}^c = n(n-1)/M(M-1)\ i \neq j$.

Un coeficiente mayor que cero indicará un aumento de eficiencia con respecto al muestreo aleatorio simple, bajo el criterio considerado, y el correspondiente porcentaje indicará la cuantía de dicho aumento. Los resultados comparativos obtenidos se exponen a continuación para cada una de las tres poblaciones y para tamaños muestrales $n = 3,4,5$, siendo n el número de conglomerados muestreados en la primera etapa. Las cantidades reflejadas en esta tabla han sido obtenidas por computación directa a partir de los valores óptimos de las funciones objetivos proporcionados por la rutina AMPL expuesta anteriormente.

Coeficiente C para U_1, U_2 y U_3 y tamaños de muestra de conglomerados $n = 3,4,5$. En todos los casos, en cada conglomerado se extraen 10 unidades finales.

n	U_1	U_2	U_3
3	78.21 %	66.68 %	34.32 %
4	87.08 %	75.39 %	38.59 %
5	91.64 %	80.10 %	42.78 %

Podemos observar como en todos los casos se ha obtenido una evidente reducción de la varianza lo que manifiesta que la metodología estudiada en este trabajo se presenta como una alternativa prometedora.

REFERENCIAS

Chambers, R. L. y Dunstan, R. (1986). «Estimating distribution functions from survey data». *Biometrika*, 73, 597-604.

Chambers, R. L., Dorfman, A. H. y Hall, P. (1992). «Properties of stimators of the finite population distribution function». *Biometrika*, 79, 577-582.

Fernández, F. R. y Mayor, J. A. (1995). *Muestreo en poblaciones finitas: curso básico*. Barcelona: E.U.B.

Fourer, R., Gay, D. M. y Kernighan, B. W. (1993). *AMPL. A Modeling Language for Mathematical Programming*. Danvers, Massachusetts: Boyd & Fraser Publishing Company.

Hill, B. M. (1968). «Posterior distribution of percentiles: Bayes' theorem for sampling from a population». *Journal of the American Statistical Association*, 63, 677-691.

Kuk, A. Y. C. (1988). «Estimation of distribution functions and medians under sampling with unequal probabilities». *Biometrika*, 75, 97-103.

Kuk, A. Y. C. y Mak, T. K. (1989). «Median estimation in the presence of auxiliary information». *Journal of the Royal Statistical Society*, Series B, 51, 261-269.

Rao, J. N. K., Kovar, J. G. y Mantel, H. J. (1990). «On estimating distribution functions and quantiles from survey data using auxiliary information». *Biometrika*, 77, 365-375.

Rao, J. N. K. (1994). «Estimating totals and distributions functions using auxiliary information in the estimation stage». *Journal of Official Statistics*, 10, 153-166.

Sedransk, J. y Meyer, J. (1978). «Confidence intervals for the quantiles of a finite population: simple random and stratified simple random sampling». *Journal of the Royal Statistical Society*, Series B, 40, 239-252.

Woodruff, R. S. (1952). «Confidence intervals for medians and other position measures». *Journal of the American Statistical Association*, 47, 635-646.

ENGLISH SUMMARY

ESTIMATING DISTRIBUTIONS FUNCTIONS USING APPROPRIATE SAMPLING DESIGNS IN TWO STAGE CLUSTER SAMPLING

J. A. MAYOR GALLEGO
M. MARTÍNEZ BLANES
Universidad de Sevilla*

In order to estimate the distribution function of a variable defined over a finite population, we can use a sampling strategy defined by the Horvitz-Thompson estimator and a sampling design. The variance of this estimation is a real function whose minimization in a suitable criterion let us find some desirable properties of the appropriate sampling designs.
In this paper we use the $\|\cdot\|_1$ norm of the variance function as a minimization criteria. This way, under the hypothesis of uniform sampling design, that is to say, with equal first order inclusion probabilities, we study a procedure to obtain appropriate designs under two stage cluster sampling.

Keywords: Sample survey, sampling design, distribution function estimation, cluster sampling

AMS Classification (MSC 2000): 62D05

*Dpto. de Estadística e Investigación Operativa. Universidad de Sevilla. Facultad de Matemáticas. c/ Tarfia s/n, 41012 Sevilla, España. E-mail:mayor@cica.es

–Received March 2001.
–Accepted December 2001.

Usually, the theory of sampling from finite populations is centered on the point estimation of some parameters as the finite population means, variances and ratios. In this paper, we consider the estimation of a functional parameter, the distribution function in relation to a numerical variable, defined over the population.

In literature we can find different approaches to this estimation problem. Chambers and Dunstan (1986) assume a model-based approach to develop an estimating procedure. Kuk (1988) studies several estimators of the distribution function under sampling with unequal probabilities, proportional to an auxiliary variable and Rao, Kovar and Mantel (1990) by means of the auxiliary information. In the same lines, we also can cite the papers of Chambers *et al.* (1992) and Rao (1994).

We propose an alternative approach based on the application of an average-type criterion to the mean square error of the distribution function estimation, in order to find the more appropriate selection probabilities of the clusters.

Let us consider a finite population $U = \{1,2 \ldots N\}$ and let Y denote the numerical survey variable of interest. Let Y_i be the value of Y for the ith population element, with ordered values $0 \leq Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(N)}$. The aim is to estimate the distribution function of the Y variable,

$$F(t) = \frac{1}{N} \sum_{i \in U} I_{[Y_i, +\infty)}(t)$$

where $I_{[Y_i,+\infty)}(t)$, $i \in U$ are the indicator functions of the $[Y_i, +\infty)$ intervals.

If we assume that s is a sample obtained from U whit a sampling design $(S, p(\cdot))$, and $\widehat{F}(t)$ is an estimator of $F(t)$, the classical way to measure the precision of this estimator is to study of the variance,

$$V[\widehat{F}(t)] = \sum_{s \in S} (\widehat{F}(t) - F(t))^2 p(s)$$

Note that the variance is a real function with different values depending on t, therefore it is not possible to use this function for a direct comparison. An alternative way to evaluate the discrepancy between $F(t)$ and $\widehat{F}(t)$ is to apply an average type criterion over the variance, considering the quantity,

$$\|V[\widehat{F}(t)]\|_1 = \int_{Y_{(1)}}^{Y_{(N)}} V[\widehat{F}(t)]\, dt$$

as such, we can search the more appropriate sampling designs minimizing $\|V[\widehat{F}(t)]\|_1$.

In section 2 of this paper we apply this approach for a sampling strategy under two-stage cluster sampling, supposing uniform sub-sampling into the clusters. Thus, if we assume that sub-sampling is performed by means of sampling designs d_i, $i \in U_c$, we obtain the following result in relation to the variance of the estimation.

Theorem. *If the sampling designs d_i, $i \in U_c$ are uniform, with respectively fixed sizes n_i, we have,*

$$\begin{aligned}\|V[\widehat{F}(t)]\|_1 &= \frac{1}{N^2}\sum_{i,j\in U_c}\sum \frac{\pi^c_{ij}}{\pi^c_i\pi^c_j}\sum_{k\in C_i,l\in C_j}\sum (Y_N - \max\{Y_k,Y_l\}) \\ &\quad -\frac{1}{N^2}\sum_{i,j\in U_c}\sum\sum_{k\in C_i,l\in C_j}\sum (Y_N - \max\{Y_k,Y_l\}) \\ &\quad +\frac{1}{N^2}\sum_{i\in U_c}\frac{1}{2\pi^c_i}\left(\sum_{k,l\in C_i}\sum |Y_k - Y_l| - \frac{N_i^2}{n_i^2}\sum_{k,l\in C_i}\sum \pi^i_{kl}|Y_k - Y_l|\right)\end{aligned}$$

Furthermore, if d_i =SRS(N_i,n_i), $i \in U_c$, that is to say, simple random sampling without replacement, the last term becomes,

$$\frac{1}{N^2}\sum_{i\in U_c}\frac{N_i}{\pi^c_i}\left(\frac{N_i}{n_i}-1\right)\frac{1}{2N_i(N_i-1)}\sum_{k,l\in C_i}\sum |Y_k - Y_l|$$

□

In order to reduce the sampling error, we develop in sections 3 and 4 a practical procedure based on linear programming, to compute the second order inclusion probabilities of the more appropriate sampling designs over the cluster population, under the super-population model,

$$\boxed{Y_k = \alpha + \beta X_i + \varepsilon_k, \quad \beta > 0, \quad E_s[\varepsilon_k] = 0 \qquad \forall k \in C_i,\ \forall i \in U_c}$$

where X is an auxiliary variable, entirely controlled.

Finally, we have applied this procedure to three artificial populations with different structures. These populations have $M = 20$ clusters, and every cluster contains $N_i = 100$ elements. The sample sizes in the first stage are $n = 3,4,5$ cluster. In the second stage, $n_i = 10$ elements are drawn by means of simple random sampling.

For the population U_1, the auxiliary variable is $X_i = 20+i$, $i = 1,\ldots,20$, and the super-population model is,

$$Y_k = 10 + 2X_i + \varepsilon_k \qquad \forall k \in C_i,\ \forall i \in U_c$$

with $\varepsilon_k \sim N(0, 3^2)$, that is to say, normal distribution with expectation $\mu = 0$ and variance $\sigma^2 = 3^2$.

For the population U_2, the auxiliary variable is $X_i = i^2$, $i = 1, \ldots, 20$, and the model,

$$Y_k = 20 + 2X_i + \varepsilon_k \qquad \forall k \in C_i,\ \forall i \in U_c$$

with $\varepsilon_k \sim N(0, 5^2)$.

For the population U_3, the auxiliary variable is X_i, $i = 1, \ldots, 20$, with values,

$$\{1, 5, 50, 51, 53, 55, 57, 100, 120, 150, 155, 160, 165, 170, 175, 180, 185, 190, 500, 1000\}$$

with the model,

$$Y_k = 20 + 2X_i + \varepsilon_k \qquad \forall k \in C_i,\ \forall i \in U_c$$

with $\varepsilon_k \sim N(0, 10^2)$.

We compare our methodology (MET) with the simple random sampling (SRS) of clusters, using the following relative efficiency measure,

$$C = 100 \times \frac{\|V[\widehat{F}(t)]\|_1^{\mathrm{SRS}} - \|V[\widehat{F}(t)]\|_1^{\mathrm{MET}}}{\|V[\widehat{F}(t)]\|_1^{\mathrm{SRS}}} \%$$

where $\|V[\widehat{F}(t)]\|_1^{\mathrm{SRS}}$ and $\|V[\widehat{F}(t)]\|_1^{\mathrm{MET}}$ are respectively the $\|\cdot\|_1$ of the variance of the compared methods. This way, we have obtained the following values of C,

n	U_1	U_2	U_3
3	78.21 %	66.68 %	34.32 %
4	87.08 %	75.39 %	38.59 %
5	91.64 %	80.10 %	42.78 %

These results show that we can consider our approach as a promising alternative, in order to reduce the sampling error estimating the finite population distribution function.

QÜESTIIÓ, vol. 26, 1-2, p. 109-122, 2002

ANÁLISIS FACTORIAL MÚLTIPLE COMO TÉCNICA DE ESTUDIO DE LA ESTABILIDAD DE LOS RESULTADOS DE UN ANÁLISIS DE COMPONENTES PRINCIPALES

E. ABASCAL FERNÁNDEZ*
Mª. I. LANDALUCE CALVO**

Una característica de los métodos factoriales es que siempre producen resultados y éstos no son una simple descripción, sino que ponen de manifiesto la estructura existente entre los datos, de ahí la necesidad de estudiar la validez de los resultados. Es preciso analizar la naturaleza de esta estructura y estudiar la estabilidad de los resultados. Consideramos que el mejor criterio es el análisis de la estabilidad de los mapas obtenidos en el análisis factorial.
El Análisis Factorial Múltiple (AFM), desarrollado por B. Escofier and J. Pagès (1992), permite el análisis simultáneo de tablas en las que un mismo conjunto de individuos se describe a través de varios grupos de variables. En este trabajo se muestra la eficacia de este método para el análisis de la estabilidad de los resultados obtenidos mediante Análisis de Componentes Principales.

Multiple Factor Analysis as a statistical technique to study the stability of the results obtained by Principal Component Analysis

Palabras clave: Análisis factorial múltiple, análisis de componentes principales y estabilidad

Clasificación AMS (MSC 2000): 62-07, 62H25

* Profesora Titular de Universidad. Departamento de Estadística e Investigación Operativa. Universidad Pública de Navarra.
** Profesora Titular de Universidad. Departamento de Economía Aplicada. Universidad de Burgos.

- Recibido en septiembre de 2000.
- Aceptado en enero de 2002.

1. INTRODUCCIÓN

Una característica de los métodos factoriales es que siempre producen resultados más o menos interpretables y éstos no son una simple descripción, sino que ponen de manifiesto la estructura existente entre los datos, de ahí la necesidad de estudiar la validez de los resultados. Es preciso analizar si representan una estructura existente entre ellos o simplemente es debida a las fluctuaciones de los datos o a la definición y codificación de las variables.

Existen diversas formas de verificar el significado de un análisis. En este trabajo, siguiendo a Lebart (1995), se considera que el mejor criterio de validación consistirá en verificar la estabilidad de las formas obtenidas en un análisis factorial. El objetivo es mostrar la eficacia del método de Análisis Factorial Múltiple (AFM) para verificar la estabilidad de los resultados de un Análisis de Componentes Principales (ACP).

2. LOS MÉTODOS DE VALIDACIÓN

Un mapa se considera estable si su forma permanece aproximadamente igual cuando se producen pequeñas alteraciones en los datos, es decir, si la orientación definida en el mismo no está determinada por aspectos aislados de los datos.

Greenacre (1993) considera dos tipos de estabilidad, interna y externa. La calidad o estabilidad interna puede verse afectada por la elección de las variables, la unidad de medida, la codificación o el peso, así como por los errores de medida. En cuanto a la estabilidad externa, estudia si los datos son válidos como representativos de una población. Se considera estable si se obtiene ésta al considerar nuevas muestras. Esta última forma de estabilidad sólo tiene sentido estudiarla cuando los datos proceden de un muestreo.

Los estudios de la estabilidad interna se realizan mediante métodos empíricos. Estos métodos trabajan sobre modificaciones de la tabla inicial y permiten verificar su estabilidad a través del mantenimiento de la configuración obtenida en el análisis, Lebart (1995). Las modificaciones que se generan van destinadas a estudiar aquellos elementos que pueden incidir sobre la calidad y estabilidad de los resultados del análisis.

Estas alteraciones de la tabla inicial se pueden producir en la definición y número de variables, o bien en perturbaciones de los datos, de diferentes formas, entre otras:

a) Por omisión de una línea para estudiar si la influencia de ésta es decisiva.

b) Agrupamiento de varias columnas.

c) En la elección de las variables que definen el problema.

d) En la codificación o definición de las variables.

e) Alterando los datos mediante la suma de un error aleatorio. Es decir, afectando los datos con perturbaciones aleatorias, de esta forma se simulan errores de medida en las variables. Así, cada observación se modifica de la forma $x_{ij}^{(r)} = x_{ij} + z_{ijr}$ siendo z_{ijr} un valor obtenido aleatoriamente de la distribución normal de media 0 y desviación típica $\frac{1}{k}s_j$. Para diferentes valores de k se obtienen nuevos valores.

Al producir estas alteraciones se generan nuevas tablas. En todos estos casos, se dispone de una tabla original y de otras que mantienen los mismos individuos, pero pueden diferenciarse en el valor de los datos, o bien en la definición o número de las variables. El objetivo será estudiar si la configuración de las representaciones gráficas de las diferentes tablas es la misma o, si por el contrario, se producen alteraciones considerables.

Los métodos empíricos utilizados habitualmente para determinar el número de factores estables, realizan los análisis por separado de cada tabla y miden la correlación entre los factores obtenidos en cada uno de ellos y los generados en el análisis de la tabla original. Si los ejes son estables los factores obtenidos en la diversas tablas estarán altamente correlacionados. En caso contrario, si las correlaciones son semejantes con factores de distinto orden, el eje es arbitrario resultado del azar, y no se considera estable.

En este trabajo se propone un método empírico para el análisis de la estabilidad de los resultados proporcionados por la técnica exploratoria Análisis de Componentes Principales (ACP). El procedimiento consiste en generar diferentes tablas mediante alguna de las alteraciones indicadas y analizarlas simultáneamente utilizando una metodología de análisis de tablas múltiples, el Análisis Factorial Múltiple (AFM). Este análisis permitirá la verificación de la estabilidad de los factores. La forma de alteración de la tabla inicial seleccionada para ilustrar la bondad del método ha sido la última de las anteriormente citadas, esto es, mediante la suma de errores aleatorios. A continuación se presenta una breve exposición de los principios del AFM.

3. EL ANÁLISIS FACTORIAL MÚLTIPLE COMO TÉCNICA DE VALIDACIÓN

El AFM, desarrollado por Escofier y Pagès, (1992), en el seno de la Escuela Francesa de Análisis de Datos, es un método factorial adaptado al tratamiento de tablas de datos en las que un mismo conjunto de individuos se describe a través de varios grupos de variables. Los grupos de variables pueden surgir de la utilización conjunta de variables de diferente naturaleza, cuantitativas y cualitativas, del empleo de tablas que provienen de otras de tres dimensiones o del manejo de un mismo conjunto de variables medidas en distintos periodos de tiempo.

El AFM se basa en la metodología de ACP y actúa en 2 etapas:

1) A cada grupo de variables se asocia una configuración euclídea o nube de individuos denominada *nube parcial*, que será analizada por separado, obteniendo los factores parciales. Es decir, realiza un ACP de cada una de las tablas X_j, donde X_j recoge la valoración de las variables del grupo $j, j = 1,2,\ldots J$, en el conjunto de individuos.
2) Realiza un ACP de la tabla global resultado de yuxtaponer las J tablas X_j. En este análisis cada tabla X_j es ponderada por el inverso del primer valor propio obtenido en el ACP de la propia tabla, $1/\lambda_j$. Esta ponderación mantiene la estructura de cada tabla, ya que todas las variables han recibido la misma ponderación, pero consigue equilibrar la influencia de los grupos, ya que la inercia máxima de cada una de las nubes de individuos, definida por los diferentes grupos, vale 1 en cualquier dirección.

El objetivo global del análisis es examinar la existencia de estructuras comunes a todas, o a parte, de las nubes parciales.

4. ANÁLISIS DE LA ESTABILIDAD MEDIANTE AFM

El AFM es una valioso instrumento para el estudio de la estabilidad de los resultados del ACP. Se considera que la tabla inicial, la analizada mediante ACP, constituye un grupo de variables medido sobre un conjunto de individuos, y cada una de las alteraciones generadas para medir la estabilidad constituyen un nuevo grupo.

El AFM permite el análisis simultáneo, equilibrando la influencia, de las tablas generadas con las distintas alteraciones y proporciona, como veremos, numerosos indicadores que permiten analizar la estabilidad. Genera además representaciones gráficas de gran poder ilustrativo.

a) El estudio de la intra-estructura o compromiso permite detectar el número de factores comunes a las tablas. Si las tablas tienen la misma estructura, es decir, si el análisis es estable, se obtendrán factores comunes a todas las tablas y, en consecuencia, el número de éstos indicará el de factores estables.

Para este objetivo el AFM proporciona como principal indicador las inercias intra de los puntos individuo. Así, aquellos individuos cuyos puntos parciales (puntos que representan a cada individuo desde los diferentes grupos) se sitúen próximos (inercia intra débil) ilustran la estructura común de las distintas tablas analizadas. Por el contrario, aquellos individuos con puntos parciales asociados estén alejados unos de otros (inercia intra alta), constituyen las excepciones a la estructura común.

b) El estudio de la interestructura analiza la proximidad entre las diferentes tablas. Se realiza a través de tres acciones:

- La representación gráfica de los grupos. La coordenada de un grupo sobre un factor es la inercia acumulada del grupo sobre el eje del AFM. Indica los grupos que han determinado en mayor medida los factores y, en consecuencia, el número de factores estables. Además, el distanciamiento de un punto de los demás indica que el grupo correspondiente es el más alejado del compromiso, lo que permite detectar qué alteraciones de la tabla han producido una estructura más alejada de la común.
- Medidas globales de relación entre los grupos, basadas en el coeficiente RV de Escoufier (Dazy, 1996). Este coeficiente se obtiene a partir de los coeficientes de correlación lineal entre dos variables cualesquiera. Su valor está comprendido entre 0 (no existe relación entre las variables de los dos grupos considerados) y 1 (las nubes que representan a los grupos son homotéticas). Para dos grupos cualesquiera j y h se tiene:

$$RV = \frac{\langle W_jD, W_hD\rangle}{\|W_jD\| \ \|W_hD\|} = \frac{\text{traza}(W_jDW_hD)}{\|W_jD\| \ \|W_hD\|}$$

  Siendo $W_jD = X_jX_j'D$ donde X_j es la matriz asociada al grupo j de variables, D una matriz diagonal que recoge los pesos de los individuos. La norma de W_jD se expresa como $\|W_jD\| = \sqrt{\sum_s \left(\lambda_s^j\right)^2}$ donde λ_s^j es el j-ésimo valor propio del análisis de la tabla j. El numerador representa los productos escalares entre los representantes de los grupos.

  Esta medida es completada con los coeficientes Lg (Escofier y Pagès, 1992).

$$L_g(W_jD, W_hD) = \langle W_jD, W_hD\rangle$$

  El producto escalar se interpreta como una medida de la relación entre los grupos. Del estudio de estos coeficientes se puede deducir o rechazar la estabilidad de los resultados obtenidos.
- Los coeficientes de correlación entre los factores parciales (obtenidos en el análisis parcial de cada nube) y los factores comunes o generales. Cuando la correlación es fuerte el factor global traduce una tendencia que está presente en todas las tablas, es decir, se trata de un factor común, obteniéndose una garantía de la estabilidad de los factores.

c) Las representaciones gráficas simultáneas. El AFM obtiene, por un lado, las representaciones gráficas de la nube de individuos caracterizados por las variables de cada grupo (individuos parciales) y por el conjunto de variables (individuo medio) sobre el mismo plano factorial, que permiten observar las proximidades entre los puntos individuo de las diferentes tablas y detectar las posibles deformaciones. Por otro

lado, obtiene la representación de las variables que permite observar las proximidades entre los puntos variables de las diferentes tablas que corresponden a un mismo concepto. El análisis de éstas permitirá estudiar la estabilidad de cada una.

5. ANÁLISIS COMPARATIVO

El procedimiento propuesto es una aproximación no paramétrica en la misma línea que la propuesta por Greenacre en 1984, en la que se analiza el concepto de estabilidad en lugar de confianza. En este sentido, el procedimiento propuesto no es alternativo, sino complementario, a otros métodos paramétricos clásicos de estudio de la estabilidad como el análisis de los intervalos de confianza para los valores propios.

Cuando se trata de analizar la estabilidad interna el procedimiento clásico consiste en eliminar un elemento, cambiarlo, o simular errores en los datos para comprobar la estabilidad. En este caso, el AFM es una metodología que además de los indicadores habituales (correlaciones entre los factores) proporciona nuevas medidas que permiten estudiar la estabilidad de los factores (coeficientes RV y Lg) y detectar los elementos perturbadores, proporcionando, además, medidas de la importancia de éstos (inercias intra).

Una característica importante del AFM son las representaciones gráficas en las que se pueden estudiar las relaciones entre las variables de las diferentes tablas, lo que permite analizar la estabilidad de las variables y estudiar cómo se ven afectadas por las diferentes codificaciones, errores, etc. Si las tablas han sido generadas por remuestreo, esta representación es una forma de obtener las zonas de confianza empíricas para cada variable, semejantes a las zonas de confianza (Convex Hulls) obtenidas para los resultados del análisis de correspondencias múltiples por Greenacre (1984). La nube que contiene todas las representaciones de la misma variable correspondiente a las diferentes réplicas constituye una zona de confianza empírica para la variable.

En el caso de ACP, se han utilizado anteriormente dos procedimientos para obtener las representaciones: analizar la tabla inicial y adjuntar las réplicas como tablas ilustrativas (Lebart, 1995) o bien, analizarlas simultáneamente mediante el método STATIS (Holmes, 1985). Las diferencias entre éstos y a su vez con el AFM radican en el criterio de obtención del espacio común y en la interpretación de las representaciones. En particular, en el procedimiento de proyectar las réplicas como ilustrativas sobre el análisis de una de ellas, significa que solamente una de las tablas ha participado en la creación de los ejes y en consecuencia, no se tiene en cuenta la estructura de relaciones entre las variables de las otras tablas.

En cuanto al método STATIS, que presenta grandes similitudes con el AFM, los ejes factoriales también se obtienen de un compromiso de todas las tablas, sin embargo

su participación no está equilibrada, ya que tienen mayor contribución aquéllas que presentan una estructura interna más fuerte, esto es, con mayores similitudes.

De este breve estudio comparativo con otros métodos utilizados para el estudio de la estabilidad, se deduce que el AFM es un valioso instrumento para la determinación del número de factores y para estudiar la estabilidad de las formas. Únicamente presenta limitaciones para el estudio cuando el elemento sospechoso de producir inestabilidad (por fuerte contribución a la obtención del factor) es un único individuo.

6. ANÁLISIS DE LA ESTABILIDAD DE LA ESTRUCTURA DE LOS SALARIOS MEDIOS DE LAS COMUNIDADES AUTÓNOMAS ESPAÑOLAS

El objetivo de esta aplicación empírica es estudiar la estabilidad de los resultados obtenidos en un ACP de la tabla que recoge la estructura de los salarios medios de las Comunidades Autónomas españolas (CCAA) en el año 1995, a través del AFM. La tabla objeto del análisis recoge los salarios medios por ramas de actividad en las Comunidades Autónomas. Las ramas de actividad consideradas son: agricultura, ganadería y pesca (AGRICULTURA), Energía (ENERGIA), industria (INDUST), construcción (CONST), servicios de mercado (SER.MERCADO), servicios de no mercado (SER.NMER).

El análisis de componentes principales de esta tabla proporciona 2 factores cuyos valores propios son superiores a 1, $\lambda_1 = 3{,}42$ y $\lambda_2 = 1{,}15$. Sin embargo, los intervalos de confianza de estos valores propios tiene intersección no vacía. Para estudiar su estabilidad se generan nuevas tablas con perturbaciones aleatorias, es decir, cada valor de la tabla original se altera mediante la adición de una perturbación generada por una distribución normal cuya varianza es una fracción de la varianza inicial de la variable. Se generan así tres nuevas tablas que corresponden a perturbaciones con varianzas $1\%S_j$, $10\%S_j$, $20\%S_j$ respectivamente y constituyen los grupos 2 a 4, siendo la tabla inicial el grupo 1.

6.1. Análisis de la intraestructura o compromiso

Este análisis consiste en el estudio de las inercias de los puntos de las nubes parciales, con respecto a su centro de gravedad. Para ello se exponen a continuación los dos primeros planos factoriales correspondientes a las nubes de individuos, CCAA (gráfico 1) y de variables, ramas de actividad, (gráfico 2).

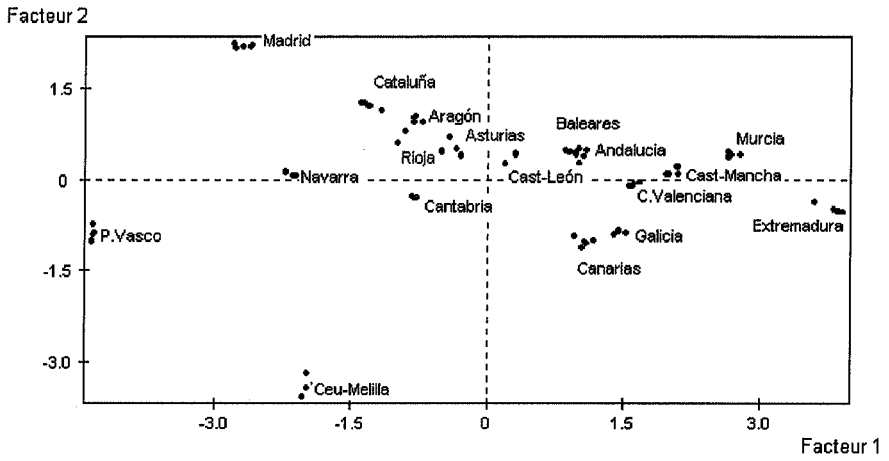


Figura 1. Plano Factorial 1-2: Comunidades Autónomas. Puntos medios y Puntos Parciales.

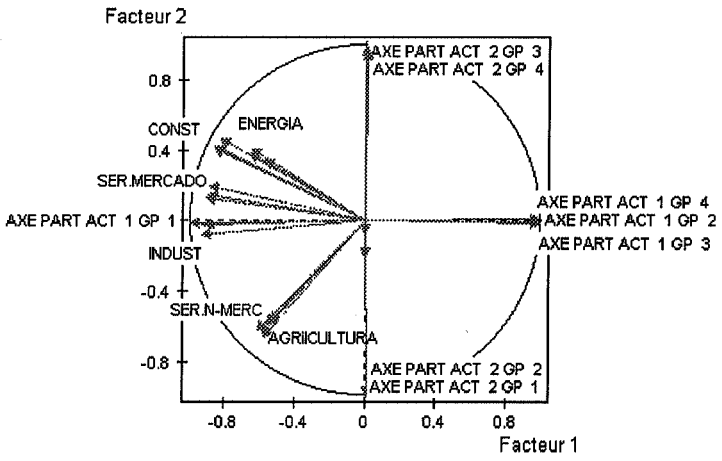


Figura 2. Plano Factorial 1-2: Variables (activas) y Ejes Parciales (suplementarios) de las 4 tablas.

En ambos gráficos se puede observar una gran proximidad entre todos los puntos que representan al mismo individuo (gráf. 1) y a la misma variable (gráf. 2), resultado que pone de manifiesto la existencia de una débil inercia intra y, como consecuencia, una elevada inercia inter. Esto es, las tablas analizadas tienen una estructura muy similar.

Tabla 1. Individuos con las mayores inercias intra.

Eje 1

Individuos	INER	ACUM
RIOJA	24.02	24.02
C. VALENCIANA	16.48	40.51
GALICIA	10.95	51.45

Eje 2

Individuos	INER	ACUM
CANARIAS	23.50	23.50
CEU-MELILLA	14.69	38.19
CAST-MANCHA	13.23	51.42

Además, en lo que se refiere a los individuos el estudio minucioso de las inercias intra, tanto de los puntos medios (tabla 1) como de los puntos parciales que representan a las regiones, permite poner de manifiesto cuáles de ellas presentan un comportamiento más heterogéneo y en qué eje factorial este comportamiento es más acusado, además de señalar cuáles son los puntos parciales responsables del mismo. Así, en el primer eje factorial son las comunidades de La Rioja, de Valencia y Galicia las que tienen las mayores inercias intra. En el segundo eje factorial son Canarias, Ceuta y Melilla y la región de Castilla-La Mancha las que presentan un comportamiento menos homogéneo. Destacar que en todos los casos es el punto-parcial asociado al grupo 4 el mayor responsable de este resultado, resultado lógico si tenemos en cuenta que estos puntos corresponden a la tabla que surge de la mayor transformación aplicada a la tabla original.

Tabla 2. Matrices de correlaciones entre los factores parciales.

	101	102	103	104	105
101	1.00				
102	0.00	1.00			
103	0.00	0.00	1.00		
104	0.00	0.00	0.00	1.00	
105	0.00	0.00	0.00	0.00	1.00

Tabla original

	101	102	103	104	105
301	**−1.00**	0.02	0.00	0.00	0.00
302	−0.02	**−1.00**	0.02	0.03	−0.01
303	0.00	0.01	**−0.99**	0.14	−0.01
304	0.00	−0.03	−0.14	**−0.98**	0.05
305	0.01	0.01	0.00	**−0.06**	−0.97

Perturbación 10%

	101	102	103	104	105
201	**1.00**	0.00	0.00	0.00	0.00
202	0.00	1.00	0.01	0.01	0.00
203	0.00	0.01	**1.00**	0.02	0.00
204	0.00	0.00	0.02	**−1.00**	0.00
205	0.00	0.00	0.00	0.00	**1.00**

Perturbación 1%

	101	102	103	104	105
401	**−0.99**	0.02	−0.01	0.01	0.02
402	−0.02	**−0.95**	0.24	0.11	0.06
403	−0.01	0.26	**0.88**	0.31	0.01
404	0.02	0.04	−0.32	**0.92**	0.01
405	0.01	0.04	0.02	0.05	**0.95**

Perturbación 20%

6.2. Análisis de la Interestructura

Es el estudio comparativo de la proximidad entre las diferentes nubes. La lectura de la matriz de correlaciones entre los factores parciales (tabla 2), pone de manifiesto la

estabilidad de los resultados obtenidos en este estudio empírico. Ello se observa tanto en las fuertes correlaciones entre los factores del mismo orden, correspondientes a las diferentes tablas, como a las correlaciones prácticamente nulas entre los factores de distinto orden. Esto es, esta matriz nos proporciona una visión previa de las similitudes entre las cuatro tablas analizadas, indicando que las primeras direcciones de variabilidad de cada grupo manifiestan estructuras comunes a los mismos.

Tabla 3. Coeficientes Lg de relación entre grupos.

	1	2	3	4
1	**1.178**			
2	1.177	**1.176**		
3	1.169	1.169	**1.168**	
4	1.167	1.166	1.161	**1.197**

Tabla 4. Coeficientes RV de relación entre grupos.

	1	2	3	4
1	1.000			
2	**1.000**	1.000		
3	**0.997**	**0.997**	1.000	
4	**0.983**	**0.983**	**0.982**	1.000

Del estudio de las matrices L y RV (tablas 3 y 4) se deduce, nuevamente, la estabilidad de los resultados obtenidos. Son grupos con una dimensionalidad parecida (se observa en los coeficientes de la diagonal principal de la matriz L) y con una estructura interna prácticamente igual (se observa en los coeficientes de la matriz RV).

Tabla 5. Correlaciones entre las variables canónicas y los factores del análisis global.

CORRELACIONES

FAC.	1	2	3	4	5
GR 1	1.00	1.00	1.00	1.00	0.99
GR 2	1.00	1.00	1.00	1.00	0.99
GR 3	1.00	1.00	0.99	0.99	0.96
GR 4	1.00	0.99	0.97	0.98	0.90

La existencia de factores comunes a todos los grupos también puede ser detectado a través del cálculo del coeficiente de correlación entre el factor global y el correspondiente a cada uno de los grupos analizados (esto es, entre las variables canónicas y las variables generales). Cuando la correlación es fuerte el factor global traduce una tendencia que está presente en todas las tablas, es decir, se trata de un factor común. En

este estudio empírico, las correlaciones son totales para los dos primeros factores y sólo a partir del tercer eje algunos coeficientes descienden levemente. Por tanto, podemos concluir que son ejes que traducen una tendencia presente en todos los grupos.

En este caso, todos estos resultados ponen de manifiesto la estabilidad de los resultados obtenidos en el ACP de la tabla original.

7. CONCLUSIONES

El AFM ha permitido estudiar las relaciones entre los factores de los análisis por separado de cada tabla, de forma semejante a otras técnicas, pero con procedimientos más directos y sencillos. Este procedimiento propuesto no es alternativo, sino complementario, a otros métodos clásicos de estudio de la estabilidad como el análisis de los intervalos de confianza para los valores propios o los métodos de remuestreo (validación cruzada, bootstrap, etc). Sin embargo, como método de validación empírica presenta buenas propiedades ya que, no sólo proporciona medidas de la estabilidad de los ejes sino también, indicadores de cómo afectan las perturbaciones en las diferentes variables a la estabilidad, así como los individuos que más contribuyen a la inestabilidad, aspectos no tratados por los métodos más habituales. Además, proporciona representaciones gráficas de las diferencias entre las tablas y entre sus estructuras sobre un marco de referencia común.

AGRADECIMIENTOS

Los autores desean agradecer los comentarios de dos evaluadores anónimos que sin duda han ayudado a la presentación de este artículo.

REFERENCIAS

Abascal, E. y Grande, I. (1989). *Métodos Multivariantes para la Investigación Comercial. Teoría, Aplicaciones y Programación BASIC.* Ed. Ariel.

Aluja, T. y Morineau, A. (1999). *Aprender de los Datos: El Análisis de Componentes Principales.* EUB Barcelona.

Dazy, F. y Le Barzic, J. F. (1996). *L´Analyse des Données Evolutives.* Technip. Paris.

Escofier, B. y Pagès, J. (1992). *Análisis factoriales simples y múltiples. Objetivos, métodos e interpretación.* Servicio editorial de la Universidad de País Vasco.

Greenacre, M. J. (1984). *Theory and Applications of Correspondence Analysis.* Academic Press London.

Greenacre, M. J. (1993). *Correspondence analysis in practique.* Academic Press London.

Landaluce, Mª. I. (1995). *Estudio de la estructura de gasto medio de las Comunidades Autónomas españolas. Una aplicación del Análisis Factorial Multiple.* Tesis doctoral. Universidad del País Vasco.

Lebart, L., Morineau, A. y Piron, M. (1995). *Statistique exploratoire multidimensionnelle.* Dunod, París.

ENGLISH SUMMARY

MULTIPLE FACTOR ANALYSIS AS A STATISTICAL TECHNIQUE TO STUDY THE STABILITY OF THE RESULTS OBTAINED BY PRINCIPAL COMPONENT ANALYSIS

E. ABASCAL FERNÁNDEZ*
Mª. I. LANDALUCE CALVO**

It is commonly accepted that Factor Analysis always get results and that these results are not only descriptive but also explanatory of the underlying structure existing in the data. There is a need of analysing the nature of this structure, studying the validity of these results.
We consider that the best validity criterion is the analysis of the stability of the maps obtained by Factor Analysis.
The Multiple Factor Analysis (MFA), developed by B. Escofier and J. Pagès (1992), allows a comparative analysis of a weighted group of tables referred to the same individuals (rows) and to the same or different variables (columns) per table. In this paper we point out, that this method gives the appropriate results and graphics to assess the stability of the results obtained by PCA.

Keywords: Multiple Factor Analysis, Principal Component and Stability

AMS Classification (MSC 2000): 62-07, 62H25

* Profesora Titular de Universidad. Departamento de Estadística e Investigación Operativa. Universidad Pública de Navarra.
** Profesora Titular de Universidad. Departamento de Economía Aplicada. Universidad de Burgos.

- Received September 2000.
- Accepted January 2002.

It is commonly accepted that Factor Analysis always get results and that these results are not only descriptive but also explanatory of the underlying structure existing in the data. There is a need of analysing the nature of this structure, studying the validity of these results.

Lebart (1995) consider that the best validity criterion is the analysis of the stability of the maps obtained by Factor Analysis.

A map is considered «stable» in two different ways: firstly at the level of the data matrix itself (internal stability) and, secondly, if the data are a valid sample representative of a population, at the level of the wider population (external stability). In this paper we only consider the first case.

We say that a map is «internally stable» if the orientation of the plane is not determined by isolated aspects of the data. Different ways of investigating this internal stability are possible. We consider modifications of the original table by adding to the data a random error. So, an observation x_{ij} is modified in the following way:

$$x_{ij}^{(r)} = x_{ij} + z_{ijr}$$

where is a normally distributed random variable with 0 mean and $\frac{1}{k}s_j$ standard deviation. The different values of k provide different tables.

The tables thus obtained and the original table are joined together forming a broad table containing the same rows and columns but with differents values.

MFA, developed by B. Escofier and J. Pagès (1992), allows a comparative analysis of a weighted group of tables referred to the same individuals (rows) and to the same or different variables (columns) per table. In this paper we point out, through an empirical application, that this method gives the appropriate results and graphics to assess the stability of the results obtained by PCA.

QÜESTIIÓ, vol. 26, 1-2, p. 123-138, 2002

UN MODELO POISSONIANO PARA PREDECIR LA MATRICULACIÓN DE VEHÍCULOS EN PAÍSES EUROPEOS

M. J. VALDERRAMA
A. M. AGUILERA
P. R. BOUZAS
Universidad de Granada*

En este artículo presentamos una modelización para la matriculación de vehículos en países de Europa mediante un proceso de Poisson Doblemente Estocástico con media aleatoria Normal truncada. Apoyándonos en trabajos previos acerca de este proceso, se amplía el estudio de características de éste. Asímismo, se hace una predicción de este proceso para los años 2000 y 2001.

A Poissonian model to forecast vehicules matriculations in european countries

Palabras clave: Proceso de Poisson Doblemente Estocástico, Distribución Normal truncada, Matriculaciones de Turismos

Clasificación AMS (MSC 2000): 60G55, 60K30

*Departamento de Estadística e I.O., Facultad de Farmacia, Campus de Cartuja, Universidad de Granada, 18071 Granada.

–Recibido en julio de 2001.
–Aceptado en noviembre de 2001.

1. INTRODUCCIÓN

El número de vehículos turismo matriculados constituye uno de los indicadores más representativos de la actividad económica de un país. Numerosos trabajos se han centrado en modelizar dicho proceso y tratar de predecir su evolución a lo largo del tiempo. Así, por ejemplo, García-Ferrer y Queralt (1998), desarollan un modelo predictivo de varias series de indicadores económicos españoles, entre ellos el número de matriculaciones, utilizando modelos de componentes no observadas con parámetros variables.

Dado que la matriculación de automóviles constituye un claro ejemplo de proceso de recuento de tipo Poissoniano, en el presente trabajo presentamos la aplicación de metodología original para modelizar dicho fenómeno en los países europeos, en la que se considera el parámetro del proceso como un término aleatorio que debe igualmente modelizarse a partir de las trayectorias muestrales. De hecho, el proceso de Poisson es insufiente en multitud de casos debido al carácter determinístico de su función de intensidad. El proceso de Poisson Doblemente Estocástico (PPDE) es una generalización del proceso de Poisson introducida por Cox (1955).

El trabajo que presentamos trata un caso particular de PPDE, aquél cuya media es una Normal truncada en cada instante de tiempo. Está principalmente enfocado a la ampliación del estudio de características de ese caso particular de PPDE y a su aplicación al caso real de proceso puntual que hemos comentado.

Más concretamente, en la Sección 2 se hace un resumen teórico del PPDE recordando aquellas características que más adelante se necesitarán. Se particulariza el caso en el que la media sea una Normal truncada en cada instante de tiempo, esquematizando los cálculos que conducen a la expresión explícita de la esperanza y varianza de este tipo de PPDE.

En la Sección 3, se muestra un método de predicción del PPDE con Normal truncada en el futuro. El método está basado en la modelización de la función paramétrica del proceso puntual y su estimación para un momento futuro. En la última sección, se aplica lo presentado en las secciones anteriores al caso real del proceso puntual número de automóviles demandados en Europa. Se muestran los resultados de cada uno de los pasos del método de predicción y se finaliza dando una predicción del proceso para el año 2000, que se compara con el dato muestral para ese año, y una predicción para el año 2001.

2. PPDE CON MEDIA ALEATORIA NORMAL

Intuitivamente, un PPDE es un proceso de recuento de Poisson cuya intensidad viene dada en cada instante por un proceso externo al proceso puntual en estudio. Notare-

mos al PPDE como $\{N(t) : t \geq t_0\}$, donde $N(t)$ indica el número de sucesos ocurridos desde t_0 hasta t. Más concretamente, sea $\{x(t) : t \geq t_0\}$ un proceso continuo a la izquierda, entonces se dice que el proceso de recuento $\{N(t) : t \geq t_0\}$ es un proceso de Poisson doblemente estocástico con proceso de intensidad $\{\lambda(t,x(t)) : t \geq t_0\}$, si para cada trayectoria de $\{x(t) : t \geq t_0\}$, $N(\cdot)$ es un proceso de Poisson con función de intensidad $\{\lambda(t,x(t)) : t \geq t_0\}$. Es decir, si $\{N(t) : t \geq t_0\}$ condicionado a $\{x(t) : t \geq t_0\}$ es un proceso de Poisson con función de intensidad $\{\lambda(t,x(t)) : t \geq t_0\}$. Al proceso $\{x(t) : t \geq t_0\}$ se le denomina proceso información. El PPDE ha sido ampliamente estudiado por Daley and Vere-Jones (1988), Snyder and Miller (1991), Valderrama *et al.* (1995), entre otros.

Mediante el método de condicionamiento, se tiene como principal estadístico de recuento que la probabilidad de que el número de puntos o sucesos que ocurran en $[t_0,t)$ sea n es igual a:

$$(1) \qquad P[N(t)=n] = E\{P[N(t)=n/x(\sigma) : t_0 \leqslant \sigma < t]\}$$
$$= E\left\{\frac{1}{n!}\Lambda(t,x(t))^n \exp[-\Lambda(t,x(t))]\right\} = G_{\Lambda}^{n)}(-1)$$

para $n = 0,1,2,\ldots$, donde $\Lambda(t,x(t))$ es la función paramétrica del PPDE, es decir $\Lambda(t,x(t)) = \int_{t_0}^{t} \lambda(\sigma,x(\sigma))d\sigma = E[N(t)/x(\sigma) : t_0 \leqslant \sigma < t]$ y $G_{\Lambda}^{n)}(s)$ es la derivada n-*ésima* de la función generatriz de momentos (que supondremos existe) de $\Lambda(t,x(t))$.

Usando nuevamente el método de condicionamiento, la función característica de un PPDE, usando la notación de Snyder y Miller (1991), resulta ser,

$$(2) \qquad M_{N(t)}(iu) = E\left\{\exp\left[(e^{iu}-1)\int_{t_0}^{t}\lambda(\sigma,x(\sigma))\,d\sigma\right]\right\}$$
$$= E\left\{\exp\left[(e^{iu}-1)\,\Lambda(t,x(t))\right]\right\} = M_{\Lambda}(e^{iu}-1)$$

donde $M_{\Lambda}(iu)$ es la función característica de $\Lambda(t,x(t))$.

Aunque se tenga esta expresión de la función característica, su evaluación no es siempre posible. Sí que a partir de ella resulta fácil observar que

$$E[N(t)/x(\sigma) : t_0 \leqslant \sigma < t] = E[\Lambda(t,x(t))]$$
$$Var[N(t)/x(\sigma) : t_0 \leqslant \sigma < t] = Var[\Lambda(t,x(t))] + E[\Lambda(t,x(t))]$$

Después de haber trabajado con distintos procesos de recuento (ej. número de hipotecas, número de efectos bancarios devueltos o impagados, etc.) parece que a menudo su media podría modelizarse mediante un modelo Gaussiano. Puesto que $E[N(t)/x(\sigma)$:

$t_0 \leqslant \sigma < t] = \Lambda(t,x(t))$ y la función paramétrica debe ser mayor o igual que cero y creciente, para evitar problemas técnicos sin apartarse demasiado del modelo Gaussiano hemos estudiado el caso de que la media, $\Lambda(t,x(t))$, se distribuya mediante una Normal truncada en cada instante de tiempo. De hecho, ha resultado ser una acertada elección como veremos más adelante.

Si la función paramétrica se distribuye mediante una Normal truncada, que notaremos mediante $\Lambda(t,x(t)) \rightsquigarrow N_T(A(t),B(t),\mu(t),\sigma(t))$, haciendo uso de la expresión de $G_\Lambda^{n)}(-1)$ para esta distribución y la ecuación (1), se puede demostrar que la función masa de probabilidad del PPDE con esa media es (Rodríguez Bouzas *et al.*, 2001)

$$P[N(t)=n] = \frac{1}{n!\sigma\sqrt{2\pi}}\left(P\left[\frac{A-\mu}{\sigma}<Z<\frac{B-\mu}{\sigma}\right]\right)^{-1}\int_A^B \Lambda^n e^{-\Lambda} e^{-\frac{1}{2}\left(\frac{\Lambda-\mu}{\sigma}\right)^2} d\Lambda$$

para $n = 0,1,2,\ldots$ donde A, B, μ y σ son los parámetros de $\Lambda(t,x(t))$, todos ellos funciones del tiempo y tal que $0 \leq A < \Lambda < B$, $\mu \in \mathbb{R}$, $\sigma \geq 0$ con A y B los puntos de truncamiento inferior y superior, respectivamente.

Usando la ecuación (11) y la función característica de $\Lambda(t,x(t))$, cuya expresión es

$$M_\Lambda(iu) = \frac{P[\frac{A-\mu-iu\sigma^2}{\sigma}<Z<\frac{B-\mu-iu\sigma^2}{\sigma}]}{P[\frac{A-\mu}{\sigma}<Z<\frac{B-\mu}{\sigma}]}\exp\left[iu\mu+\frac{1}{2}(iu)^2\sigma^2\right]$$

se tiene que la función característica del PPDE que nos ocupa es de la forma

$$M_{N(t)}(iu) = C(iu,t)\cdot\exp\left[\mu(e^{iu}-1)+\frac{1}{2}(e^{iu}-1)^2\sigma^2\right] \tag{3}$$

donde

$$C(iu,t) = \frac{P\left[A-(e^{iu}-1)\sigma^2<\xi<B-(e^{iu}-1)\sigma^2\right]}{P[A<\xi<B]}$$

y ξ es una variable con distribución $N(\mu,\sigma)$, recordando que μ y σ son funciones del tiempo. La función $C(iu,t)$ que aparece en la ecuación (12) es sólo una función del tiempo dados unos μ y σ particulares. Observando además la exponencial que aparece en la misma expresión, se puede escribir abreviadamente

$$M_{N(t)}(iu) = C(iu,t)\,M_\xi\left(e^{iu}-1\right) \tag{4}$$

También se puede interpretar la función característica del proceso como el producto de $C(iu,t)$ por la función característica de un proceso de Poisson $\{N^*(t);t \geq t_0\}$ con

intensidad gaussiana $\{\lambda^*(t,x(t)); t \geq t_0\}$ tal que

$$E[\lambda^*(t,x(t))] = \mu'(t) \text{ y } \int_{t_0}^{t}\int_{t_0}^{t} R(u,v)\,du\,dv = \frac{1}{2}\sigma^2(t)$$

donde $R(u,v)$ es la función de covarianza de $\lambda^*(t,x(t))$, ver el artículo de Valderrama *et al.* (1995), por lo que tendríamos

$$M_{N(t)}(iu) = C(iu,t)\,M_{N^*(t)}(iu) \tag{5}$$

Ambas formas de escribir la función característica del PPDE con media Normal truncada, nos ayudan notablemente a la hora de hacer los cálculos para hallar la media y varianza del proceso, que serían imposibles de realizar usando sus correspondientes definiciones, puesto que los momentos de orden 1 y 2 de la variable ξ son ya por todos conocidos, así como los del proceso $N^*(t)$ calculados en el citado artículo.

Usando (4), tenemos que la esperanza se calcularía

$$E[N(t)] = E[\Lambda(t)] = \left.\frac{\partial M_{N(t)}(iu)}{\partial iu}\right|_{i0} = \left.\frac{\partial M_{\Lambda(t)}(e^{iu}-1)}{\partial iu}\right|_{i0}$$
$$= \left.\frac{\partial C(iu,t)}{\partial iu}\right|_{i0} M_\xi\left(e^{i0}-1\right) + C(i0,t)\left.\frac{\partial M_\xi(e^{iu}-1)}{\partial iu}\right|_{i0}$$

y para la varianza del proceso necesitamos el momento de segundo orden

$$E\left[N^2(t)\right] = \left.\frac{\partial^2 M_{N(t)}(iu)}{\partial iu^2}\right|_{i0} =$$
$$\left.\frac{\partial^2 C(iu,t)}{\partial iu^2}\right|_{i0} M_\xi\left(e^{i0}-1\right) + 2\left.\frac{\partial C(iu,t)}{\partial iu}\right|_{i0}\left.\frac{\partial M_\xi(e^{iu}-1)}{\partial iu}\right|_{i0} +$$
$$+ C(i0,t)\left.\frac{\partial^2 M_\xi(e^{iu}-1)}{\partial iu^2}\right|_{i0}$$

Puesto que $\xi \rightsquigarrow N(\mu,\sigma)$, es fácil observar que

$$M_\xi\left(e^{i0}-1\right) = 1; \quad \left.\frac{\partial M_\xi(e^{iu}-1)}{\partial iu}\right|_{i0} = \mu \quad \text{y} \quad \left.\frac{\partial^2 M_\xi(e^{iu}-1)}{\partial iu^2}\right|_{i0} = \mu^2 + \mu + \sigma^2$$

se comprueba también que

$$C(i0,t)=1; \qquad \left.\frac{\partial C(iu,t)}{\partial iu}\right|_{i0}=\frac{\sigma}{\sqrt{2\pi}}\frac{e^{-\frac{1}{2}\left(\frac{B-\mu}{\sigma}\right)^2}-e^{-\frac{1}{2}\left(\frac{A-\mu}{\sigma}\right)^2}}{P[A<\xi<B]} \quad \text{y}$$

$$\left.\frac{\partial^2 C(iu,t)}{\partial iu^2}\right|_{i0}=\frac{\sigma}{\sqrt{2\pi}}\frac{(A-\mu+1)e^{-\frac{1}{2}\left(\frac{A-\mu}{\sigma}\right)^2}-(B-\mu+1)e^{-\frac{1}{2}\left(\frac{B-\mu}{\sigma}\right)^2}}{P[A<\xi<B]}$$

Tenemos ya los cálculos necesarios para obtener la expresión de la esperanza y la varianza. Aunque hemos hecho los cálculos para la descomposición de la función característica del proceso como la función $C(iu,t)$ por la función característica de ξ en $e^{iu}-1$, resultaría igual haber tomado la segunda interpretación (ecuación 5) que hacíamos anteriormente. Queda, por tanto, que la esperanza y la varianza de un PPDE con media aleatoria Normal truncada son los siguientes:

(6) $$E[N(t)]=\frac{\sigma}{\sqrt{2\pi}}\frac{f(B)-f(A)}{P[A<\xi<B]}+\mu$$

y

(7) $$Var[N(t)]=$$
$$\frac{\sigma[(A-\mu+1)f(A)-(B-\mu+1)f(B)]}{\sqrt{2\pi}P[A<\xi<B]}-\left[\frac{\sigma[f(B)-f(A)]}{\sqrt{2\pi}P[A<\xi<B]}\right]^2+\sigma^2+\mu$$

donde ξ se distribuye como una $N(\mu,\sigma)$ y $f(x)=\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$. Recordamos de nuevo que A,B,μ y σ son todas ellas funciones del tiempo.

3. PREDICCIÓN DEL PPDE CON MEDIA NORMAL TRUNCADA

Una vez conocidos de forma teórica los estadísticos del PPDE, esta sección trata el problema de modelizar y predecir este tipo de procesos puntuales teniendo conocimiento de r trayectorias muestrales. Se propone un método de estimación de la distribución de $\Lambda(t,x(t))$ en un instante t_1, conocido el proceso en un intervalo de tiempo pasado $[t_0,T]$, $T<t_1$, a partir de lo cual se puede predecir la distribución del proceso puntual en ese instante futuro. Esto nos permitirá estudiar las características del proceso en ese mismo instante de tiempo.

El método de predicción de la distribución del PPDE en t_1 se basa en la idea de modelizar la evolución de $\Lambda(t,x(t))$ en el tiempo. En el primer paso del método, se necesitará conocer la estimación de la distribución Normal truncada a partir de una muestra. Por

tanto, estudiaremos la estimación de sus parámetros antes de adentrarnos en el método de predicción del PPDE.

El artículo de Mittal y Dahiya (1987) explica que los estimadores máximo verosímiles (EMV) para los parámetros de la distribución Normal truncada, $N_T(A,B,\mu,\sigma)$, pueden ser infinito con probabilidad positiva. Por esta razón, proponen otra forma de estimar esos parámetros, la estimación modal bayesiana, usando una densidad a priori para el parámetro $\frac{1}{\sigma^2}$. En la literatura bayesiana, es común usar una densidad chi-cuadrado cuando no hay información acerca del parámetro, así la densidad sería

$$f(\sigma^2) = f(\theta) = c(\nu)\,\theta^{-\frac{\nu-2}{2}} \exp\left(\frac{-1}{2\theta}\right), \theta > 0$$

donde $c(\nu) = 2^{-\frac{\nu}{2}} \left[\Gamma(\frac{\nu}{2})\right]^{-1}$ es la densidad a priori para μ. La función de verosimilitud modificada es ahora

$$L_m(\mu,\sigma/\Lambda_1,\ldots,\Lambda_r) = \frac{c_2\,\sigma^{2-\nu} \exp\left[-\left(\sum_{i=1}^{r}(\Lambda_i-\mu)+1\right)/2\sigma^2\right]}{\left[\int_A^B \exp\left(-(y-\mu)^2/2\sigma^2\right) dy\right]^r}$$

donde c_2 es una función de ν. En Mittal and Dahiya (1987) se propone $\nu = 4$ como valor óptimo para realizar la estimación.

La estimación modal bayesiana da como resultado los estimadores máximo verosímiles modificados (EMVM). En el artículo citado, se prueba que los EMVM existen con probabilidad uno y notablemente mejores que los EMV. Estos estimadores modificados se desarrollan en la siguiente proposición.

Proposición 1. *Sea Λ una variable aleatoria distribuida según una Normal truncada, $N_T(A,B,\mu,\sigma)$, con μ y σ desconocidos.*

Entonces, los EMVM de la distribución Normal truncada verifican el siguiente sistema de ecuaciones no lineales:

(8)

$$\overline{\Lambda} = \frac{\int_A^B y \exp[-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2]dy}{\int_A^B \exp[-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2]dy}$$

$$S^2 + \frac{1}{r} = \frac{2\sigma^2}{r} + \frac{\int_A^B y^2 \exp[-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2]dy}{\int_A^B \exp[-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2]dy} - \left(\frac{\int_A^B y \exp[-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2]dy}{\int_A^B \exp[-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2]dy}\right)^2$$

donde $\overline{\Lambda}$ y S son la media y la desviación típica muestrales, respectivamente, y r el tamaño muestral.

En esta proposición, se han dado estimadores para μ y σ pero no para A ni B. No existe en la literatura solución para la estimación de todos los parámetros de una distribución Normal truncada, siempre se suponen conocidos los límites de truncamiento. Nosotros proponemos escoger A y B como el mínimo y máximo de los valores de la muestra, de esta forma se hará en la aplicación. Se ha implementado la resolución del sistema de ecuaciones no lineales para la estimación de μ y σ (ecuación 8) usando el método de Newton-Raphson para su utilización en distintas aplicaciones como la presentada en este trabajo.

El método de predicción del PPDE en un instante de tiempo futuro que en este trabajo vamos a utilizar para la modelización y predicción del número de vehículos matriculados en países europeos puede ya exponerse de forma abreviada en los siguientes tres pasos.

1. Estimar los parámetros de la distribución Normal truncada perteneciente a la media del PPDE para cada $t \in [t_0, T]$, aplicando la implementación de la estimación de parámetros de la Normal truncada mencionada anteriormente.
2. Comprobar, para cada $t \in [t_0, T]$, que la media del proceso puntual se modeliza adecuadamente por la Normal truncada encontrada en el primer paso. Para ello ha sido también implementado el test de Kolmogorov-Smirnov para la bondad de ajuste de la distribución Normal truncada.
3. Una vez aceptado, que $\Lambda(t, x(t))$ es una particular Normal truncada para cada $t \in [t_0, T]$, tenemos entonces unos valores diferentes para A, B, μ and σ para cada t. Se estudia ahora cómo esos parámetros evolucionan en el tiempo estableciendo las funciones $A(t), B(t), \mu(t)$ y $\sigma(t)$. Con el conocimiento de estas funciones se puede extrapolar los parámetros de $N_T(A(t_1), B(t_1), \mu(t_1), \sigma(t_1))$, que es la distribución de $\Lambda(t_1, x(t_1))$. Esta predicción de $\Lambda(t_1, x(t_1))$ nos permite calcular la distribución del propio PPDE, así como sus características en ese instante.

Los tres pasos del método de predicción del PPDE han sido implementados en notebooks de *Mathematica 3.0*, el cual usa el método gaussiano adaptativo para calcular numéricamente las integrales que aparecen en la función de densidad de $\Lambda(t, x(t))$, el estadístico de recuento del PPDE, el sistema de ecuaciones no lineales, etc.

4. MODELIZACIÓN DEL NÚMERO DE MATRICULACIONES EN PAÍSES EUROPEOS

Como ya adelantamos anteriormente, vamos a modelizar el proceso puntual real número de matriculaciones en países europeos usando como modelo el PPDE con media aleatoria un Normal truncada. Poseemos los datos del número de matriculaciones en dieciséis países europeos desde 1979 a 1999, pero para que éstos sean comparables trabajaremos con ese número de matriculaciones por cada 10000 habitantes.

Tabla 1. N.° medio de matriculaciones acumuladas de 1979 a 1999 en España.

1979	1980	1981	1982	1983	1984	1985
13.9	26.8	37.9	49.7	61.8	73.1	85.5
1986	1987	1988	1989	1990	1991	1992
100.4	120.2	143.6	168.3	189.8	209.4	230.8
1993	1994	1995	1996	1997	1998	1999
247.2	267.3	285.7	306.2	329.3	356.4	388.1

Tabla 2. Normales truncadas estimadas para los 21 años y sus estadísticos experimentales de K-S.

Λ en el año	Distribución	D_{15}
1979	$N_T(8.3, 35.9, 22.7, 7.5)$	0.095
1980	$N_T(11.9, 70.1, 43.9, 15.3)$	0.081
1981	$N_T(16.6, 100.7, 65.3, 22.7)$	0.102
1982	$N_T(24.4, 131.3, 87.4, 29.6)$	0.079
1983	$N_T(30.7, 161.6, 111.0, 37.5)$	0.078
1984	$N_T(37.9, 192.7, 133.2, 44.3)$	0.073
1985	$N_T(47.0, 224.6, 158.6, 51.8)$	0.088
1986	$N_T(55.1, 258.8, 186.8, 61.2)$	0.082
1987	$N_T(61.2, 294.3, 214.3, 70.6)$	0.094
1988	$N_T(67.6, 514.8, 235.2, 93.7)$	0.135
1989	$N_T(76.2, 544.4, 260.9, 100.6)$	0.132
1990	$N_T(86.9, 578.4, 285.5, 107.9)$	0.131
1991	$N_T(101.3, 611.7, 307.8, 114.3)$	0.122
1992	$N_T(117.9, 645.1, 330.1, 120.7)$	0.130
1993	$N_T(130.4, 669.6, 349.3, 126.2)$	0.135
1994	$N_T(139.9, 693.7, 371.5, 132.3)$	0.137
1995	$N_T(150.7, 717.5, 393.9, 138.1)$	0.156
1996	$N_T(162.4, 742.8, 419.4, 144.3)$	0.154
1997	$N_T(175.6, 778.2, 446.0, 150.2)$	0.138
1998	$N_T(190.1, 812.5, 474.9, 156.8)$	0.137
1999	$N_T(211.1, 846.8, 504.2, 161.8)$	0.147

Fácilmente se calcula a partir de los datos las medias acumuladas por año para cada país, es decir $\Lambda(1979), \ldots, \Lambda(1999)$, con lo que preparamos los datos para realizar el primer paso del método de predicción. Los países estudiados han sido Alemania,

Austria, Bélgica, Checoslovaquia, Dinamarca, España, Finlandia, Francia, Grecia, Holanda, Irlanda, Italia, Luxemburgo, Noruega, Gran Bretaña y Suecia. Mostramos como ejemplo en la Tabla 1 las medias acumuladas para España desde 1979 a 1999.

Tomando los valores de las medias acumuladas de todos los países en un cierto año, t, tenemos una muestra de 16 datos de $\Lambda(t,x(t))$, que intentaremos modelizar mediante una Normal truncada que estimamos mediante la implementación de la resolución de las ecuaciones de los EMVM (ecuación 8). La Tabla 2 muestra las Normales truncadas estimadas, $N_T(A,B,\mu,\sigma)$, para cada uno de los veintiún años.

Una vez concluido el primer paso del método propuesto en la sección anterior, cada una de estas distribuciones ha sido contrastada mediante el test de bondad de ajuste de Kolmogorov-Smirnov según el segundo de los pasos. Todas ellas daban un buen ajuste a los datos muestrales por lo que se han aceptado como estimadoras de las distribuciones de cada $\Lambda(t,x(t))$, $t \in [t_0,T]$. El estadístico experimental del test de Kolmogorov-Smirnov (D_r) para cada Normal truncada puede también observarse en la Tabla 2. Puede comprobarse que todos los estadísticos experimentales son menores que el valor crítico, 0.338, para $\alpha = 0.05$.

Una vez estimadas todas las Normales truncadas, tenemos veintiún valores «observados» de las cuatro funciones del tiempo, $A(t),B(t),\mu(t)$ y $\sigma(t)$, parámetros de la distribución de la media del PPDE. Para completar el tercer paso del método se hace una regresión para cada una de esas funciones sobre los datos que nos han proporcionado las Normales truncadas encontraremos la expresión general de la media, $\Lambda(t,x(t)) \rightsquigarrow N_T(A(t),B(t),\mu(t),\sigma(t))$. Las regresiones más adecuadas para los cuatro parámetros han resultado ser:

$$
\begin{aligned}
A(t) &= (2.57+0.57z)^2 \\
B(t) &= z/\left(-z15\cdot 10^{-5}+0.02\right) \\
\mu(t) &= z/\left(-z65\cdot 10^{-6}+0.04\right) \\
\sigma(t) &= z/\left(-z69\cdot 10^{-5}+0.13\right)
\end{aligned}
\tag{9}
$$

donde $z = t - 1978$. Las gráficas de los ajustes pueden verse en las Figuras 1, 2, 3 y 4.

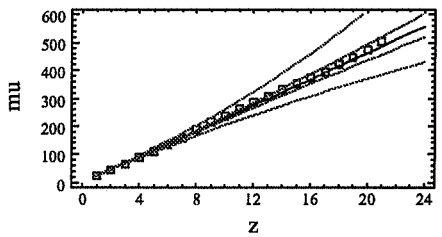


Figura 1. Regresión del parámetro A ($R^2 = 0.9972$).

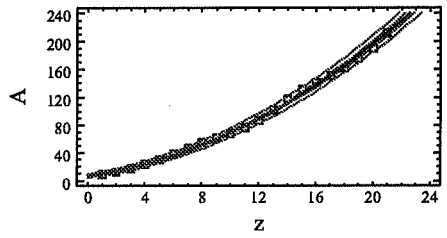


Figura 2. Regresión del parámetro B ($R^2 = 0.9948$).

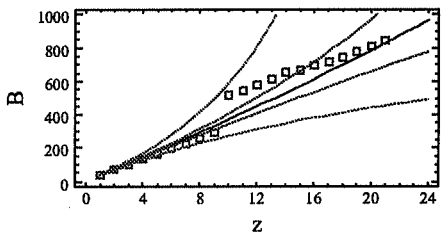


Figura 3. Regresión del parámetro μ ($R^2 = 0.9993$).

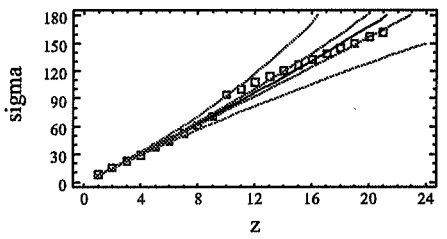


Figura 4. Regresión del parámetro σ ($R^2 = 0.9992$).

A partir del conocimiento de las funciones parámetro (9) de la Normal truncada general, se puede extrapolar a un instante futuro. Por ejemplo, se tiene que el PPDE en el año 2000 tiene una media que se distribuye mediante la Normal truncada, $N_T(229.7, 875.4, 510.4, 187.3)$. Teniendo en cuenta las ecuaciones (13) y (14) se obtiene también que la

esperanza del PPDE en el año 2000 y su varianza son

$$E[N(2000)] = 524.89 \text{ y } Var[N(2000)] = 23376.7.$$

Mediante la desigualdad de Chebyschev podemos decir que la probabilidad de que el número de matriculaciones en un país Europeo esté en el intervalo $[219.102, 830.678]$ es menor o igual que 0.75. Puesto que el número medio acumulado de matriculaciones hasta 1999 fue de 481.36 y la $E[N(2000)] = 524.89$, el número de matriculaciones en el año 2000 puede estimarse en aproximadamente 43.53.

Utilizando los datos disponibles para el año 2000, hemos calculado el valor muestral del número medio de matriculaciones hasta el 2000, resultando ser 540.92. Dado que el valor de la desviación típica que obteníamos mediante el método de predicción para el PPDE en el 2000 era 152.89, consideramos que la estimación obtenida de la media (524.89) es muy precisa. Debido a esto y al hecho de que las regresiones (ecuación 9) son tan buenas y suaves podemos extender la predicción al año 2001. La distribución de $\Lambda(2001)$ es la Normal truncada $N_T(247.4, 920.7, 534.4, 197.04)$, y a partir de ella y la ecuaciones (13) y (14) se obtiene que

$$E[N(2001)] = 551.84 \text{ y } Var[N(2001)] = 25495.1.$$

AGRADECIMIENTOS

Los autores queremos expresar nuestra gratitud a los dos referees que han revisado nuestro trabajo por sus interesantes sugerencias.

Este trabajo ha sido realizado dentro del Proyecto BFM2000-1466, financiado por la Dirección General de Investigación del Ministerio de Ciencia y Tecnología.

REFERENCIAS

Cox, D. R. (1955). «Some Statistical Methods Connected with Series of Events». *J. Royal Statistical Society B.* 17, 129-164.

Daley, D. J. and Vre-Jones, D. (1988). *An Introduction to Point Processes*, Springer-Verlag, New York.

García Ferrer, A. and Queralt, R. A. (1998). «Can univariate models forecast turning points in seasonal economic time series?». *International Journal of Forecasting*, 14, 433-446.

Mittal, M. M. y Dahiya, R. C. (1987). «Estimating the parameters of a Doubly Truncated Normal Distribution ». *Communications in Statistics. Simula*, 16, 141-159.

Rodríguez Bouzas, P., Aguilera, A. M. y Valderrama, M. J. (2001). «Forecasting a Class of Doubly Stochastic Poisson Processes». *Statistical Papers. En prensa.*

Snyder, D. L. and Miller, M. I. (1991). *Random Point Processes in Time and Space*, 2nd edition. Springer Verlag, New York.

Valderrama, M. J., Jiménez, F., Gutiérrez, R. y Martínez-Almécija, A. (1995). «Estimation and Filtering on a Doubly Stochastic Poisson Process». *Aplied Stochastic Model and Data Analysis*, 11, 13-24.

ENGLISH SUMMARY

A POISSONIAN MODEL TO FORECAST VEHICULES MATRICULATIONS IN EUROPEAN COUNTRIES

M. J. VALDERRAMA
A. M. AGUILERA
P. R. BOUZAS
Universidad de Granada*

In this paper, we present a methodology to model the number of domestic car registrations in European countries by a doubly stochastic Poisson process which mean is a truncated Normal variable. On the basis of previous work about this statistical model, we extend the study of its moments. Also, we forecast the mentioned real process for 2000 and 2001.

Keywords: Doubly stochastic Poisson process, truncated Gaussian distribution, car registrations

AMS Classification (MSC 2000): 60G55, 60K30

*Departamento de Estadística e I.O., Facultad de Farmacia, Campus de Cartuja, Universidad de Granada, 18071 Granada.

–Received July 2001.
–Accepted November 2001.

The number of domestic passenger car registrations is one of the most representative indicators of a country economic activity. This paper presents an original methodology to model this indicator in european countries. The statistical model is a doubly stochastic Poisson process (DSPP) that is a Poisson process in which the parameter is considered to be random. We study the particular case where its mean is a truncated Normal distribution in each instant of time. We also show a forecasting method for this process.

We will note the DSPP as $\{N(t);t \geq t_0\}$, where $N(t)$ is the number of ocurrences from t_0 to t. $\{N(t);t \geq t_0\}$ is called a DSPP with intensity $\{\lambda(t,x(t));t \geq t_0\}$ where $\{x(t);t \geq t_0\}$ is an external process iff for any sample path of $x(\cdot)$, $N(\cdot)$ is a Poisson process with intensity $\lambda(\cdot)$.

Using the conditioning method, it is obtained the probability mass function:

$$P[N(t)=n] = E\{P[N(t)=n/x(\sigma) : t_0 \leqslant \sigma < t]\} = G_{\Lambda}^{n)}(-1) \tag{10}$$

for $n = 0,1,2,\ldots$ where $\Lambda(t,x(t))$ is the parametric function of the DSPP, it is $\Lambda(t,x(t)) = \int_{t_0}^{t} \lambda(\sigma,x(\sigma))\,d\sigma$ and $G_{\Lambda}^{n)}(s)$ is the n$-th$ derivative of the moment generating function of $\Lambda(\cdot)$. Using again the conditioning method, the characteristic function of $N(\cdot)$ results

$$M_{N(t)}(iu) = E\left\{\exp\left[\left(e^{iu}-1\right)\Lambda(t,x(t))\right]\right\} = M_{\Lambda}(e^{iu}-1) \tag{11}$$

where $M_{\Lambda}(iu)$ is the characteristic function of $\Lambda(\cdot)$. The evaluation of $M_{N(t)}(iu)$ is habitually difficult if not imposible.

Studying different examples of counting processes, it usually seems that their means could be modeled by a Gaussian model but as $\Lambda(\cdot)$ has to be a non-decreasing function equal or greater than zero, we have considered a truncated Normal variable for each instant of time. We will note that $\Lambda(\cdot)$ is distributed as a truncated Normal like $\Lambda(t,x(t)) \rightsquigarrow N_T(A(t),B(t),\mu(t),\sigma(t))$. In this case, equation (10) becomes

$$P[N(t)=n] = \frac{1}{n!\sigma\sqrt{2\pi}}\left(P\left[\frac{A-\mu}{\sigma} < Z < \frac{B-\mu}{\sigma}\right]\right)^{-1} \int_{A}^{B} \Lambda^n e^{-\Lambda} e^{-\frac{1}{2}\left(\frac{\Lambda-\mu}{\sigma}\right)^2} d\Lambda$$

for $n = 0,1,2,\ldots$ and A,B,μ and σ are the parameters of $\Lambda(\cdot)$, all of them functions of time with $0 \leq A < \Lambda < B$, $\mu \in \mathbb{R}$ and $\sigma \geq 0$.

It can also be shown that equation (11) has the form

$$M_{N(t)}(iu) = C(iu,t)\,M_{\xi}\left(e^{iu}-1\right) \text{ with} \tag{12}$$

$$C(iu,t) = \frac{P\left[A-(e^{iu}-1)\sigma^2 < \xi < B-(e^{iu}-1)\sigma^2\right]}{P[A<\xi<B]}$$

where $\xi \rightsquigarrow N(\mu,\sigma)$, μ and σ functions of time. This interpretation of $M_{N(t)}(iu)$ allows us to calculate the mean and variance of the process. Due to (12) and after some manipulations, we have that

$$E[N(t)] = E[\Lambda(t)] = \frac{\sigma}{\sqrt{2\pi}} \frac{f(B) - f(A)}{P[A < \xi < B]} + \mu \tag{13}$$

$$Var[N(t)] = \frac{\sigma[(A-\mu+1) f(A) - (B-\mu+1) f(B)]}{\sqrt{2\pi} P[A < \xi < B]} -$$

$$\left[\frac{\sigma[f(B) - f(A)]}{\sqrt{2\pi} P[A < \xi < B]}\right]^2 + \sigma^2 + \mu \tag{14}$$

where again $\xi \rightsquigarrow N(\mu,\sigma)$ and $f(x) = \exp[-\frac{1}{2}(\frac{x-\mu}{\sigma})^2]$.

In this paper, we propose a method for estimating $N(\cdot)$ in t_1, known r sample paths of the counting process in $[t_0,T]$, $T < t_1$.

The forecasting method can be exposed now in these three steps:

1. Estimate the truncated Normal distributions of the means of the DSPP for each $t \in [t_0,T]$ using the modified maximum likelihood estimators.
2. Prove, for each $t \in [t_0,T]$, the truncated Normal estimated in Step 1 fits the data by a Kolmogorov-Smirnov test.
3. Once accepted all the truncated Normals in Step 2, it should be studied the evolution in time of A, B,μ and σ, stablishing the functions $A(t)$, $B(t)$, $\mu(t)$ and $\sigma(t)$. These functions are used to forecast $\Lambda(t_1,x(t_1))$ and so, $N(t_1)$.

The model studied has been applied to the number of domestic passenger car registrations per 10000 inhabitants in sixteen european countries from 1979 to 1999. The countries are Germany, Austria, Belgium, Checoslovaquia, Denmark, Spain, Finland, France, Greece, Holland, Ireland, Italy, Luxemburg, France, Great Britain and Sweede.

Investigació Operativa

QÜESTIIÓ, vol. 26, 1-2, p. 141-171, 2002

ON SUPERLINEAR MULTIPLIER UPDATE METHODS FOR PARTIAL AUGMENTED LAGRANGIAN TECHNIQUES*

E. MIJANGOS*

The minimization of a nonlinear function with linear and nonlinear constraints and simple bounds can be performed by minimizing an augmented Lagrangian function, including only the nonlinear constraints. This procedure is particularly interesting in case that the linear constraints are flow conservation equations, as there exist efficient techniques to solve nonlinear network problems. It is then necessary to estimate their multipliers, and variable reduction techniques can be used to carry out the successive minimizations. This work analyzes the possibility of estimating these multipliers using Newton-like methods. Several procedures are put forward which combine first and second-order estimation, and are compared with each other and with the Hestenes-Powell multiplier estimation by means of computational tests.

Keywords: Partial Augmented Lagrangian, Variable Reduction, Lagrange Multipliers, Superlinear-order Estimation, Nonlinear Network

AMS Classification (MSC 2000): 90C30, 90C35

* Department of Applied Mathematics and Statistics and O.R., Zientzi Fakultatea (UPV/EHU), P.O. Box 644, 48080-Bilbao, Spain. Fax 34 944648500, e-mail: mepmifee@lg.ehu.es.

* Partially supported by CICYT grant TAP1999-1075-C0201.

– Received Setember 2000.
– Accepted November 2001.

1. INTRODUCTION AND MOTIVATION

Consider the linearly and nonlinearly constrained problem (**EGP**):

$$
\begin{array}{lll}
(1) & \text{minimize} & f(x) \\
(2) & \text{subject to:} & Ax = b \\
(3) & & c(x) = 0 \\
(4) & & l \leq x \leq u,
\end{array}
$$

where

1. $f : \mathbb{R}^n \longrightarrow \mathbb{R}$. $f(x)$ is nonlinear and twice continuously differentiable on the feasible set defined by constraints (2) and (4).
2. A is an $m \times n$ matrix and b an m-vector.
3. $c : \mathbb{R}^n \longrightarrow \mathbb{R}^r$, is such that $c = [c_1, \cdots, c_r]^t$, $c_i(x)$ being linear or nonlinear and twice continuously differentiable on the feasible set defined by constraints (2) and (4) $\forall i = 1, \cdots, r$.

Throughout this work the gradient of f at x is defined as the column vector $\nabla f(x)$, and matrix $\nabla c(x) = [\nabla c_1(x), \cdots, \nabla c_r(x)]$ is the transpose of the *Jacobian* of c at x, though here, for convenience, it will simply be called Jacobian.

The idea is to solve **EGP** by solving a series of approximate problems whose solutions «converge» on the solution of the original problem in a well-defined sense.

In practice, it is only necessary to solve a finite number of approximate problems to arrive at an acceptable approximate solution to the original problem.

Real problems with this same structure exist and have a high dimensionality (e.g., when A is a node-arc incidence matrix). They often need to be solved and it is important to find the procedure that will solve them with the highest efficiency. The short-term hydroelectric power generation optimization problem [5, 32] and the short-term hydrothermal coordination of electric power generation [16] are examples of this type.

One possible way to address problem **EGP** would be to use of Projected Lagrangian methods [12]. This would lead to solution of a series of subproblems of the type:

$$
\begin{array}{ll}
\text{minimize} & \Phi^k(x) \\
\text{subject to:} & Ax = b \\
& l \leq x \leq u \\
& \nabla c(x^k)^t x = -c(x^k) + \nabla c(x^k)^t x_k
\end{array}
$$

where $\Phi^k(x)$ is either a quadratic approximation to the partial Lagrangian function, or a nonlinear function based on the partial augmented Lagrangian, as the well-known MINOS 5.5 package [27] does. When A is a node-arc incidence matrix this subproblem is a nonlinear network flow problem with linear side constraints, which could be solved by specialized techniques of primal partitioning for network flow problems [19]. This possibility has already been explored [15].

In this work **EGP** is solved using partial augmented Lagrangian techniques [1, 2, 3, 6, 30], as is done in [4, 7, 21, 22, 23, 24], where only the general constraints (3) are included in the Lagrangian.

There is a class of problems for which these techniques have potential interest, in particular those that incorporate *flow conservation equations* as linear constraints (2), as there are efficient techniques for solving nonlinear network problems, see [9, 31]. In [21, 22, 23, 24] the constraints (3) are the *side constraints* of a nonlinear network flow problem. The application of these techniques consists of two fundamental steps. The first is the solution of problem (**EGS**):

$$\begin{array}{lll} & \underset{x}{\text{minimize}} & L_\rho(x,\mu) \\ (5) & \text{subject to} & Ax = b \\ & & l \le x \le u, \end{array}$$

where $\rho \in \mathbb{R}$, such that $\rho > 0$, and $\mu \in \mathbb{R}^r$ are fixed,

$$L_\rho(x,\mu) = f(x) + \mu' c(x) + \frac{1}{2}\rho c(x)' c(x).$$

Should the solution $\tilde{x}$ obtained by solving **EGS** be infeasible with respect to (3), the second step, which is the updating of the estimation μ of the Lagrange multipliers of constraints (3), is carried out —also updating, if necessary, the penalty coefficient ρ— followed by a return to the first step. Should $\tilde{x}$ be feasible (or the violation of constraints (3) sufficiently small) the procedure ends.

It is of paramount importance that the estimation of multipliers μ be as accurate as possible, otherwise the convergence of the algorithm can be severely handicapped, as shown in [2].

In practice there are two first-order procedures to estimate μ. On the one hand the method put forward by Hestenes [18] and Powell [29]

$$\tilde{\mu} = \mu + \rho c(\tilde{x}),$$

and on the other hand μ_L obtained through the classical solution to the system of Kuhn-Tucker necessary conditions

$$(6) \qquad \nabla f(\tilde{x}) + \nabla c(\tilde{x})\mu + A'\pi + \lambda = 0$$

by least squares, as suggested in [12]. A study of the viability of using these multiplier estimation techniques together with the solution of problem **EGS** (5) by the Murtagh and Saunders's active set technique [26] was carried out in [21], using the method suggested by Hestenes & Powell, and in [24], by solving the Kuhn-Tucker necessary conditions using least squares.

An interesting alternative to these methods is the second-order multiplier estimation for inexact minimization suggested by Tapia in [30], and whose convergence analysis is due to Bertsekas in [2].

In this paper the applicability of the Tapia's second-order multiplier estimate is analyzed when subproblem (5) is solved using variable reduction techniques. As a consequence, under variable reduction techniques an algorithm that is a combination of superlinear and first-order multiplier methods is designed to improve the results obtained using first-order estimates over large-scale nonlinear network problems with side constraints [21, 24]. In the same algorithm, it is shown how to compute the superlinear-order multiplier estimate efficiently by means of suitable matrix computations. Finally, these techniques are extended to problem (**IGP**):

(7)	minimize	$f(x)$
(8)	subject to	$Ax = b$
(9)		$\alpha \leq c(x) \leq \beta$
(10)		$l \leq x \leq u,$

using the results of the work by Mijangos and Nabona in [21], which is based on the theory developed by Bertsekas (1982) in [2] for extending the multiplier methods obtained for equality constrained problems to one-sided inequality constrained problems, without need of using slack variables. Numerical tests over nonlinear network problems with side constraints are carried out, several algorithmic alternatives being compared with each other and with the Hestenes-Powell multiplier estimation.

The paper is organised as follows. Section 2 analyzes the applicability of these techniques to solve problem **EGP** when subproblem (5) is solved by the Murtagh and Saunders procedure [26]. Also, in Section 2, an algorithm combining superlinear and first-order multiplier estimate techniques is put forward and various practical issues are analyzed. Section 3 extends these techniques to problem **IGP**. Section 4 offers some details about the implementation. Section 5 presents computational tests. Finally, Section 6 offers the conclusions.

2. APPLICABILITY OF THE NEWTON-TYPE METHODS

There follows a new analysis of the applicability of the inexact Newton-like methods in combination with variable reduction techniques.

Let us consider any update formula U of μ in the solution of **EGP** by means of multiplier methods, where the subproblem to solve is (5). The first-order optimality conditions to be fulfilled by the optimizer $(x^*,\mu^*,\pi^*,\lambda^*)$ at the end of this sequential procedure, for all $\rho > 0$ large enough, give rise to the following system

$$\nabla_x L_\rho(x,\mu) + A^t\pi + \lambda = 0 \tag{11}$$

$$c(x) = 0 \tag{12}$$

$$Ax = b, \tag{13}$$

where

$$\nabla_x L_\rho(x,\mu) = \nabla f(x) + \nabla c(x)[\mu + \rho c(x)].$$

Throughout this work we assume that $(x^*,\mu^*,\pi^*,\lambda^*)$ is an optimizer at the end of the sequential procedure satisfying the strict complementarity condition.

Likewise, we assume that given a pair (μ,ρ), ρ is greater than a certain threshold (see Lemma 1.25 in [2]) in order to make sure the full rank of the reduced Hessian of the augmented Lagrangian in the subproblem **EGS**, see (5). Since this work is concerned with the solution of problems where the aforementioned reduced Hessian is dense in general, the highest dimension of this matrix is limited to 500×500.

Let be $\tilde{x} = x(\mu,\rho)$ an optimizer of this subproblem. It is obvious that the only information we have about the active variables at x^* and the bound where they are fixed is given by the solution of subproblem **EGS**.

Newton's method for solving the system consisting of (11–13) and the equations of the active variables at $\tilde{x}$ involves solving the linear system

(14)

n		r	m	$\hat{t}$		
$\nabla^2_{xx} L_\rho(\tilde{x},\mu)$		$\nabla c(\tilde{x})$	A^t	0	Δx	$-\nabla_x L_\rho(\tilde{x},\mu)$
				$\mathbb{I}$		$-A^t\pi - \lambda$
$\nabla c(\tilde{x})^t$					$\Delta\mu$	$-c(\tilde{x})$
A			0		$\Delta\pi$	0
0	$\mathbf{1}$				$\Delta\lambda$	0

where $\hat{t}$ stands for the number of variables fixed at their bounds in the solution of subproblem (5).

Let $\nabla^2\mathcal{L}$ denote the matrix of coefficients of (14), which should be nonsingular. Let us assume that matrix $\nabla^2_{xx} L_\rho(\tilde{x},\mu)$ is nonsingular and A has full row rank; then, the necessary and sufficient condition for matrix $\nabla^2\mathcal{L}$ to be nonsingular is that

(15)
$$\widehat{A}(\tilde{x}) = \begin{array}{|c|c|c|l} \hline \nabla c_B(\tilde{x})^t & \nabla c_S(\tilde{x})^t & \nabla c_N(\tilde{x})^t & \}r \\ \hline B_A & S_A & N_A & \}m \\ \hline \multicolumn{2}{|c|}{0} & \mathbf{1} & \}\widehat{t} \\ \hline \end{array}$$

has full row rank, where the second row section contains the matrix A partitioned into three submatrices, B_A and S_A corresponding to the last basic and superbasic submatrices after using the Murtagh and Saunders method to solve subproblem **EGS** (5), N_A being the columns of A corresponding to the $\widehat{t}$ active variables in $\tilde{x}$. The partition of $\nabla c(\tilde{x})^t$ is the same as that of A. The variables associated with columns N_A are precisely those variables whose bound is predicted to be active at the optimizer. However, it could happen that the submatrix

$$BS = \begin{array}{|c|c|} \hline \nabla c_B(\tilde{x})^t & \nabla c_S(\tilde{x})^t \\ \hline B_A & S_A \\ \hline \end{array}$$

does not have full row rank, which could be due to one of two reasons:

- either $\nabla c(\tilde{x})^t$ does not have full row rank in spite of $\nabla c(x^*)^t$ having it;
- or $\nabla c(\tilde{x})^t$ has full row rank, but $[\nabla c_B(\tilde{x})^t \quad \nabla c_S(\tilde{x})^t]$ does not have it, because the prediction about the variables whose bound will be active at the optimizer is not totally correct.

Both cases can occur if $\tilde{x}$ is not sufficiently close to x^*.

In the second case we can know (or hope) that $\nabla c(\tilde{x})^t$ has full row rank either because $\tilde{x}$ is already close enough to the optimizer, or due to the structure of the problem, or because the rank is checked explicitly. Then, it may be interesting to enlarge BS by adding columns of the submatrix

(16)
$$N = \begin{array}{|c|} \hline \nabla c_N(\tilde{x})^t \\ \hline N_A \\ \hline \end{array}$$

until a matrix is obtained with full row rank. However, this would mean excluding these variables from being bound active variables and forcing their λ^* multiplier to be zero, against the prediction in $\tilde{x}$. If these variables were not suitably chosen (assuming that «suitable» ones exist), it could happen that the Δx displacement associated with them would not be feasible, and thus all the computational effort expended in solving system (14) would have been lost. Therefore, when $\tilde{x}$ is sufficiently close to x^*, and in order to complete the row rank of BS (if necessary), we suggest enlarging BS by adding

the columns of N whose associated active variables have a multiplier estimate zero or almost zero, see Appendix A in [24].

2.1. Multiplier update in problem EGP

In practice, one does not need to solve system (14). Let $\overline{Z}$ be the following matrix:

$$\overline{Z} = \begin{bmatrix} Z_A^t & 0 \\ 0 & \mathbf{1} \end{bmatrix}, \tag{17}$$

where the unit matrix has dimension $r+m+\hat{t}$ and Z_A is the variable reduction matrix:

$$Z_A = \begin{bmatrix} -B_A^{-1}S_A \\ \mathbb{1} \\ 0 \end{bmatrix}, \tag{18}$$

such that S_A is enlarged with columns of N_A if it is necessary for BS to have full row rank.

Premultiplying both sides of (14) by $\overline{Z}$ and, since there is a vector Δx_S and only one such that

$$\Delta x = Z_A \Delta x_S, \tag{19}$$

using this last expression we are led from system (14) to the *reduced Newton system*:

(20)

	s	r			
	$Z_A^t \nabla_{xx}^2 L_\rho(\tilde{x},\mu) Z_A$	$Z_A^t \nabla c(\tilde{x})$	Δx_S	$=$	$-Z_A^t \nabla_x L_\rho(\tilde{x},\mu)$
	$\nabla c(\tilde{x})^t Z_A$	0	$\Delta\mu$		$-c(\tilde{x})$

Since $\nabla^2 \mathcal{L}$ is nonsingular and Z_A has full column rank, the coefficient matrix of system (20) is also nonsingular, hence this system has a single solution $(\Delta x_S, \Delta\mu)$.

By first calculating $\Delta\mu$ we find that for problem **EGP** the U_{INW} update is given by

$$\overline{\mu} = \mu + [J^t H^{-1} J]^{-1} [c(\tilde{x}) - J^t H^{-1} (Z_A^t \nabla_x L_\rho(\tilde{x},\mu))] \tag{21}$$

where $J = Z_A^t \nabla c(\tilde{x})$ is the *reduced Jacobian* in $\tilde{x}$ and H is either the exact reduced Hessian of the augmented Lagrangian —i.e., $Z_A^t \nabla_{xx}^2 L_\rho(\tilde{x},\mu) Z_A$— or an approximation of it (e.g., quasi-Newton or discrete Newton).

We should not forget that J must have full column rank so that the system matrix of (20) will be nonsingular and, in consequence, will have an inverse.

2.2. Solution of the reduced Newton system

The procedure that we are going to describe makes it possible to *switch* from the first-order update to a superlinear-order update «when the circumstances allow and suggest it». The final aim of this combination is to increase its convergence rate as far as possible, at the same time keeping the global convergence of the multiplier method, see [2].

The rules considered in the previous section are complemented with others added in order to check the superlinear convergence at the end of the algorithm. Moreover, three possibilities are considered in relation to the Hessian H (here reduced Hessian) of the augmented Lagrangian, corresponding to the following cases:

- it is the exact Hessian at x^k,
- it is its approximation by finite differences, or
- it is a quasi-Newton approximation (e.g., BFGS).

Let $x^k = x(\mu^k, \rho^k)$ be the solution of the subproblem **EGS** for pair (μ^k, ρ^k), see (5).

As a consequence of §2.1, if for $x = x^k$ the submatrices of the coefficient matrix and the right-hand side of system (20) are denoted as follows:

$$\begin{aligned} H_k &= Z_A^t \nabla_{xx}^2 L_{\rho^k}(x^k,\mu^k) Z_A \\ J_k &= Z_A^t \nabla c(x^k) \\ h_k &= Z_A^t \nabla_x L_{\rho^k}(x^k,\mu^k), \end{aligned}$$

the solution of (20) at x^k is given by the initial calculation of

$$\Delta\mu = (J_k^t H_k^{-1} J_k)^{-1} (c(x^k) - J_k^t H_k^{-1} h_k),$$

μ^{k+1} being calculated by means of

$$\mu^{k+1} = \mu^k + \Delta\mu,$$

then $\bar{h}_k = Z_A^t \nabla_x L_{\rho^k}(x^k, \mu^{k+1})$ is obtained, and finally

$$\Delta x_S = -H_k^{-1}\bar{h}_k, \tag{22}$$

from which the remaining components of Δx are obtained by using expression (19). This Δx may be used in the first iteration of the solution of the new **EGS** subproblem (see (5)) as search direction d, provided that it is feasible with regard to the simple bounds of subproblem **EGS**. However, the reason for this section is to compute the superlinear-order update of μ, when this is possible, and using it if it is reliable enough. Therefore, assuming that H_k is either the exact projected Hessian, or a good enough approximation of it, and that we have a Cholesky factorization $H_k = R_k^t R_k$, the following algorithm is designed. Let $D_k = J_k^t H_k^{-1} J_k$ and $U_O(x, \mu, \rho) = \mu + \rho c(x)$.

Algorithm 1

Step 1. Compute the first-order multiplier estimation U_O and obtain a projected Jacobian J_k having full column rank. Set

$$\tilde{\mu} = U_O(x^k, \mu^k, \rho^k).$$

If T_{nw0} (*activation rule*) (see §2.3) a projected Jacobian J_k having full column rank is obtained, if it exists, by means of the method suggested in Appendix A of [24]. If either T_{nw0} is not true, or J_k does not exist in spite of this rule being true, the algorithm finishes with the update U_O by default.

Step 2. Compute Cholesky factorization $D_k = \bar{R}_k^t \bar{R}_k$.

(i) The matrix M_k is computed by calculating its columns $M_k^{(i)}$, for $i = 1, \ldots, r$, which is done by successively solving system

$$R_k^t M_k^{(i)} = J_k^{(i)},$$

where $J_k^{(i)}$ denotes the i-th column of J_k.

(ii) The QR factorization of M_k is carried out by obtaining an upper triangular matrix $\bar{R}_k$ such that

$$\begin{bmatrix} \bar{R}_k \\ 0 \end{bmatrix} = Q_k M_k,$$

Q_k being a suitable orthogonal matrix.

Step 3. Compute $\bar{c}_k = c(x^k) - J_k^t H_k^{-1} h_k$.

(i) Solve $R_k^t R_k p = h_k$.

(ii) Compute $\bar{c}_k = c(x^k) - J_k^t p$.

Step 4. Solve $D_k \Delta\mu = \bar{c}_k$. By solving $\bar{R}_k^t \bar{R}_k \Delta\mu = \bar{c}_k$.

Step 5. Compute the superlinear-order update of μ. Set

$$\mu^{k+1} = U_{INW}(x^k, \mu^k, \rho^k),$$

where

$$U_{INW}(x^k, \mu^k, \rho^k) = \mu^k + \Delta\mu.$$

Step 6. Check the validation rule. If

$$T_{nw1} := \max_{i=1,\ldots,r} \frac{|\mu_i^{k+1} - \widetilde{\mu}_i|}{|\widetilde{\mu}_i|} \leq \tau_{nw1},$$

(see §2.3) the algorithm finishes with update U_{INW}, achieved in step 5. Otherwise, it ends with $\mu^{k+1} = \widetilde{\mu}$.

2.3. Activation and validation rules

One issue that is interesting to consider is whether it is worth attempting to compute the superlinear-order update when the current point is not close enough to an optimizer. The main reason for not doing so is the high computational cost of obtaining a projected Jacobian J_k having full column rank in the first step of Algorithm 1 (if possible). We run the risk of finding that the current projected Jacobian J_k does not have full column rank, in which case we have wasted all the effort taken to compute it. One way of controlling the closeness to (x^*, μ^*) is given by the following test:

$$T_{nw0} := T_x \text{ and } T_\mu,$$

where, for a previously fixed tolerance τ_x,

$$T_x := \overline{\tau} \leq \tau_x$$

is a test carried out at the end of the solution of subproblem **EGS** (5) in order to know the closeness of the current point to x^*, $\overline{\tau}$ being a tolerance used in the definition of the optimality tolerance for subproblem **EGS** (for further details see the definition of $\overline{\tau}$ in Step 1 of the Algorithm 2 for solving problem **IGP** in §3.1); and

$$T_\mu := \|\mu^k - \mu^{k-1}\| < (1 + \|\mu^k\|)\tau_\mu,$$

where τ_μ is a previously fixed tolerance.

In the event of T_{nw0} being false the superlinear-order update is not performed and $\mu^{k+1} = \widetilde{\mu}$, which is the first-order update U_O.

The test

$$T_{nw1} := \max_{i=1,\ldots,r} \frac{|\mu_i^{k+1} - \widetilde{\mu}_i|}{|\widetilde{\mu}_i|} \leq \tau_{nw1},$$

is based on the fact that the second-order update is rarely correct if it is very far from the first-order update, see [11]. Therefore, there is some level of agreement between the two estimations.

2.4. A more numerically-stable alternative

In this alternative the second-order update is performed in two steps, as suggested by Bertsekas in proposition 2.8 of [2].

The first step consists of

- computing the first-order estimation of μ^* given by

$$\widetilde{\mu} = U_O(x^k, \mu^k, \rho^k),$$

- taking $\widehat{H}_k = Z_A^t \nabla^2_{xx} L_0(x^k, \widetilde{\mu}) Z_A$, where $L_0(x^k, \widetilde{\mu})$ means the Lagrangian function (i.e., the augmented Lagrangian function for $\rho = 0$) with $\mu = \widetilde{\mu}$,
- and computing a modified Cholesky factorization of $\widehat{H}_k$ (see [12]), obtaining $\widehat{H}_k = \widehat{R}_k^t \widehat{R}_k$.

Note that in Algorithm 1 we replace R_k with $\widehat{R}_k$ and that

$$H_k = \widehat{H}_k + \rho^k J_k J_k^t.$$

In the second step Algorithm 1 is applied replacing the Step 5 with

$$U_{INW}(x^k, \mu^k, \rho^k) = \widetilde{\mu} + \Delta\mu. \tag{23}$$

It is easy to see that the solution obtained with the two procedures is the same, see §2.3.2 in [2].

The fundamental difference between both methods lies in that instead of the reduced Hessian (or an approximation of it) of the augmented Lagrangian function $L_{\rho^k}(x^k, \mu^k)$, we require the reduced Hessian (or an approximation of it) of the Lagrangian function $L_0(x^k, \widetilde{\mu})$, where $\widetilde{\mu}$ also appears instead of μ^k. This precludes using the BFGS approximation of $\nabla^2_{xx} L_{\rho^k}(x^k, \mu^k)$ in each k-th major iteration. However, if we have the analytic reduced Hessian of the Lagrangian function for pair $(x^k, \widetilde{\mu})$ (or an approximation of it, e.g., by finite differences) then the procedure considered in this section can be applied.

Since $\widetilde{\mu}$ is a better estimate than μ^k —using known results for Newton's method— we expect that (23) will yield a vector μ^{k+1} which is closer to μ^* than μ^k.

Moreover an important advantage of this procedure is that although the condition number of

$$J_k^t \widehat{H}_k^{-1} J_k$$

is present in the error bound of the system to solve, here it is multiplied by a $\|\Delta\mu\|$, which is generally small compared with that obtained directly from step 5 without modifications, and this reduces the effect of the condition number with regard to the direct application of the Algorithm 1 —see §2.7.2 in [13]—, as can be observed in the computational tests later.

2.5. Issues concerning the quasi-Newton multiplier update

A quasi-Newton method (here BFGS) is normally used when the number of superbasic variables is no greater than a number previously given $\bar{s}$, which is chosen so that the computational effort of computing the search direction in this way will be smaller than carrying out this calculation using another efficient method (e.g., the Truncated Newton [8]). Here, $\bar{s} = 500$ by default. If $s > \bar{s}$ it is clear that we will not have this factored approximation, hence we will necessarily either have to compute the reduced Hessian of the augmented Lagrangian directly, or approximate it by means of finite differences according to the procedure explained below. In large-scale problems with many degrees of freedom the possibility of $s > \bar{s}$ is also an issue to take into account at the moment of implementing the quasi-Newton techniques for the superlinear-order update, though it is not a crucial question for the aims of this work. Nevertheless, we must not forget that there are other algorithmic choices that help to solve this problem as for example: the L-BFGS method of Nocedal [25, 28] and the partitioned quasi-Newton method of Griewank and Toint [14].

When using a BFGS approximation in the solution $\widetilde{x}$ of subproblem **EGS**, see (5), H will be positive definite and one can directly update its Cholesky factorization. However, the matrix J that we have in $\widetilde{x}$ may not have full column rank. Therefore, it will be necessary to take columns of N (defined by (16)) to complete its rank, and the reduction matrix (18) will thus have to be modified as a consequence of enlarging S_A with columns of N_A, which makes the current quasi-Newton update H useless as an approximation of $Z_A^t \nabla^2_{xx} L_\rho(\widetilde{x},\mu) Z_A$ in system (20). Hence the quasi-Newton approximation can only be directly exploited when the reduced Jacobian J at $\widetilde{x}$ does not require that columns of N be taken in order to complete the column rank of BS and, in consequence, the rank of J. Otherwise, H could be updated by adding nonbasic variables associated with the columns of N to the superbasic set one by one and updating for each new superbasic variable the Cholesky factorization of H (as in [26]). It is worth bearing in mind when choosing these nonbasic variables that it is appropriate to show preference for those whose multiplier estimate is zero or small enough, in order to avoid the vector Δx derived from the superlinear-order update being infeasible. Let $\widehat{N}_A$ denote the submatrix constituted by these columns of N_A; thus, a new variable reduction matrix $\overline{Z}_A$ is defined by

$$\overline{Z}_A = [Z_A \quad Z_{\widehat{N}}] = \begin{bmatrix} -B_A^{-1}S_A & -B_A^{-1}\widehat{N}_A \\ \mathbf{1} & 0 \\ 0 & \mathbf{1} \\ 0 & 0 \end{bmatrix}. \tag{24}$$

Two rules can be used to exploit the current factorization of the quasi-Newton approximation of the reduced Hessian of the augmented Lagrangian.

1. *Updating of the Cholesky by finite differences.* Denote

$$\widehat{Z}_A^{(i)} = [Z_A, z_1, \cdots, z_i] = [\widehat{Z}_A^{(i-1)}, z_i],$$

for $i = 1, \cdots, m_1$, where m_1 is the column number of $\widehat{N}_A$ and z_i denotes the i-th column of $Z_{\widehat{N}}$ such that $\widehat{Z}_A^{(0)} = Z_A$ and $\widehat{Z}_A^{(m_1)} = \overline{Z}_A$.

The procedure used here consists in repeating successively from $i = 1$ to $i = m_1$ the classical method suggested in [26] for updating the Cholesky factor R when adding a new superbasic variable, in which $\overline{Z}$, Z and z are replaced, for each i, by $\widehat{Z}_A^{(i)}$, $\widehat{Z}_A^{(i-1)}$ and z_i respectively.

2. *Another alternative to keep the positive definiteness.* A new approximation $\widetilde{H}_k$ is taken such that:

$$\widetilde{H}_k = \begin{bmatrix} H_k & 0 \\ 0 & \mathbf{1} \end{bmatrix},$$

where each element of $\mathbf{1}$ correspond to a variable whose corresponding column in A belong to $\widehat{N}_A$. Therefore, in order to factor $\widetilde{H}_k = \widetilde{R}_k^t \widetilde{R}_k$ it is enough to set

$$\widetilde{R}_k = \begin{bmatrix} R_k & 0 \\ 0 & \mathbb{1} \end{bmatrix}.$$

2.6. Approximation of the reduced Hessian of the augmented Lagrangian by finite differences

Next, an approximation of the reduced Hessian of the augmented Lagrangian is calculated by extending the usual lines in the approximation procedure of the Hessian by means of finite differences (see [12]).

Denote $g(x) = \nabla_x L_\rho(x,\mu)$ and $G(x) = \nabla^2_{xx} L_\rho(x,\mu)$. The Taylor-series expansion of g about x allows us to write $g(x+\sigma d) \simeq g(x) + \sigma G(x)d$. Premultiplying by the variable

reduction matrix Z_A^t and taking into account that $d = Z_A d_S$, we obtain

$$Z_A^t g(x+\sigma Z_A d_S) \simeq Z_A^t g(x) + \sigma Z_A^t G(x) Z_A d_S,$$

from which

$$Z_A^t G(x) Z_A d_S \simeq \frac{Z_A^t g(x+\sigma Z_A d_S) - Z_A^t g(x)}{\sigma}.$$

Let us consider a given $\sigma > 0$, sufficiently (but not too) small, $d_S = e_i$ is successively taken for $i = 1, \cdots, s$, where e_i is the unit vector that has the one in the i-th position, the remaining entries being zero, and s is the number of columns of matrix Z_A. Then, we can obtain an approximation of the i-th column of $Z_A^t G(x) Z_A$, which is denoted by $\overline{H}^{(i)}$, by means of

$$\overline{H}^{(i)} = \frac{Z_A^t g(x+\sigma Z_A e_i) - Z_A^t g(x)}{\sigma},$$

thus achieving $\overline{H} \simeq Z_A^t G(x) Z_A$.

Once $\overline{H}$ is computed a symmetrical approximation H is obtained by doing $H_{ij} = = 1/2(\overline{H}_{ij} + \overline{H}_{ji})$ in order to calculate each entry (i,j) of H; in passing, $\overline{H}_{ij}$ and $\overline{H}_{ji}$ are compared, and if the difference between them is excessive, this approximation must be rejected and a more suitable σ must be chosen. One classic possibility is to take $\sigma = \sqrt{\varepsilon_M}$ (ε_M=machine precision).

3. EXTENSION OF THE NEWTON-LIKE METHODS TO PROBLEM IGP

In [21] the original multiplier method is extended to problems of the **IGP** type (see (7–10) in §1), specializing its performance in the case of nonlinear network flows with side constraints. There an algorithm for solving this problem using the aforementioned method is put forward, which is summarized in §3.1.

When we have $\alpha = \beta = 0$, problem **IGP** becomes problem **EGP**. As can be seen in §2, in the solution of this problem, subproblem **EGS** (5) must be solved and the vector μ updated.

Here the extension of update U_{INW} of §2.1 to problem **IGP** for $\alpha < \beta$ is considered. At each iteration of the multiplier method we solve subproblem **(IGS)**:

$$\begin{array}{lll} & \underset{x}{\text{minimize}} & L_\rho(x,\mu) \\ (25) & \text{subject to} & Ax = b \\ & & l \leq x \leq u, \end{array}$$

where $\rho > 0$ and μ are fixed,

$$L_\rho(x,\mu) = f(x) + \sum_{j=1}^{r} p_j[c_j(x), \mu_j, \rho], \tag{26}$$

p_j being defined by

$$p_j[c_j(x),\mu_j,\rho] = \mu_j\varphi_j[c_j(x),\mu_j,\rho] + \frac{1}{2}\rho|\varphi_j[c_j(x),\mu_j,\rho]|^2, \tag{27}$$

such that

$$\varphi_j[c_j(x),\mu_j,\rho] = \begin{cases} c_j(x)-\beta_j & \text{if } \mu_j+\rho[c_j(x)-\beta_j] > 0 \\ c_j(x)-\alpha_j & \text{if } \mu_j+\rho[c_j(x)-\alpha_j] < 0 \\ -\mu_j/\rho & \text{otherwise.} \end{cases} \tag{28}$$

Replacing in (27) $\varphi_j[c_j(x),\mu_j,\rho]$ with its expression in (28) and operating we get the function $p(t,\mu,\rho)$ defined by Bertsekas in §3.2 of [2], which is continuously differentiable with respect to t and twice continuously differentiable for all t except $t = \alpha-\mu/\rho$ and $t = \beta-\mu/\rho$, see [2].

In [21] the following proposition is shown.

Proposition 1 (a) *If f and c are continuous in a subset $\mathcal{X}$ of $\mathbb{R}^n$, $L_\rho(\cdot,\mu)$ is continuous in $\mathcal{X}$ for all μ, and $\rho > 0$.*

(b) *If f and c are continuously differentiable in an open subset $\mathcal{X}$ of $\mathbb{R}^n$, $L_\rho(\cdot,\mu)$ is continuously differentiable in $\mathcal{X}$ for all μ, and $\rho > 0$.*

(c) *Let $K = \{1,\ldots,r\}$. If f and c are twice continuously differentiable in an open subset $\mathcal{X}$ of $\mathbb{R}^n$, $L_\rho(\cdot,\mu)$ is twice continuously differentiable in the set*

$$\begin{aligned} \hat{\mathcal{X}}_{\mu,\rho} = \quad & \{x \mid x\in\mathcal{X},\ \mu_j+\rho[c_j(x)-\beta_j] \neq 0,\ j\in K\} \\ & \cap\{x \mid x\in\mathcal{X},\ \mu_j+\rho[c_j(x)-\alpha_j] \neq 0,\ j\in K\} \end{aligned} \tag{29}$$

for all μ, and $\rho > 0$.

On the other hand, let E denote the *state vector* defined for a given pair (μ,ρ) as follows:

$$E_j = \begin{cases} +1, & \text{if } c_j[x(\mu,\rho)]-\beta_j > -\mu_j/\rho \\ -1, & \text{if } c_j[x(\mu,\rho)]-\alpha_j < -\mu_j/\rho \\ 0, & \text{otherwise.} \end{cases} \tag{30}$$

Let $\hat{c}(x)$ denote the active contraints of $c(x)$ (we say that $c_j(x)$ is active in the sense that $E_j \neq 0$; see (30)) and $\hat{\mu}$ its corresponding multiplier vector. Moreover, the zero superscript (e.g., as in $\Delta\mu^0$ and μ^0/ρ) means that the symbol with which is associated refers to the inactive constraints (in the sense that $E_j = 0$), and $\hat{\varphi}$ stands for the current value

of the vector function constituted by the φ_j functions (defined in (28)) corresponding to those constraints $c_j(x)$ which are active. Using this notation, let us consider any update formula U of μ in the solution of **IGP** by means of multiplier methods, where the subproblem to solve is **IGS** (25). In **IGP** the first-order multiplier estimate is given as

$$U_0(x,\mu_j,\rho) = \begin{cases} \mu_j + \rho\varphi_j[c_j(x),\mu,\rho], & \text{if } E_j \neq 0 \\ 0, & \text{otherwise.} \end{cases} \tag{31}$$

Let us assume that $(x^*,\mu^*,\pi^*,\lambda^*)$ is an optimizer at the end of the sequential procedure satisfying the strict complementary condition, then there exists a neighborhood $N(\mu^*)$ such that for $\mu \in N(\mu^*)$, vector $x(\mu,\rho)$ belongs to $\widehat{\mathfrak{X}}_{\mu,\rho}$, see (29).

First-order optimality conditions to be fulfilled by the optimizer $(x^*,\mu^*,\pi^*,\lambda^*)$ for all $\rho > 0$ large enough, are —in addition to the complementary slackness condition over λ^* (strict by hypothesis)— system

$$\begin{aligned} \nabla_x L_\rho(x,\mu) + A^t\pi + \lambda &= 0 \\ \varphi[c(x),\mu,\rho] &= 0 \\ Ax &= b, \end{aligned} \tag{32}$$

where

$$\nabla_x L_\rho(x,\mu) = \nabla f(x) + \nabla c(x)\Big(\mu + \rho\varphi[c(x),\mu,\rho]\Big).$$

Let $\tilde{x} = x(\mu,\rho)$ be a solution of subproblem **IGS**, see (25). Bearing in mind that both φ and $\nabla_x L_\rho(x,\mu)$ are differentiable with respect to x in $\widehat{\mathfrak{X}}_{\mu,\rho}$, the solution of system (32) by means of Newton's method at $x = \tilde{x}$ means solving the system

(33)

n	r_0	$\hat{r}$	m	$\hat{t}$		
$\nabla^2_{xx} L_\rho(\tilde{x},\mu)$	0	$\nabla\hat{c}(\tilde{x})$	A^t	0	Δx	$-\nabla_x L_\rho(\tilde{x},\mu)$
				$\mathbb{1}$		$-A^t\pi - \lambda$
0	$\frac{-\mathbb{1}}{\rho}$				$\Delta\mu^0$	μ^0/ρ
$\nabla\hat{c}(\tilde{x})^t$		0			$\Delta\hat{\mu}$	$-\hat{\varphi}$
A					$\Delta\pi$	0
0 $\mathbb{1}$					$\Delta\lambda$	0

where $\frac{-\mathbb{1}}{\rho}$ stands for the diagonal matrix whose nonzero entries are equal to $-1/\rho$. Note that necessarily $n > m + \hat{r} + \hat{t}$ and $r_0 + \hat{r} = r$.

As in the case of system (14), we can premultiply system (33) with matrix $\overline{Z}$ defined as in (17), consider that the matrix of coefficients is nonsingular and use the equality (19), in order to finally obtain the system

(34)
$$\begin{array}{ccc|c|c} s & r_0 & \widehat{r} & & \\ \hline Z_A^t \nabla^2_{xx} L_\rho(\widetilde{x},\mu) Z_A & 0 & Z_A^t \nabla \widehat{c}(\widetilde{x}) & \Delta x_S & -Z_A^t \nabla_x L_\rho(\widetilde{x},\mu) \\ \hline 0 & \frac{-1}{\rho} & & \Delta\mu^0 & \mu^0/\rho \\ \hline \nabla \widehat{c}(\widetilde{x})' Z_A & & 0 & \Delta\widehat{\mu} & -\widehat{\varphi} \end{array} = ,$$

which is called the *reduced Newton system associated with problem* **IGP**. The solution of this system provides $(\Delta x_S, \Delta\mu^0, \Delta\widehat{\mu})$.

Note that subsystem associated with submatrix $-\mathbb{1}/\rho$ is fully independent of the rest, its solution being

$$\Delta\mu^0 = -\mu^0,$$

where μ^0 is the value of the multiplier vector of the inactive constraints before it is updated, from which we find that the multiplier estimate of these constraints after being updated is zero. Therefore, the system to solve is the same as (20), but replacing $\Delta\mu$, $\nabla c(\widetilde{x})$, and $c(\widetilde{x})$ with $\Delta\widehat{\mu}$, $\nabla\widehat{c}(\widetilde{x})$, and $\widehat{\varphi}$ respectively. The solution is performed by means of the Algorithm 1, but with the above substitutions. Therefore for problem **IGP** we have

(35)
$$U_{INW}(x,\mu_j,\rho) = \begin{cases} \mu_j + \Delta\widehat{\mu}_j & \text{if } E_j \neq 0 \\ 0, & \text{otherwise.} \end{cases}$$

3.1. Algorithm for solving problem IGP

The techniques for the partial elimination of constraints will be used so as to relax only the general constraints (9) of problem **IGP** (see (7–10)), i.e. both constraints (8) and (10) being explicitly maintained. In this way, for each pair (μ^k, ρ^k), the following partial augmented Lagrangian is obtained:

(36)
$$L_{\rho^k}(x,\mu^k) = f(x) + \sum_{j=1}^{r} p_j[c_j(x), \mu_j^k, \rho^k],$$

where the function p_j is defined as in (27–28). This gives rise to the subproblem ($\mathbf{IGS_k}$) —see the definition of **IGS** in (25)—, which will be solved approximately for each k *major iteration* by Murtagh and Saunders's active set technique. The minimization procedure for $\mathbf{IGS_k}$ will be interrupted when vector $x^k = x(\mu^k, \rho^k)$ satisfies a predetermined stopping criterion which becomes more rigorous as k increases, so that the minimization will be *asymptotically exact*, as indicated in §2.5 of [2]. In the development of the algorithm, for each k major iteration, the *state vector* E^k (see (30)) and the *violation vector* V^k are used, where

$$V_j^k = \begin{cases} c_j(x^k) - \beta_j, & \text{if } E_j^k = +1 \\ c_j(x^k) - \alpha_j, & \text{if } E_j^k = -1 \\ 0, & \text{otherwise.} \end{cases} \tag{37}$$

Algorithm 2 (Multiplier method associated with IGP)

Step 0. Initialization. μ^0, ρ^0, γ_0 and the auxiliary vector V^{-1} are appropriately selected; set $k = 0$ and tolerances ε_{opt} and τ are set to sufficiently small values.

Step 1. Solution of subproblem. Subproblem $\mathbf{IGS_k}$ is solved by Murtagh and Saunders's active set technique. The subproblem is considered solved when the algorithm associated with that active set technique stops for an optimality tolerance equal to τ_k, which is defined as

$$\tau_k = \bar{\tau} * \nu(\pi_k),$$

where

$$\bar{\tau} = \max\{\bar{\varepsilon}_k, \varepsilon_{opt}\}, \quad \bar{\varepsilon}_k = \min\{\varepsilon_k, \gamma_k \|V^k\|\}, \text{ and}$$

$$\nu(\pi_k) = \max\left\{1, \frac{\|\pi_k\|_1}{\sqrt{m}}\right\},$$

π_k being the multiplier vector associated to the constraints $Ax = b$ at the local minimum obtained for $\mathbf{IGS_k}$, x^k. ($\nu(\pi)$ is a function used by MINOS 5.5 [27].)

Step 2. Feasibility of x^k with respect to the general constraints. If $T_{opt} := \bar{\tau} \leq \varepsilon_{opt}$,

(a) if at x^k,

$$\|V^k\|_\infty < \tau \|c(x^k)\|,$$

is fulfilled, go to step 7,

(b) otherwise, go to step 3.

If T_{opt} is not true, ε_k and γ_k are updated so that $\{\varepsilon_k\} \longrightarrow 0$ and $\{\gamma_k\} \longrightarrow 0$ by means of

$$\varepsilon_{k+1} = \theta_1 \varepsilon_k$$

$$\gamma_{k+1} = \theta_2 \gamma_k,$$

such that $0 < \theta_i < 1$ for $i = 1,2$. Go to step 3.

Step 3. Multiplier update. The vector μ that calculates the general constraint multipliers is updated following the expresion

$$\mu_j^{k+1} = U(x^k, \mu_j^k, \rho^k),$$

where U means any update formula of μ.

Step 4. Update of penalty parameter. The penalty parameter ρ^k is updated if the violation of any active constraint j has not been sufficiently reduced. That is, if there exists any j, such that $E_j^k \neq 0$, for which it is verified that $|V_j^k| > \frac{1}{\bar{\gamma}}|V_j^{k-1}|$, then set

$$\rho^{k+1} = \xi \rho^k,$$

where ξ and $\bar{\gamma}$ are such that $\xi, \bar{\gamma} \in (1, 10]$ with $\xi > \bar{\gamma}$.

Step 5. Set $k = k + 1$.

Step 6. Return to step 1 with x^{k-1} as the initial feasible point for the subproblem $\mathbf{IGS_k}$.

Step 7. Optimum solution found. The algorithm is stopped: a local minimum of problem **IGP** has been found.

The implementation of this algorithm when it is specialized for networks gives rise to the code PFNRN. More information about this algorithm and its implementation (for $U = U_0$ in Step 3, see (31)) can be found in [21].

4. IMPLEMENTATION

The algorithm that is implemented here has Algorithm 1 as an alternative Step 3 in Algorithm 2. The implementation in Fortran-77, called PFNRN03, was designed mainly to solve large-scale **IGP** type problems (see (7–10)) when these are nonlinear network problems with nonlinear side constraints, given the existence of efficient algorithms to solve this kind of problem when the side constraints are linear. It makes use of the network structure to improve its efficiency, taking into account the specific procedures for nonlinear networks flows [9, 31] and Murtagh and Saunders's active set procedure using a spanning tree as the basis matrix of the network constraints. PFNRN03 also exploits the sparsity of the Jacobian and of the Hessian (as defined by the user).

In this code, the user can set up the level of accuracy of the optimizer by means of the parameter ε_{opt}, whose default value is 10^{-6}. Note that $\{\bar{\tau}\} \longrightarrow \varepsilon_{opt}$ as $\{\varepsilon_k\} \longrightarrow 0$ —i.e., an asymptotically exact minimization is used, as Bertsekas suggests in [2].

For Step 2 the user can also set up a level of feasibility accuracy for the side constraints by giving a value to the parameter τ. This value is 10^{-5} by default. Also, the values of ε_0 and γ_0 can be modified, $\varepsilon_0 = 0.01$ and $\gamma_0 = 0.25$ being the default ($\theta_1 = 0.5, \theta_2 = 0.25$ are fixed).

In Step 3, when it is performed by Algorithm 1, the values of the tolerances in the activation and validation rules are $\tau_x = \varepsilon_{opt}$, $\tau_\mu = 0.1$, and $\tau_{nwl} = 1.0$.

The last values have been chosen in order to have a sufficiently large amount of major iterations where the superlinear-order estimation is used, it allows us to evaluate the performance of this estimator. The initial multiplier estimate vector μ^0 is the null vector by default, although it is worth using information on this vector when it is available, given the substantial improvement that it may bring to bear on the convergence of the algorithm, as was found by means of numerical experiments.

In Step 4 the default values are $\rho^0 = 0.1$ and $\xi = 2.0$. These values can also be modified by the user. Likewise, $\bar{\gamma} = 0.9\xi$, as Bertsekas suggests in [2]. Note that if $|V_j^k| \leq \frac{1}{\bar{\gamma}}|V_j^{k-1}|$, ρ does not grow in the k-th iteration. Moreover, when the current μ^k is sufficiently close to μ^* —i.e., $\|\mu^{k+1} - \mu^k\|_1 < \eta_0(1 + \|\mu^k\|_1)$ for a small enough scalar η_0 (here $\eta_0 = 0.01$)— $\xi = 1$, so ρ^k stops growing in order not to increase the ill-conditioning at the end of the execution. In practice this rule is very stringent, hence ρ is sufficiently large when it is true. This technique is used by MINOS 5.5 to reduce ρ or to set it to zero, see [27].

Furthermore, by default, the constraints are initially multiplied by scale factors that make the norm of $\nabla c_j(x^0)$ equal to 1, for $j = 1, \ldots, r$. The user may eliminate this scaling.

This implementation includes the two following choices with respect to Algorithm 1:

1. using this algorithm without changes, giving rise to the update U_{S1}
2. using the algorithmic alternative given in §2.4, giving rise to the update U_{S2}.

In addition, with regard to the reduced Hessian H_k of the augmented Lagrangian function at x^k three different choices have been considered:

(H) the exact reduced Hessian, if it is positive definite, or otherwise the positive definite approximation achieved by means of the modified Cholesky method (see §4.4.2.2 in [12]);

(D) an approximation by finite differences achieved from projected gradients as is put forward in §2.6; and

(Q) a quasi-Newton approach by means of the BFGS update, using the first rule given in §2.5 when the projection J_k of the Jacobian over the manifold defined by the current nonbasic variables does not have full column rank. We must not forget that BFGS update is only used when the number of superbasic variables s is not greater than $\bar{s}$ ($\bar{s} = 500$ by default). In this implementation the alternative method to the BFGS is the Truncated Newton [8]), see §2.5.

In consequence, by combining the two type of updates (U_{S1} and U_{S2}) and the above three choices for the projected Hessian we have several updates that are different in practice, and which are denoted by expressions such as U_{S1H}, U_{S1D} and U_{S1Q}, where these mean that update U_{S1} is calculated by using the alternatives H, D and Q respectively. The same is true for update U_{S2}. If $s > \bar{s}$ when using U_{S1Q}, the alternative update is U_O. Nevertheless, for U_{S2} it is not possible to compute U_{S2Q} because we do not have a quasi-Newton approximation of the reduced Hessian of the Lagrangian function (whereas we do have one of the augmented Lagrangian function).

In this package there are other parameters that can also be modified by means of a specification file, see [20].

The code PFNRN03 and its user's guide [20] can be accessed *for academic purposes* via *anonymous ftp* at `ftp-eio.upc.es` in directory: `pub/onl/codes /pfnrn`. (Comments, suggestions, or description of any trouble or abnormal situations experienced when using it reported to `mepmifee@lg.ehu.es` will be appreciated.)

Code PFNRN03 contains the solver PFNL for nonlinear network flow problems, which is used in Step 1 (subproblem solution) of the Algorithm 2. Any other nonlinear network flow code could have been employed instead.

5. COMPUTATIONAL TESTS

In order to evaluate the efficiency and robustness of code PFNRN03 a series of computational tests are performed, which consist of solving nonlinear network flow problems with linear and nonlinear side constraints with this code in its various versions (depending on the kind of U_S) and comparing it with the results obtained when the multiplier estimation is carried out by means of U_O. All these tests were performed on a Sun Sparc 10/41 work station under UNIX.

Two types of problems are considered: real and artificial. The real problems solved are of Short-Term Hydrothermal Coordination of Electricity Generation [21]. They correspond to problems whose name starts with «P», and they have *two-sided inequality*

nonlinear side constraints. The objective function in strongly nonlinear, with a high computational cost.

Table 1. Test problems.

test	arcs	nodes	s.c.	j.e.n.	a.s.c	sb.a.
D12e2	1524	360	180	183	3	75
D13e2	1524	360	360	364	9	89
D14e2	1524	360	36	538	31	151
D15e2	1524	360	36	5387	4	107
D16e2	1524	360	36	96	20	83
D13n1	1524	360	360	364	38	681
D21e2	5420	1200	120	135	2	30
D22e2	5420	1200	600	656	13	46
D23e2	5420	1200	120	135	17	238
D31e1	4008	501	5	20	1	63
D32e1	4008	501	50	199	5	83
D31e2	4008	501	5	20	1	60
D32e2	4008	501	50	199	5	65
D41e2	12008	1501	15	182	9	172
D51e1	18000	3000	30	526	8	58
D52e1	18000	3000	150	27118	16	166
P4103	3591	1072	63	3071	10	17
P4101	3591	1072	63	3071	32	63
P4203	5187	1548	91	4443	10	17
P4201	5187	1548	91	4443	32	65
P4403	7581	2262	133	6501	10	20
L13e2	1524	360	360	363	7	67
L21e2	5420	1200	120	138	5	27
L32e2	4008	501	120	138	4	61
L52e1	18000	3000	30	527	10	57
XA48	2256	697	240	1915	145	265

The model that gives rise to the problem «xa» is presented in [16] as a single network model for Hydrothermal Scheduling obtained by joining the hydrogeneration optimization network and a series of thermal subnetworks (one for each interval), all of them with a common sink node. The objective function to be minimized is quadratic with both *equality and inequality linear side constraints.*

Problems created from DIMACS random network generators [10] are used as artificial problems: generators Rmfgen, Gridgen and Grid-on-Torus have been employed, linear side constraints being created from the *Di2no* random generator [17], and they give rise to the problems whose first letter is an «L». The nonlinear side constraints for the DIMACS networks are generated through the *Dirnl* random generator described in [21],

and they give rise to whose first letter is an «D». Several types of objective function used to create problems with DIMACS networks. The last two letters in the problem names starting with L or D indicate the type of objective function: «n1» functions are of the Namur family used in [31] and «e1» and «e2» functions belong to the EI01 family, which was designed by Heredia in [17] (see also [21]) with a view to obtaining problems with a moderate number of superbasic variables at the optimizer.

Table 1 shows the characteristics of the problems resulting from these models. The names of the test problems are given in the *test* column. The *arcs* and *nodes* columns give us, respectively, the number of arcs and nodes in the network, whilst *s.c.* and *j.e.n.* give, respectively, the total number of the problem's side constraints and the number of non-null entries in their Jacobian. Finally, the *a.s.c.* and *sb.a.* give, respectively, the number of side constraints that are active and the number of variables that are superbasic at the optimum (superbasic variables with respect to the nonlinear network subproblem that is solved in each major iteration of Algorithm 2).

Next there are two subsections dedicated to the computational results. In the first the results for the various versions of U_S are compared with each other and with U_O. The second part provides only the efficiency evaluation of the quasi-Newton update U_{S1Q} with respect to the first-order update U_O, though for a greater number of test problems than in the first part.

5.1. Comparison of results

Each test problem was solved by using the five types of update U_S considered here. The only aspect that is modified in each solution of any test problem is the type of update U_S with which it is solved. The results have been grouped into two different tables, 2 and 3, as the multiplier update will be U_{S1} or U_{S2}. Each table presents the results for the three types of H_k (projected Hessian matrix of the augmented Lagrangian function, or an approximation of it, in the k-th major iteration) that are used in this implementation —i.e., H, D and Q, see §4— except for U_{S2Q}, given that the quasi-Newton approximation of the projected Hessian of the Lagrangian function is not available. The place of this latter type of update in this table is occupied by the update U_O, which is used as a reference.

In each table, the name of the test problem is given under the heading «TEST»; the columns «Mit», «mit» and «cpu» are used to give, respectively, the major iteration number, the minor iteration number (total number of iterations used to solve the consecutive network subproblems **IGS** (25); see §3.1) and the CPU seconds spent by the execution.

Table 2. Results for update U_{S1} and various H_k.

TEST	U_{S1H}			U_{S1D}			U_{S1Q}		
	Mit	mit	cpu	Mit	mit	cpu	Mit	mit	cpu
D12e2	18	890	27.4	18	890	21.0	14	790	14.4
D13e2	15	2028	45.7	15	2018	45.9	13	1938	30.9
D13n1	14	3652	5582.8	14	3660	750.5			(†)
D23e2	15	3124	344.9	15	3121	231.9	13	3194	161.7
D32e1	35	3802	265.8	35	3802	201.2	9	3951	116.5
D51e1	130	1277	6981.8	121	1285	594.9	8	1116	91.7
P4203			(⋆)	12	4038	428.5	10	3167	328.1
P4403	8	3934	722.2	12	5563	1022.4	9	4780	804.0
L13e2			(‡)			(‡)	9	508	6.1
L21e2	42	1801	119.6	42	1801	76.9	10	1724	22.6
L32e2			(‡)			(‡)	7	856	20.1
L52e1	6	4096	250.5	6	4096	222.5	6	4117	221.8

Table 3. Results for update U_{S2} and various H_k, and U_O.

TEST	U_O			U_{S2H}			U_{S2D}		
	Mit	mit	cpu	Mit	mit	cpu	Mit	mit	cpu
D12e2	25	1184	15.6	18	869	22.5	18	869	17.6
D13e2	20	2497	36.3	15	2064	45.7	15	2052	38.3
D13n1	18	3869	478.3	14	3662	5581.7	14	3649	792.8
D23e2	33	4948	279.0	14	3103	306.6	14	3103	194.9
D32e1	28	5114	136.8	35	4437	280.0	35	4437	218.2
D51e1	12	1170	97.0	8	1099	202.2	8	1099	93.5
P4203	7	3071	303.0	6	2859	277.7	7	2895	291.3
P4403	7	3681	547.4	9	4331	716.3	8	4560	748.4
L13e2	16	537	6.4	8	506	6.7	8	506	5.7
L21e2	39	1848	24.2	14	1963	53.7	14	1963	40.4
L32e2	9	868	21.5	6	849	25.8	6	849	23.2
L52e1	6	4116	219.6	6	4116	252.1	6	4116	221.3

Table 2 shows the results for update U_{S1} (which is directly derived from the Algorithm 1) with the three choices of the matrix H_k considered in §4. Note that the running times of U_{S1H} are greater, in nearly all cases, than those of U_{S1D}, at the same time as the running times of U_{S1D} are greater than those of U_{S1Q} in all cases. The reason of this lies in the definition of H_k and in the way in which this matrix is computed. If H is the authentic reduced Hessian of the augmented Lagrangian one must compute the Hessian of the augmented Lagrangian —$\nabla^2_{xx}L_\rho(x,\mu)$— and then its projection —$Z_A^t\nabla^2_{xx}L_\rho(x,\mu)Z_A$—,

whereas if H is an approximation by finite differences achieved from projected gradients as is put forward in §2.6, H is directly computed column by column. On the other hand, if H is a quasi-Newton approximation (see §4), it is supplied together with the solution of the subproblem.

In this table (†) means that problem «D13n1» cannot be solved by using a quasi-Newton approach because in this problem the number of superbasic variables is almost 700, higher than $\bar{s}$ (here 500), and when this happens the method used is the Truncated Newton (see §4), which leads to this approximation being unavailable in problem «D13n1». In addition, (‡) means that the execution stops at the end of a large number of major iterations (e.g., in «L13e2» more than 480) since it does not succeed in improving the current point, as a consequence of a bad updating of the multipliers, which can be attributed to numerical stability difficulties in the solution of Step 4 in Algorithm 1. However, this trouble does not appear in the solution of the same problem by U_{S2}, as can be seen in Table 3, which confirms the greater numerical stability of the update U_{S2}, see the definition of U_{S2} in §4. In «P4203» ($\star$) means that the execution stops at the end of 14 major iterations and 4533 minor iterations as a result of the bad convergence caused by an inaccurate estimation of the multipliers. This in turn is a consequence of the reasons given for (‡), together with the fact that, when using the modified Cholesky factorization, the Frobenius norm of the modification of the projected Hessian used to make it into positive definite is of the order of 10^{-3}. However, Table 3 shows how the results for this problem improve when U_{S2} is used. In Table 3 the symbols (†), (‡) and ($\star$) (if they appear) mean the same as in Table 2.

As is shown in Tables 2 and 3, and as is expected in theory, if we compare for each problem the number of minor iterations when U_S is used, it is generally lower than when U_O is used, whereas, in contrast, the running times are higher. This last happens because of the high computational cost of calculating the superlinear-order estimate of the multiplier vector with regard to the low cost of using U_O. Moreover, it could happen that in Algorithm 1 condition T_{nw0} would hold, but T_{nw1} would not; then all the computations carried out in this algorithm would have been wasted.

On the other hand, the estimation of the multipliers in problems «P4203» and «P4403» is not very good because of the two following reasons: the projected Hessian of the augmented Lagrangian function is indefinite in these problems and the aforementioned issues of numerical stability (e.g., in «P4403» the Frobenius norm of the modification matrix of the modified Cholesky factorization has occasional values of the order of 10^7).

The high number of major iterations in problem «D51e1» when it is solved by using U_{S1H} and U_{S1D} is due to the active side constraints having —in some cases— a very low activity, and matrix H_k being indefinite but almost positive semidefinite. Hence the modification matrix of the Cholesky factorization has a Frobenius norm higher than zero, which yields numerical stability problems. However, these difficulties are not encountered when using U_{S2H} and U_{S2D} due to the reasons given above.

An important effect noticed in many of these tests, and which occurs when using any U_S superlinear-order update, is that although the initial conditions are the same, the final value of the penalty parameter required for the convergence of the algorithm can be much smaller than that obtained when using the first-order update; e.g., in problem «D23e2», $\rho^* = 33554432$ for U_O and, however, $\rho^* = 40$ for any U_S, the initial values being always $\rho^0 = 10$ and $\xi = 2$, see Algorithm 2 in §3.1. Hence, as a consequence of performing the multiplier estimation with U_S updates, the ill-conditioning is reduced when solving subproblem **IGS** (25) in each major iteration. This effect has two numerical consequences: firstly, the algorithm converges faster when a U_S update is used, reducing the number of minor iterations (as is shown in these tables); and secondly, the execution with U_O of some tests breaks down because of ill-conditioning, whereas with U_S it finished succesfully.

Table 4. Comparison of efficiency with regard to U_O.

TEST	U_{S1H}	U_{S1D}	U_{S1Q}	U_{S2H}	U_{S2D}
D12e2	0.57	0.74	**1.08**	0.69	0.89
D13e2	0.79	0.79	**1.17**	0.79	0.95
D13n1	0.09	0.64	(†)	0.09	0.60
D23e2	0.81	**1.20**	**1.73**	0.91	**1.43**
D32e1	0.51	0.68	**1.17**	0.49	0.63
D51e1	0.01	0.16	**1.06**	0.48	**1.04**
P4203	(⋆)	0.71	0.92	**1.09**	**1.04**
P4403	0.76	0.54	0.68	0.76	0.73
L13e2	(‡)	(‡)	**1.05**	0.96	**1.12**
L21e2	0.20	0.31	**1.07**	0.45	0.60
L32e2	(‡)	(‡)	**1.07**	0.83	0.93
L52e1	0.88	0.97	0.99	0.87	0.99

In order to give a clearer idea of the efficiency of the different U_S updates in comparison with the update U_O, Table 4 has been constructed (where the symbols (†), (‡) and (⋆) mean the same as in Table 2). For this same reason we have used bold type for all those efficiency values that are one or greater than one, where each efficiency value (ev) is given by the following ratio:

$$ev(XY) = \frac{\text{CPU time using } U_O}{\text{CPU time using } U_{SXY}}, \tag{38}$$

X being either 1 or 2, and Y being H, D or Q, excluding $XY = 2Q$ for the reasons given at the beginning of this section. Therefore, $ev(XY) \geq 1.00$ means that the efficiency of U_{SXY} is at least as good as that of U_O, and hence, the greater $ev(XY)$, the greater the efficiency of U_{SXY} with respect to U_O. Below each heading U_{SXY} we find the different values of $ev(XY)$ for each test problem.

Since the interesting results for the update U_{S1Q} at Table 4, we compare in a more detailed way the efficiency of this update with respect to the first-order update U_O in the following section, in order to decide whether it is worth using U_{S1Q} instead of U_O.

5.2. Efficiency of the update U_{S1Q} regarding U_O

Here we consider the results of Table 5, where the headings are the same as those used in previous tables. The symbol (†) means that the execution was stopped due to ill-conditioning, and (‡) means that since it is not possible to obtain a projected Jacobian having full rank in any major iteration, update U_{S1Q} offers the same results as update U_O. The reason that this Jacobian matrix cannot be found is that some of the active constraints have normals that are orthogonal to the subspace spanned by the columns of matrix Z_A —see (18)— in each major iteration.

Table 5. Efficiency of U_{S1Q} with regard to U_O.

TEST	U_O			U_{S1Q}			$ev(1Q)$
	Mit	mit	cpu	Mit	mit	cpu	
D12e2	25	1184	15.6	14	790	14.4	1.08
D13e2	20	2497	36.3	13	1938	30.9	1.17
D14e2	26	5485	147.5	20	4006	117.1	1.26
D15e2			(†)	19	1909	56.87	(†)
D16e2	850	3966	85.4	16	2345	39.0	2.19
D21e2	22	48669	1155.1	17	2238	40.1	28.81
D22e2	19	5306	119.5			(‡)	(‡)
D23e2	33	4948	279.0	13	3194	161.7	1.73
D31e1	10	1253	29.1	7	1085	30.9	0.94
D32e1	28	5114	136.8	9	3951	116.5	1.17
D31e2	9	947	24.9	7	535	20.0	1.25
D32e2			(†)	17	3661	110.9	(†)
D41e2	32	2842	301.5	10	2143	265.2	1.14
D51e1	12	1170	97.0	8	1116	91.7	1.06
D52e1	10	8388	1258.5	11	8511	1362.9	0.92
P4103	10	2814	165.3	12	2852	177.6	0.93
P4101	8	23687	1520.3	10	21929	1355.3	1.12
P4203	7	3071	303.0	10	3167	328.1	0.92
P4201	7	22845	2159.7	8	22901	2161.1	1.00
P4403	7	3681	547.4	9	4780	804.0	0.68
L13e2	16	537	6.4	9	508	6.1	1.05
L21e2	39	1848	24.2	10	1724	22.6	1.07
L32e2	9	868	21.5	7	856	20.1	1.07
L52e1	6	4116	219.6	6	4117	221.8	0.99
XA48	175	1465	103.8	27	1296	89.5	1.16

Note that except for the problems «D22e2» and «P4403», the efficiency of update U_{S1Q} is similar to or higher than that of U_O, reaching an efficiency value of 28.81 for problem «D21e2» and a value of 2.19 for problem «D16e2». The mean value of the computed efficiencies excluding the highest and lowest values is 1.16. This is not very much larger than one, but we must consider it together with the higher robustness of U_{S1Q}. As an example of this, see the problems «D15e2» and «D32e2» in Table 5, in both cases the execution was stopped due to ill-conditioning. In fact, for «D32e2» when this stopped the penalty parameter value was 65538, whereas using U_{S1Q} the execution normally finished with $\rho_S^* = 4$. Therefore, in the face of the classical difficulty of the augmented Lagrangian techniques the U_S updates are a good alternative, in particular it is worth taking update U_{S1Q} (rather than the first-order update U_O) seriously into account when using augmented Lagrangian techniques with large scale problems, specially with nonlinear networks with nonlinear side constraints.

6. CONCLUSIONS

In this paper the author has put forward techniques for implementing superlinear-order multiplier estimations using the least computational effort.

The projected Jacobian may not have full rank —although it will at the optimum— if the prediction of the active variables at the optimizer provided in the solution of the augmented Lagrangian subproblem is not totally correct, because of the fact that the current point is not sufficiently close to the optimum. Therefore, caution should be exercised when computing the multiplier estimation either to ensure proximity to the optimum or otherwise to avoid using a Newton method, as in this case we do not have the help of the line search.

An algorithm is designed that allows us to solve nonlinear problems with linear constraints, simple bounds, and two-sided nonlinear constraints, by combining superlinear and first-order multiplier methods together with variable reduction techniques. This procedure is particularly interesting in case that the linear constraints are flow conservation equations, as there exist efficient techniques to solve nonlinear network problems.

An implementation of the designed algorithm is put forward, which is specialized for network problems and offers various choices for computing a superlinear-order multiplier estimation. As a result the code PFNRN03 is obtained.

The computational results confirm the quasi-Newton update as a better and more robust alternative than the first-order multiplier estimation to solve large-scale nonlinear network problems with linear or nonlinear side constraints. In addition, we must take into account that since the pure network constraints are the only constraints that are explicitly maintained, whether or not the subproblem solved iteratively is a pure net-

work subproblem only has a significant effect on the total CPU time, but not on the comparison of the efficiency of the different kinds of superlinear-order multiplier estimation used. Furthermore, from the computational tests it is also clear that the update U_{S2} has a greater numerical stability than U_{S1} if we do not include the quasi-Newton case. Another important practical result caused by performing the multiplier estimation by means of superlinear-order updates is the reduction of the difficulties associated with ill-conditioning, as they become unnecessary to increase the penalty parameter to the same extent as when the first-order update is used, and this makes the first more robust than the second.

REFERENCES

[1] Bertsekas, D. P. (1977). «Approximation procedures based on the method of multipliers». *J. Optim. Theory Appl.*, 23, 487-510.

[2] Bertsekas, D. P. (1982). *Constrained Optimization and Lagrange Multiplier Methods.* Academic Press, New York.

[3] Bertsekas, D. P. (1995). *Nonlinear Programming.* Athena Scientific, Belmont, Massachusetts.

[4] Bertsekas, D. P. (1998). *Network Optimization: Continuous and Discrete Models.* Athena Scientific, Belmont, Massachusetts.

[5] Brännlud, H., Sjelvgren, D. and Bubenko, J. A. (1988). «Short Term Generation Scheduling with Security Constraints». *IEEE Transactions on Power Systems*, 3, 1, 310-316.

[6] Conn, A. R., Gould, N. I. M. and Toint, Ph. L. (1991). «A globally convergent augmented Lagrangian algorithm for optimization with general constraints and simple bounds». *SIAM J. Numer. Anal.*, 28(2), 545-572.

[7] Conn, A. R., Gould, N. I. M., Sartenaer, A. and Toint, Ph. L. (1996). «Convergence properties of an augmented Lagrangian algorithm for optimization with a combination of general equality and linear constraints». *SIAM J. Optim.*, 6(3), 674-703.

[8] Dembo, R. S., Eisenstat, S. C. and Steihaug, T. (1982). «Inexact Newton methods». *SIAM J. Numer. Anal.*, 19, 400-408.

[9] Dembo, R. S. (1987). «A primal truncated newton algorithm with application to large-scale nonlinear network optimization», *Math. Program. Stud.*, 31, 43-71.

[10] DIMACS (1991). *The first DIMACS international algorithm implementation challenge: The bench-mark experiments.* Technical Report, DIMACS, New Brunswick, NJ, USA.

[11] Gill, P. E. and Murray, W. (1979). «The computation of Lagrange-Multiplier estimates for constrained minimization». *Math. Program.*, 17, 32-60.

[12] Gill, P. E., Murray, W. and Wright, M. H. (1981). *Practical Optimization.* Academic Press, New York.

[13] Golub, G. H. and Van Loan, Ch. F. (1996). *Matrix computations.* (Third edition) Johns Hopkins, Baltimore and London.

[14] Griewank, A. and Toint, Ph. L. (1982). «Partitioned variable metric updates for large structured optimization problems». *Numer. Math.*, 39, 119-137.

[15] Heredia, F. J. and Nabona, N. (1991). *Large scale nonlinear network optimization with linear side constraints.* EURO XI, European Congress on Operational Research, Aachen, Germany, July.

[16] Heredia, F. J. and Nabona, N. (1995). «Optimum Short-Term Hydrothermal Scheduling with Spinning Reserve through Networks Flows». *IEEE Trans. on Power Systems*, 10(3), 1642-1651.

[17] Heredia, F. J. (1995). *Optimització de Fluxos No Lineals en Xarxes amb Constriccions a Banda. Aplicació a Models Acoblats de Coordinació Hidro-Tèrmica a Curt Termini.* Ph. D. Thesis, Statistics and Operations Research Dept., Universitat Politècnica de Catalunya, Barcelona.

[18] Hestenes, M. R. (1969). «Multiplier and gradient methods». *J. Optim. Theory Appl.*, 4, 303-320.

[19] Kennington, J. L. and Helgason, R. V. (1980). *Algorithms for network programming.* John Wiley and Sons, New York.

[20] Mijangos, E. (1997). *PFNRN03 user's guide.* Technical Report 97/05, Dept. of Statistics and Operations Research. Universitat Politècnica de Catalunya, Barcelona.

[21] Mijangos, E. and Nabona, N. (1996). *The application of the multipliers method in nonlinear network flows with side constraints.* Technical Report 96/10, Dept. of Statistics and Operations Research. Universitat Politècnica de Catalunya, Barcelona.

[22] Mijangos, E. and Nabona, N. (1998). «On the first-order estimation of multipliers from Kuhn-Tucker systems in multiplier methods using variable reduction». In: P. Kall and H.-J. Lüthi Editors, *Operations Research Proceedings 1998*, pp. 63-72. Springer-Verlag, Zürich.

[23] Mijangos, E. and Nabona, N. (1999). «On the compatibility of the classical first order multiplier estimates with the variable reduction techniques when there are nonlinear inequality constraints». *Qüestiio*, 23(1), 61-83.

[24] Mijangos, E. and Nabona, N. (1999). «On the first-order estimation of multipliers from Kuhn-Tucker systems». *Comput. Oper. Res.* To appear.

[25] Mijangos, E. (2000). *Superlinear-order multiplier estimation in nonlinear networks with nonlinear constraints.* EURO XVII, European Congress on Operational Research, Budapest, Hungary, July.

[26] Murtagh, B. A. and Saunders, M. A. (1978). «Large-scale linearly constrained optimization». *Math. Program.*, 14, 41-72.

[27] Murtagh, B. A. and Saunders, M. A. (1983-1998). *MINOS 5.5. User's guide.* Report SOL 83-20R, Department of Operations Research, Stanford University, Stanford, CA, USA, Revised July 1998.

[28] Nocedal, J. (1980). «Updating quasi-Newton matrices with limited storage». *Math. Comput.*, 35, 773-782.

[29] Powell, M. J. D. (1969). «A method for nonlinear constraints in minimization problems». In: *Optimization* (R. Fletcher, ed.), pp. 283-298. Academic Press, New York.

[30] Tapia, R. A. (1977). «Diagonalized multiplier methods and quasi-Newton methods for constrained optimization». *J. Optim. Theory. Appl.*, 22(2), 135-194.

[31] Toint, Ph. L. and Tuyttens, D. (1990). «On large scale nonlinear network optimization». *Math. Program.*, 48, 125-159.

[32] Wang, S. J., Shahidehpour, S. M., Kirschen, D. S., Mokhtari, S. and Irisarrri, G. D. (1994). *Short-Term Generation Scheduling with Transmission and Environement Constraints Using an Augmented Lagrangian Relaxation.* IEEE/PES Summer Meeting, San Francisco, CA, USA, 94 SM 572-8 PWRS.

Estadística Oficial

QÜESTIIÓ, vol. 26, 1-2, p. 175-196, 2002

ALIMENTACIÓN DE MODELOS CUANTITATIVOS CON INFORMACIÓN SUBJETIVA: APLICACIÓN DELPHI EN LA ELABORACIÓN DE UN MODELO DE IMPUTACIÓN DEL GASTO TURÍSTICO INDIVIDUAL EN CATALUNYA

J. LANDETA RODRÍGUEZ
J. MATEY DE ANTONIO
V. RUÍZ HERRÁN
O. VILLARREAL LARRINAGA
Universidad del País Vasco*

El presente artículo presenta un estudio realizado para el Institut d'Estadística de Catalunya, donde se ha aplicado una técnica que trabaja con información subjetiva (el Método Delphi) para obtener unos datos que puedan ser empleados en la alimentación parcial de un modelo matemático, que se nutre principalmente de datos estadísticos (objetivos), con el fin de incrementar la utilidad global del modelo.
El artículo justifica la utilización en determinadas circunstancias de la información subjetiva, describe el método Delphi y plantea una serie de proposiciones referentes a la utilidad, empleo y mejora de esta técnica que se ven refrendadas en el caso expuesto.

Supplying of subjective information to quantitative models: the Delphi Method applied to a model of attribution of the individual tourist expense in Catalunya

Palabras clave: Método Delphi, información subjetiva, tecnologías de la información

Clasificación AMS (MSC 2000): 93C41

* Universidad del País Vasco/Euskal Herriko Unibertsitatea. Instituto de Economía Aplicada a la Empresa. Facultad de Ciencias Económicas y Empresariales. Avda. Lehendakari Aguirre, 83. 48015 Bilbao

– Recibido en marzo de 2001.
– Aceptado en enero de 2002.

1. INTRODUCCIÓN

1.1. Objeto del artículo

En la práctica académica y empresarial es muy habitual necesitar la estimación de valores futuros (o presentes) de variables de las que se desconocen sus datos pasados o, en el caso de conocerlos, se desconfía de su calidad o de la pervivencia de los patrones de comportamiento en sus tendencias observadas con anterioridad.

Las variables estimadas en las condiciones referidas pueden ser útiles en sí mismas (demanda esperada de un producto de nueva creación, por ejemplo) o pueden aportar utilidad al ser incorporadas junto con otras variables a un modelo matemático que integre diferentes variables y relaciones, alimentadas con datos objetivos o subjetivos.

Este último caso es el que se plantea en el presente artículo; la aplicación del Método Delphi, una técnica que trabaja con información subjetiva, para obtener unos datos que puedan ser empleados en la alimentación parcial de un modelo matemático, que se nutre principalmente de datos objetivos obtenidos a partir de un proceso de encuestación, con el fin de incrementar la utilidad global del modelo.

1.2. La información subjetiva y el método Delphi

La *información subjetiva*, es decir, aquella que es obtenida a partir de la filtración de sucesos, experiencias y datos acumulados por el individuo o grupo a través de sus creencias, expectativas u opiniones, constituye una fuente de información alternativa o complementaria a la objetiva que, en ocasiones, es la única disponible o utilizable.

La justificación de su utilización responde al espíritu que guía a las ciencias aplicadas. Estas ciencias comparten la finalidad de ser instrumentos eficaces para la sociedad y, en nuestro caso concreto, de incrementar la utilidad práctica de los decisores, privados o públicos, mediante la reducción de la incertidumbre creciente que envuelve sus acciones de toma de decisiones. Decisiones que, por otra parte, deben ser adoptadas en su momento concreto, independientemente de la existencia o no de información objetiva disponible.

Los modelos y técnicas basadas en este tipo de información están guiados por este objetivo referencial, y deben ser considerados como válidos en la medida que sus outputs proporcionen al decisor más utilidad que la ausencia de ellos, es decir, que su aplicación aporte al decisor una información que lo coloque en una situación más favorable que la que procede de la simple intuición y suerte.

El acudir a la información subjetiva y, por ende, al juicio de los expertos, no implica renunciar a la metodología científica, sino entenderla de una manera menos restringida. La forma de conseguirlo debe ser a través del desarrollo de técnicas que hagan aparecer este tipo de información de la forma más explícita, razonada y sistemática posible, garantizando su efectividad y el nivel más elevado alcanzable de objetividad. A tal fin se debe actuar básicamente en tres campos de acción (Helmer 1983:58):

a) Incidiendo en la mejora en la selección de las fuentes de información (los expertos) más apropiadas para cada caso concreto, fijando criterios de selección para ello.

b) Ayudando al desenvolvimiento eficaz de la actividad del experto, principalmente mediante la definición precisa de la situación para la que se le requiere, el suministro de la información pertinente y la facilitación de la interacción con otros expertos.

c) Desarrollando métodos de previsión y/o obtención de información que establezcan una metodología de actuación que garantice un alto nivel de calidad en la información conseguida mediante su aplicación.

Es este el contexto que justifica la aparición y pervivencia de las técnicas de decisión y previsión de grupo basadas en información subjetiva.

El *método Delphi* se encuadra dentro de este conjunto de técnicas, ya que es una técnica de investigación social que tiene como objeto la obtención de una opinión grupal fidedigna a partir de un grupo de expertos.

Es un método de estructuración de la comunicación entre un grupo de personas que pueden aportar contribuciones valiosas para la resolución de un problema complejo.

Fue concebido en los años cincuenta con fines militares y a partir de la década de los sesenta ha sido utilizado en los ámbitos académicos y empresariales. Ha sido empleado principalmente como técnica de previsión y consenso en situaciones de incertidumbre, en las que no es posible acudir a otras técnicas basadas en información objetiva. Sus principales características son las siguientes:

- Es un proceso iterativo. Como mínimo, los expertos deben ser consultados dos veces sobre la misma cuestión, de forma que puedan volver a pensar su respuesta ayudados por la información que reciben de las opiniones del resto de los expertos.
- Mantiene el anonimato de los participantes, o al menos de sus respuestas, ya que éstas van directamente al grupo coordinador. Ello permite poder desarrollar un proceso de trabajo en grupo con unos expertos que no coinciden ni temporal ni espacialmente, y además busca evitar las influencias negativas que en las respuestas individuales pudieran tener factores relativos a la personalidad de los expertos participantes.

- Feedback controlado. El intercambio de información entre los expertos no es libre, sino que se realiza a través del grupo coordinador del estudio, con lo que se elimina toda información que no sea relevante.
- Respuesta estadística de grupo. Todas las opiniones forman parte de la respuesta final. Las preguntas están formuladas de forma que se pueda realizar un tratamiento cuantitativo y estadístico de las respuestas.

Para su empleo es necesario constituir un equipo coordinador del estudio y contar con la colaboración de un grupo apropiado de expertos.

El equipo coordinador debe estar formado por un reducido número de especialistas conocedores de la técnica Delphi y de la realidad objeto de estudio, a fin de poder aplicar el método correctamente e interpretar adecuadamente las opiniones y aportaciones del grupo de expertos colaboradores.

El grupo o panel de expertos es el eje central del método, en tanto que son los que proveen la información que, después del correspondiente proceso de iteración, interacción y agregación, se convertirá en la opinión grupal y, por consiguiente, en el output de la investigación.

El método Delphi ha atravesado distintas épocas desde su nacimiento en los años 50. Desde una primera etapa de secretismo que acompañó a su génesis con fines militares, ha avanzado a través de sucesivas etapas de novedad, popularidad, crítica y reexamen, hasta llegar a la fase actual de empleo continuo, pero relativamente escaso, en la que permanece desde los años 80 (Rieger, 1986; Landeta, 1999; p. 38).

La utilidad social de esta técnica y el valor de sus aportaciones metodológicas siguen vigentes y, en determinadas circunstancias, se compara favorablemente con las técnicas que comparten su ámbito de actuación (grupos cara a cara, grupo nominal, brainstorming), lo que la convierte en un técnica que puede ser racionalmente empleada cuando se dan las condiciones adecuadas para su uso (Landeta, 1999; cap. 5).

Es más, se puede afirmar que esta técnica recobra vigencia y utilidad, ya que sus características definitorias (Dalkey y Helmer, 1963) lo hacen especialmente adecuado para la situación actual. En efecto, el hecho de poder contar con el conocimiento, información e intuición de un número elevado de personas o expertos, interactuando de una manera estructurada y sistemática en lugares geográficamente distantes y en momentos temporales no exactamente coincidentes, con el fin de afrontar procesos de previsión o de decisión en situaciones de incertidumbre, es ahora aún más necesario que hace décadas.

La velocidad de los cambios, que hace que en muchas ocasiones no se den las condiciones necesarias para aplicar otras técnicas de fundamento científico más sólido, y el espectacular desarrollo y difusión de las tecnologías de la información (TI), que ha acelerado y simplificado el proceso de intercambio de información y ha permitido contar

con la participación de mejores expertos, independientemente de su radicación en cada momento, son dos elementos adicionales que justifican la actualidad[1] de esta técnica.

Sin embargo, ciertas debilidades metodológicas que presenta, como son entre otras, la limitada información que se puede intercambiar y, sobre todo, la duración prolongada de un ejercicio Delphi, y la consiguiente dificultad para mantener la motivación y la fidelidad del panel de expertos[2], han hecho que en la práctica el número de aplicaciones sea relativamente reducido, y circunscrito casi exclusivamente al mundo académico y en especial, al referido a las ciencias de la salud, siendo su aplicación real en la empresa muy escasa.

Por esta razón, el desarrollo de las TI puede abrir una nueva vía de reactivación en el empleo de esta técnica, sobre todo en el ámbito empresarial, al acortar notablemente la duración de la aplicación y facilitar el intercambio de información y feedback entre los participantes.

1.3. Finalidad

Partiendo de los antecedentes presentados, este artículo tiene como finalidad verificar las tres proposiciones siguientes:

- Las técnicas grupales de captura y proceso de información subjetiva pueden ser utilizadas favorablemente en la alimentación de modelos estadísticos que habitualmente se nutren casi exclusivamente de datos «objetivos».

Este artículo transmite una experiencia positiva del empleo de una técnica basada en información subjetiva (el método Delphi). Esta experiencia ha puesto de manifiesto, por

[1]Tomando como referencia la base de datos *Dissertation Abstracts*, Mac Spirs 2.3, desde 1997 se han realizado al menos 152 tesis doctorales que han empleado el método Delphi. En el ámbito nacional la base de datos Teseo recoge desde 1987, 22 tesis doctorales que han empleado este mismo método. Así mismo, numerosos artículos de diferentes disciplinas científicas dan cuenta de la utilización de esta técnica. La base de datos *Abi Inform Global Ed.* recoge desde 1998 14 artículos publicados en diversas revistas económicas del mundo. La base de datos *Psycinfo* muestra 29 artículos de psicología. Y Medline, base de datos de revistas de medicina, recoge 168 artículos relacionados con las ciencias de la salud que han utilizado el método Delphi. Este empleo continuo en los círculos científicos manifiesta indirectamente la permanencia de su vigencia social y metodológica. Aunque, su aplicación al ámbito empresarial sigue siendo aún muy escasa.

[2]M. J. Bardecki (1984; p. 288) afirmaba que, a vista de los estudios publicados hasta esa fecha, el porcentaje medio de abandonos por parte de los expertos una vez iniciado el proceso oscilaba entre el 20 y el 30%. Hay que tener en cuenta que este porcentaje está extraído a partir de los trabajos publicados, y puesto que sólo se publican las experiencias exitosas, el porcentaje real de abandonos probablemente será mayor. A modo de ejemplo de estudio en el que el índice de abandono se ha situado en estos niveles puede consultarse Arregui, Villarreal y otros, (1996).

tanto, un campo de aplicación de la metodología Delphi poco usual[3], fuera de las más habituales orientaciones prospectivas y normativas[4], que es el de la «colaboración» con las técnicas tradicionales de obtención de información para la decisión, y que presenta unas posibilidades de desarrollo muy grandes en unos entornos cada vez más inciertos en los que la disposición y fiabilidad de datos objetivos presenta dificultades a menudo insalvables.

- Las TI pueden acortar la duración de un proceso Delphi, contribuyendo a incrementar simultáneamente su calidad.

Esta hipótesis, planteada ya hace muchos años (Turoff, 1971 y 1972), no ha sido suficientemente contrastada en entornos próximos al nuestro, o al menos son escasos los resultados de aplicaciones Delphi con TI que hayan sido dados a conocer[5]. La razón principal que explica esta falta de estudios publicados es que la difusión de las TI no había alcanzado todavía la suficiente amplitud como para permitir plantear un proceso interactivo en círculos no académicos que pudiera ser seguido telemáticamente por la totalidad de los expertos implicados. Por esta razón el «diálogo postal» ha seguido siendo el vehículo de comunicación casi exclusivo para los procesos Delphi que se llevan a cabo.

En la experiencia que se presenta, la utilización del correo electrónico para desarrollar el proceso de interacción con los expertos permitió una mayor rapidez y control de los resultados intermedios y finales, y una participación efectiva de los expertos seleccionados. Se superaron de esta forma los habituales problemas que surgen en este tipo de investigaciones, derivados de las distancias y emplazamientos deslocalizados de los componentes del panel de expertos y de la lentitud de los procesos de obtención y tratamiento de la información. Como consecuencia de todo ello, la calidad de los resultados finales fue destacable.

- El método Delphi es una técnica de obtención de información que cobra mayor utilidad y relevancia cuando es utilizado en entornos inciertos y con agentes deslocalizados.

Es una proposición derivada de las dos anteriores. Como ya hemos afirmado, en este contexto las alternativas para la obtención de información válida disminuyen y esta técnica adquiere un valor relativo mayor.

---

[3]No hemos encontrado publicaciones recientes que revelen aplicaciones similares a la que damos a conocer en este trabajo, aunque Preble (1983) presenta algunos estudios que sí comparten la misma filosofía de aplicación.

[4]Como aplicación de la metodología Delphi con orientación prospectiva puede consultarse Arregi, Villarreal y otros (1996) Y con orientación normativa puede consultarse Landeta, Matey y otros (1992).

[5]La revisión bibliográfica efectuada nos ha mostrado una aplicación realizada utilizando una web como instrumento de transmisión de información y de comunicación (foros interactivos) (Asociación Española de Ingenieros de Comunicación).

En este artículo presentamos en primer lugar, brevemente, el trabajo empírico referido, indicando su origen y los motivos de su realización. A continuación definimos el objeto y la finalidad del estudio para describir y justificar seguidamente la metodología Delphi empleada. La descripción de los agentes participantes y las fases o rondas seguidas permitirá interpretar adecuadamente el desarrollo del trabajo. A continuación mostramos los resultados obtenidos, valorando su calidad y limitaciones. Finalmente presentaremos las conclusiones relacionadas con los objetivos referidos en esta introducción.

2. PRESENTACIÓN DEL TRABAJO EMPÍRICO

El trabajo de investigación ha sido realizado por los profesores del Instituto de Economía Aplicada a la Empresa (IEAE) de la Universidad del País Vasco/Euskal Herriko Unibertsitatea (UPV/EHU) firmantes de esta ponencia, contando con la colaboración de un equipo de técnicos del Institut d'Estadística de Catalunya (IDESCAT) y del Departament d'Indústria, Comerç i Turisme de la Generalitat de Catalunya coordinados por Josep Maria Bas y Enric Pareta.

El origen de este estudio radica en la propuesta de colaboración que nos fue formulada por el IDESCAT, en virtud de nuestro conocimiento sobre la técnica Delphi, para valorar la viabilidad de aplicar la referida técnica en la estimación del gasto mínimo en el que incurren los turistas que visitan Catalunya y, en caso de ser factible, llevar a cabo el estudio correspondiente. Después de una fase de análisis por parte del equipo del IEAE de la información adicional enviada por el IDESCAT, de varias conversaciones telefónicas y de una visita a Barcelona en la que se reflexionó en grupo sobre la aplicabilidad del Delphi como técnica útil para obtener la información que necesitaba el instituto catalán, la propuesta se formalizó en un contrato de investigación entre la Universidad del País Vasco y el IDESCAT.

Según la información que nos ha sido facilitada por el IDESCAT, este estudio es consecuencia derivada de una iniciativa conjunta llevada a cabo desde 1997 por el Departament d'Indústria, Comerç i Turisme de la Generalitat de Catalunya y el Institut d'Estadística de Catalunya dentro del ámbito de la producción de estadísticas de turismo, destacando en la misma la realización periódica de dos encuestas telefónicas sobre la demanda turística, referidas a los viajes de los españoles a Catalunya y de los viajes de los catalanes dentro y fuera de su Comunidad.

Ambas encuestas aparecen incluidas en la Ley del Plan de Estadística de Catalunya y siguen unas pautas semejantes en lo relativo al ámbito de investigación, definición de las variables estadísticas básicas, formato de los cuestionarios, diseño de las muestras, selección de los entrevistados, operaciones de campo, tabulación, análisis de los resultados, etc.

La información obtenida puede ser agrupada en cinco grandes apartados: nivel de actividad (número de turistas, viajes y pernoctaciones), ámbitos territoriales de procedencia y destino, características de los viajes (motivos, organización, alojamiento, transporte, grado de satisfacción, etc.), características de los viajeros (sexo, edad, estado civil, nivel de estudios y profesión) y gasto turístico (manutención, alojamiento, transporte y otros gastos).

De estos cinco apartados, es el quinto, el del gasto turístico, el único que presenta ciertas dificultades en la obtención de datos, ya que la proporción de no respuestas o de respuestas no fiables es notablemente mayor que en las otras categorías. Las causas pueden ser atribuidas a un posible desconocimiento del tema por parte del entrevistado, a que el gasto es un dato de difícil memorización, a que la encuesta telefónica exige una respuesta inmediata, difícil en una cuestión que implica cálculos numéricos, o simplemente a una negativa más o menos explícita a facilitar información de contenido económico.

Una vía posible contemplada por los responsables del IDESCAT para superar esta falta de información en el apartado de gasto turístico, ha sido la confección de un modelo de imputación de un gasto mínimo en aquellos viajes en los que los datos de este apartado económico fueran inexistentes o no fiables.

Después de analizar diversas opciones[6], los responsables del IDESCAT consideraron que el Método Delphi podría ser una técnica válida para obtener la información necesaria con la que alimentar el modelo de imputación de gasto mínimo que habían concebido, como solución a su problema de falta habitual de datos económicos en las encuestas turísticas.

Este trabajo ha iniciado, por tanto, una interesante y novedosa experiencia científica de colaboración entre la Universidad y la Administración de dos Comunidades distintas pero afines, y también de colaboración entre la Administración y los ciudadanos implicados en la realidad turística analizada, dado el protagonismo activo que los mismos han tenido en la concepción y desarrollo de este trabajo Delphi.

3. OBJETO Y FINALIDAD DEL ESTUDIO

Este trabajo estudia el gasto turístico mínimo de los visitantes a Catalunya provenientes del resto de las comunidades autónomas del estado español. El gasto turístico por per-

[6]Con anterioridad al empleo del Método Delphi en la estimación del gasto mínimo, el IDESCAT había recurrido para este fin a técnicas econométricas avanzadas, en colaboración con otra universidad catalana, pero los resultados no habían sido lo suficientemente satisfactorios.

sona es analizado a partir de su descomposición en cuatro fuentes de gasto: alojamiento, manutención, transporte y otros gastos (de bolsillo).

La finalidad del estudio ha sido la obtención de unas estimaciones fidedignas del gasto mínimo en el que incurren los turistas que visitan Catalunya, mediante la interacción de un conjunto de expertos del ámbito turístico empleando una metodología Delphi. No se ha pretendido obtener todas las estimaciones que condujeran a una cuantificación del gasto mínimo individual del visitante, sino sólo aquéllas que no puedan ser conseguidas a partir de los datos disponibles o de fuentes más objetivas.

El fin último de las estimaciones proporcionadas por este trabajo es completar la alimentación de un modelo de imputación de gasto turístico mínimo[7] que pueda ser utilizado por el IDESCAT en caso de ausencia de respuesta fiable a esta cuestión en las encuestas periódicas que elabora.

4. METODOLOGÍA; JUSTIFICACIÓN DE LA ELECCIÓN DE LA TÉCNICA

Como ya ha sido indicado, el tipo de información que requería el IDESCAT y el Departament d'Indústria, Comerç i Turisme de la Generalitat de Catalunya a menudo no podía ser obtenida directamente a partir de las encuestas realizadas a los usuarios. Era necesario, por tanto, acudir a fuentes indirectas.

La estimación de algunos componentes del gasto turístico podían realizarse favorablemente utilizando información objetiva disponible, como podía ser el caso del gasto de transporte en automóvil o en otros medios de transporte con tarifas conocidas y estables. Sin embargo, para la mayoría de los componentes del gasto turístico no existían datos objetivos fiables que pudieran ser utilizados con o sin transformación previa.

La información de más calidad sobre estos aspectos integrantes del gasto turístico, a priori, era la poseída por los expertos del sector, pero a menudo de manera parcial e inconexa.

Era necesario, por tanto, emplear una técnica que obtuviera, procesara, completara e integrara la información que reside en estos expertos, a la vez que los hiciera reflexionar e interactuar, pero siendo conscientes que por las características y radicación

[7]El objetivo del modelo es que cuando un turista no informe satisfactoriamente del gasto en el que ha incurrido en su visita a Catalunya, en virtud de su localidad de origen, del medio de transporte que ha utilizado, de dónde se ha alojado, de cómo se ha mantenido, de la zona que ha visitado (Pirineo, Costa, Interior o Barcelona) y de la temporada en la que ha acudido (datos todos ellos que son obtenidos sin dificultad en el cuestionario telefónico), el Idescat pueda imputar un gasto mínimo lógico que sustituya a su respuesta no válida.

de su actividad no iban a poder trabajar físicamente juntos de una manera habitual o permanente.

Todas estas limitaciones llevaron a los responsables del IDESCAT y a este equipo de investigación a considerar el método Delphi como una alternativa de trabajo potencialmente apropiada para el objeto y condiciones de este estudio.

5. DESARROLLO DEL TRABAJO

5.1. Agentes

- Grupo coordinador

  El equipo coordinador ha estado compuesto por los autores de este artículo y por cuatro técnicos provenientes del IDESCAT y del Departament d'Indústria, Comerç i Turisme de la Generalitat de Catalunya (dos por organismo): Enric Pareta, Monserrat Pérez, Cèlia Gomà y Josep María Bas.

  Los primeros con la función de dirigir todo el proceso: confección del cuestionario y del manual de procedimiento, análisis e interpretación de las respuestas, gestión y control del feedback y confección del informe final de resultados.

  Los segundos con la función de seleccionar el grupo de expertos, realizar el seguimiento logístico de las operaciones, mantener los contactos pertinentes con el grupo de expertos, aportar su conocimiento sobre el sector turístico en la formulación e interpretación de las respuestas y velar por el ajuste del desarrollo del trabajo a las necesidades del IDESCAT.

- Panel de expertos

  La selección de los integrantes del panel de expertos se realizó siguiendo criterios básicos de conocimiento de la oferta y demanda turística catalana, así como de accesibilidad para el grupo coordinador. Otro criterio que fue planteado desde el principio fue el de dar preferencia a aquellos expertos que dispusieran de conexión electrónica real con el exterior.

  Se buscó que en la composición del panel estuvieran representados los diferentes agentes que intervienen o conocen el sector turístico catalán: tour operadores, agencias de viajes, oficinas de turismo, representantes de los sectores hostelero, restauración, apartamentos y campings, viajeros intensivos en Cataluña, especialistas universitarios y consultores turísticos.

  Finalmente fueron seleccionados 14 expertos, una dimensión asumible en términos de trabajo y riesgo, dadas las características precursoras de este trabajo y el riesgo profesional que asumían las entidades promotoras del mismo. Todos los expertos accedieron a participar de forma voluntaria y gratuita en este estudio.

5.2. Fases

- Elaboración del primer cuestionario

  Partiendo del posible modelo de tabla para imputar el gasto mínimo de los viajes que nos fue facilitado por el IDESCAT, y después de varias propuestas y contra-propuestas en el seno del grupo coordinador, llegamos a una categorización de los items (noventa en total) de las diferentes áreas, que pretendía, por un lado, clarificar lo máximo posible los conceptos de gasto manejados (manutención, alojamiento, transporte y otros gastos), y por otro agrupar coherentemente las diferentes alternativas de destino y temporada de los turistas en Cataluña.

  Las preguntas del cuestionario estaban diseñadas de tal forma que las respuestas dadas por los expertos pudieran ser analizables estadísticamente. En este sentido, el grupo coordinador, dado el elevado número de items a responder, optó por solicitar a los expertos solamente un dato cuantitativo por item, descartando otras posibilidades más complejas (intervalos, máximo y mínimo, tripletas de confianza, etc.)

  Así mismo, en cada uno de los cinco grandes apartados del cuestionario (ver anexo) se les requería a los expertos cualquier información o comentario que consideraran relevante sobre las cuestiones planteadas y las respuestas dadas (información cualitativa).

  Debido a la amplia variedad de componentes de gasto que contemplamos en este estudio, y al elevado grado de especialización de algunos de los expertos, el equipo coordinador ofreció a cada experto la posibilidad de autoevaluarse sobre su grado de conocimiento o dominio (elevado, medio o superficial). Esta autovaloración nos ha permitido dar una ponderación diferente en los resultados grupales a las opiniones de los expertos según su grado de conocimiento manifestado.

  El pre-test del cuestionario fue realizado por los miembros catalanes del equipo coordinador en sus organizaciones de origen.

- Desarrollo de las iteraciones Delphi

  Una vez elaborado el primer cuestionario, el IDESCAT organizó una reunión el día 16 de noviembre de 2000 en Barcelona, a la que asistieron casi todos los panelistas, los miembros del grupo coordinador catalanes, Jon Landeta en representación del equipo del IEAE y autoridades del propio IDESCAT y del Departament d'Indústria, Comerç i Turisme de la Generalitat de Catalunya.

  En esta reunión se les presentó el proyecto a los expertos, se les explicó el proceso y la metodología, se disiparon las dudas que les surgieron y se acordó el calendario a cumplir. En el transcurso de la reunión también se les entregó el primer cuestionario. Esta reunión cumplió a la perfección los objetivos que impulsaron su organización: presentar el proyecto y la metodología, aclarar dudas y expectativas y, sobre todo, implicar al grupo de expertos en el desarrollo del proyecto.

El modo de envío de las respuestas por parte de los expertos, de los cuestionarios posteriores y del feedback se efectuó utilizando el correo electrónico como instrumento de comunicación, abandonando el clásico correo postal.

Todos los expertos del panel continuaron hasta el final del estudio.

El feedback o retroalimentación, estructurado por el grupo coordinador, que recibieron los expertos con los cuestionarios de las dos rondas siguientes fue de dos tipos:

- Información estadística: datos estadísticos por item, obtenidos a partir de la distribución de las respuestas (estimación individual en la ronda anterior —dato específico de cada experto—, mediana, cuartil inferior, cuartil superior, mínimo y máximo).
- Información adicional. Este tipo de información, cualitativa o cuantitativa, proveniente en su origen de los propios expertos, ha tenido una doble utilidad. Por un lado, ha permitido especificar mejor la formulación de varias preguntas (a partir del segundo cuestionario), con el fin de garantizar que todos los expertos hayan respondido verdaderamente a la misma cuestión. Y por otro lado, ha hecho posible que los datos objetivos, argumentos y reflexiones realizados hayan sido conocidos por los demás expertos y valorados en sus respectivas respuestas.

El proceso finalizó con la recepción del tercer cuestionario del último de los 14 expertos el 25 de diciembre de 2000.

El grupo coordinador analizó estadísticamente los datos de las respuestas de este tercer cuestionario y a partir de ellos y de la percepción de la realidad estudiada y compartida por coordinadores y panelistas se elaboró el correspondiente informe.

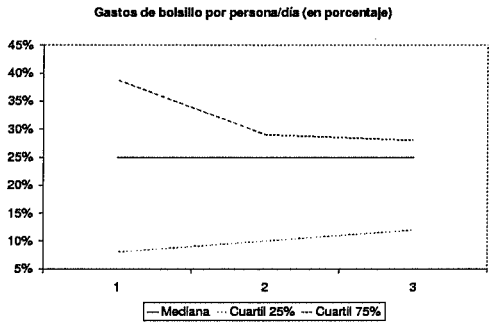
6. RESULTADOS

6.1. Análisis cuantitativo de los resultados

A efectos del tratamiento estadístico, por pregunta entendemos cada una de las cuestiones sobre las que los expertos han tenido que pronunciarse (item). En este sentido, si bien los expertos sólo han tenido que estimar el gasto mínimo en términos de manutención, alojamiento, transporte y otros gastos (4 componentes de gasto), las subdivisiones de estos componentes según zona geográfica y temporada han elevado a 90 el número real de preguntas que los expertos han tenido que responder.

Para cada pregunta e iteración se calcularon indicadores de tendencia central (mediana y media ponderada) y de dispersión/consenso (cuartiles ponderados). También se representaron gráficamente las evoluciones respectivas. Los resultados grupales de la

3ª ronda son presentados en el anexo de forma sintética. Un ejemplo de los resultados cuantitativos y gráficos de una pregunta se recogen en la figura 1.



	Cuest. 1	Cuest. 2	Cuest. 3
Mediana	25.00%	25.00%	25.00%
Cuartil 25%	8.00%	10.00%	12.00%
Cuartil 75%	38.75%	29.00%	28.00%
Media Pond.	23.00%	22.61%	22.92%
Var. RIR		47.00%	12.00%
Var. Res. Exp.		92.86%	50.00%
Consenso	123.00%	76.00%	64.00%

Figura 1. Gasto mínimo en gastos de bolsillo por persona y día (en porcentaje sobre el gasto diario total).

Para reflejar el *grado de estabilidad* alcanzado en cada ronda, y considerando el tamaño del panel y el tipo de respuestas, hemos optado por dos medidores diferentes:

- Variación del recorrido intercuartílico relativo: mide la estabilidad de la respuesta grupal y se obtiene calculando la diferencia entre el recorrido intercuartílico relativo de la respuesta grupal de dos rondas sucesivas.

$$\begin{aligned} \text{Variación RIR}k/k-1 &= \text{RIR}k - RIRk - 1 \\ k &= \text{número ordinal de ronda} \\ \text{RIR} &= \big((Q3 - Q1)/\text{Me}\big) \times 100 \end{aligned}$$

- Proporción de expertos que modifican su estimación: mide la estabilidad de las respuestas individuales. Indica el porcentaje de expertos que modifican su respuesta individual en cada ronda.

Var. Res. Exp. = (Nº de expertos que modifican su respuesta/Nº total de expertos) × 100

Y por su parte, como indicador del *grado de consenso* exhibido por los expertos en sus respuestas a cada pregunta, hemos utilizado el recorrido intercuartílico relativo (RIR).

Debido a que por las condiciones del estudio se había delimitado a priori el número máximo de rondas (tres), no se ha considerado necesario prefijar unos niveles previos de estabilidad o de consenso que fueran a ser utilizados como criterio de finalización del proceso iterativo.

6.2. Valoración de la calidad de los resultados

Un aspecto fundamental en la utilización de una determinada técnica o metodología para la realización de una investigación, como la desarrollada en el caso que nos ocupa, es la calidad de los resultados que a través de aquella se obtienen. Entendemos que la investigación se puede considerar exitosa en la medida en que los resultados que se obtengan de ella sean válidos y fiables.

La validez deberá ser valorada por el IDESCAT, en la medida que estos resultados le sean útiles para alimentar el referido modelo de imputación de gasto. Nos consta que así es.

La fiabilidad, en cambio, está muy relacionada con la forma en la que la investigación ha sido llevada a cabo. Los aspectos que nos permiten valorar positivamente la fiabilidad de los resultados, son los siguientes:

— Estabilidad y calidad del panel (autoevaluación)
— Tiempo transcurrido entre rondas
— Comentarios recogidos de los expertos (información cualitativa)
— Estabilidad de los resultados entre rondas
— Consenso/convergencia de opiniones.

Veamos brevemente cada uno de ellos:

- *Estabilidad y calidad del panel (autoevaluación)*

La calidad del grupo de expertos deriva del escrupuloso proceso de selección seguido y ha sido claramente demostrada por la información cualitativa que se ha manejado durante el proceso. Además, el empleo de las TI ha permitido contar con la colaboración de dos expertos «deslocalizados» durante el periodo de realización del estudio, pues estaban viajando por diversos lugares (uno de ellos por América).

Relacionado con este tema no podemos dejar de hacer referencia a los distintos niveles de conocimiento y dominio que poseían los diferentes expertos en las distintas áreas y aspectos a tratar. La ponderación de las opiniones individuales a partir de sus propias autovaloraciones nos ha permitido optimizar la calidad individual de los expertos, sin por ello perder una visión global e integral de la investigación[8].

[8] Aunque no son muchas las aplicaciones Delphi que emplean diferentes ponderaciones al grado de conocimiento de los expertos, ni los estudios que demuestren concluyentemente la mejora de la calidad derivada de esta forma de proceder, lo cierto es que no tenemos noticia de trabajos que demuestren que incide negativamente en los resultados, pero sí de estudios que parecen mostrar lo contrario: que mejora la precisión de la respuesta grupal: Dalkey (1970), Rowe y Wright (1996).

En cuanto al cumplimiento formal del grupo de expertos, hay un dato absolutamente clarificador y es que la totalidad de los expertos que iniciaron el proceso en la primera ronda lo finalizaron, participando en la tercera y última ronda. Este hecho permite manejar muestras perfectamente comparables, lo cual, adicionalmente cualifica significativamente las comparaciones que se quieran hacer entre las sucesivas rondas. Téngase en cuenta que el abandono de algún experto haría cambiar los resultados grupales, sin necesidad de variación en las respuestas individuales.

- *Tiempo transcurrido entre rondas*

La fase completa de envío y recepción de cuestionarios del estudio se ha desarrollado en un periodo total de cuarenta días para las tres rondas. Es un tiempo excepcionalmente corto que facilita el mantenimiento del interés por parte de los expertos y posibilita habilitar un único escenario temporal para la estimación de los diferentes componentes de gasto. Téngase en cuenta que en un periodo largo se pueden dar circunstancias que hagan variar la realidad a analizar y, por tanto, las estimaciones de los panelistas.

Como ya hemos referido anteriormente, el empleo de las TI ha contribuido a este logro.

- *Comentarios recogidos de los expertos (información cualitativa)*

La incorporación de la información cualitativa al proceso ha sido extensa y relevante, lo cual entendemos que ha sido clave para lograr los niveles de consenso y estabilidad logrados en la tercera ronda. La heterogeneidad y motivación del panel de expertos, y la facilidad y rapidez de la retroalimentación lo han facilitado.

- *Estabilidad de los resultados entre rondas*

La aplicación ortodoxa del método habría requerido prefijar unos niveles mínimos de estabilidad para cada pregunta y sólo después de haberla alcanzado cesar en las iteraciones, pero las condiciones de realización del trabajo lo impedían.

No obstante, los resultados obtenidos parecen bastante aceptables. En cualquier caso, sí podemos constatar que el grado de estabilidad en las respuestas entre las rondas 2 y 3 ha sido muy superior al habido entre las rondas 1 y 2. En un 70% de las respuestas de la última ronda la estabilidad grupal ha sido superior a la de la ronda anterior (es decir, menor variación del RIR) y en un 80% de las mismas, la estabilidad individual ha sido también mayor (menor proporción de expertos que han modificado su respuesta)

- *Consenso/convergencia de opiniones*

En términos generales, nuestra valoración sobre el grado de consenso alcanzado es positiva, en el sentido de que, más del 50% de los items han llegado a consensos que

implican dispersiones inferiores al 20%. Es cierto que casi un 15% de los items presentan dispersiones entre el 40% y el 88 %, pero ello no impide que el conjunto de los resultados muestre un grado de consenso que le confiere una elevada fiabilidad.

Por otra parte, es importante destacar que en 89 de las 90 preguntas el grado de consenso final ha sido superior al exhibido en la primera ronda.

Por consiguiente, a tenor de los valores obtenidos en los medidores seleccionados, estimamos que los resultados finales obtenidos en este estudio poseen un elevado nivel de fiabilidad y aceptabilidad.

6.3. Limitaciones del estudio

A pesar de la valoración positiva de los resultados alcanzados, hemos de reconocer que este estudio presenta ciertas limitaciones, derivadas principalmente de las condiciones en las que se ha tenido que llevar a cabo, aparte de las intrínsecas que tiene el juicio subjetivo como fuente de información y previsión (Landeta 1992 y 1994). A continuación relacionamos las dos más relevantes:

- Número de rondas prefijado. El hecho de contar con unos recursos y plazos limitados condicionó la necesidad de prefijar un número máximo de rondas. Esta limitación posiblemente ha impedido alcanzar un mayor grado de consenso en algunas preguntas en las que se ha observado que todavía no se había alcanzado una estabilidad importante en las respuestas de los expertos[9].
- Número de expertos. El carácter pionero del estudio aconsejaba, por prudencia, dotarlo de unas dimensiones manejables por el equipo coordinador y por la institución contratante. No obstante, si se hubiera podido contar con un número mayor de expertos por cada área y zona de gasto, posiblemente el resultado habría sido más preciso.

Otras limitaciones, a priori más evidentes, como han podido ser las presupuestarias (los expertos no percibían retribución alguna) o temporales (las tres fases se desarrollaron tan sólo en cinco semanas), consideramos que no han tenido influencia significativa en los resultados, ya que han sido compensadas con el interés e ilusión que han depositado en este estudio tanto los expertos colaboradores como el propio equipo coordinador.

[9]Sin embargo, lo cierto es que son pocos los estudios publicados (y menos aún los no publicados) que llevan a cabo más de tres rondas, siendo lo más habitual en los Delphis realizados con fines profesionales una duración de dos rondas.

7. CONCLUSIONES

- El método Delphi, una técnica que este año cumple medio siglo de vida, sigue teniendo vigencia en los ámbitos académicos y profesionales. El recurso al conocimiento de los expertos sigue estando plenamente justificado y el contexto incierto en el que con cada vez mayor frecuencia deben tomarse importantes decisiones acrecienta la necesidad de su concurso.
- La conjunción de planteamientos de estudio y análisis de la realidad que conjuguen tratamiento de datos objetivos con la complementación de los mismos a partir de datos subjetivos obtenidos mediante aplicaciones Delphi u otras técnicas basadas en información subjetiva, se revela como un campo con gran potencial de desarrollo.
- Las nuevas TI proporcionan al método Delphi una importante vía para mejorar sus resultados, en el sentido de que permiten acortar la duración del proceso, con la consiguiente mejora en la calidad de las aportaciones de los expertos, en su mayor motivación y en la reducción del número de abandonos, y, por otra parte, permitiéndoles tomar parte en los procesos Delphi independientemente de dónde estén localizados en cada momento.
- Esta conjunción de necesidad de técnicas de consenso para la actuación coordinada y eficiente de empresas e individuos deslocalizados con la mayor rapidez y facilidad de empleo del Delphi conseguida gracias a las TI, puede abrir una reactivación del uso de esta técnica en ámbitos empresariales.
- La adaptación de esta técnica a las necesidades empresariales, apoyada por las TI, y la combinación eficiente de técnicas cualitativas y cuantitativas para la obtención de información para la decisión, se constituyen en dos áreas abiertas a la investigación que pueden aportar importantes contribuciones al progreso científico y social.

8. BIBLIOGRAFÍA

Arregui Ayastuy, G., Vallejo Alonso, B. y Villarreal Larrínaga, O. (1996). «Aplicación de la metodología Delphi para la previsión de la integración española en la Unión Económica y Monetaria». *Investigaciones Europeas de Dirección y Economía de la Empresa*, 2, 2, 13-38.

Asociación Española de Ingenieros de Comunicación: www.gtic.ssr.upm.es/encuestas/teleco.htm

Bardecki, M. J. (1984). «Participants' response to the Delphi Method: An attitudinal perspective», *Technological Forecasting and Social Change*, 25, 281-292.

Dalkey, N. y Helmer, O. (1963). «An experimental application of the Delphi Method to the use of experts», *Management Science*, 9, 458-467.

Dalkey, N., Brown, B. y Cochran, S. W. (1970). «The Delphi Method IIIuse the self-ratings to improve group estimates». *Technological Forecasting*, 1, 283-291.

Dalkey, N. C. y Helmer, O. (1963). «An experimental application of the Delphi method to the use of experts». *Management Science*, 9, 458-467.

Landeta, J. (1992). *Información subjetiva para la decisión: el método Delphi.* Tesis doctoral. Facultad de Ciencias Económicas y Empresariales de la Universidad del País Vasco. Bilbao.

Landeta, J. (1994). «Previsión basada en información subjetiva: su utilidad y carácter científico». *Cuadernos de Gestión*. Mayo, 89-106.

Landeta, J. (1999). *El método Delphi. Una técnica de previsión para la incertidumbre*. Ed. Ariel. Barcelona.

Landeta, J., Matey, J. y otros (1992). «Evaluación del Plan de Ordenación y Reforma de Ambulatorios. Estudio Delphi». *Le Management des Entreprises dans l'espace Economique Europeen*, 147-172. II Congreso Hispano-Francés de Economía y Dirección de Empresas, Burdeos.

Preble, J. F. (1983). «Public sector use of de Delphi Technique». *Technological Forecasting and Social Change*, 3, 1, 75-88.

Rieger, W. R. (1986). «Directions in Delphi developments: Dissertations and their quality». *Technological Forecasting and Social Change*, 29, 195-204.

Rowe, G. y Wright, G. (1996). «The impact of task characteristics on the performance of structured group forecasting techniques». *International Journal of Forecasting*, 12, 73-89.

Rowe, G. y Wright, G. (1999). «The Delphi technique as a forecasting tool: issues and analysis». *International Journal of Forecasting*, 15, 353-375.

Turoff, M. (1971). «Delphi and its potencial impact on information systems». *Proceedings of Fall Joint Computer Conference*, 317-326.

Turoff, M. (1972). «Delphi conferencing: computer based conferencing with anonimity». *Technological Forecasting and Social Change*, 3, 159-204.

9. ANEXO: TABLA DE LOS RESULTADOS GRUPALES DE LA 3ª RONDA (SÓLO MEDIANAS)

Pregunta 1: MANUTENCIÓN			
	BCN	**Costa A./Pirineo A.**	**Costa B./Pirineo B./Int.**
Apartamento, cámping y casa rural/Elaboración Propia		1.250	1.100
Hotel (sólo alojamiento)/Consumo Externo	2.800	3.000	2.800

Pregunta 2.A: ALOJAMIENTO						
	BCN	**Costa/A**	**Costa/B**	**Pirineo/A**	**Pirineo/B**	**Interior**
5 estrellas	20.500	19.000	15.000	16.000	12.000	(*)
4 estrellas	14.500	10.500	7.500	10.000	7.125	7.000
3 estrellas	8.000	6.500	4.500	7.000	4.500	4.350
Hasta 2 estrellas	3.500	3.500	2.500	3.500	2.500	2.500

Pregunta 2.B: ALOJAMIENTO					
	Costa/A	**Costa/B**	**Pirineo/A**	**Pirineo/B**	**Interior**
Apartamento	2.900	1.600	2.600	1.500	1.500
Cámping	1.250	1.000	1.150	1.050	1.000
Casa rural	2.600	2.000	2.800	2.000	2.000

Pregunta 3: MANUTENCIÓN + ALOJAMIENTO						
	BCN	**Costa/A**	**Costa/B**	**Pirineo/A**	**Pirineo/B**	**Interior**
5 estrellas	27.000	29.000	20.000	22.000	19.000	
4 estrellas	22.000	16.000	11.000	15.300	12.000	11.750
3 estrellas	14.000	9.500	7.500	10.500	8.000	8.000
Hasta 2 estrellas	7.000	6.200	5.000	6.500	5.700	5.000

Pregunta 4: TRANSPORTE					
	600 Km. (Ida y vuelta)	**1.200 Km. (Ida y vuelta)**	**1.800 Km. (Ida y vuelta)**	**Baleares**	**Canarias**
Avión/regular	23.000	26.000	35.000	17.500	33.950
Avión/charter	15.000	18.000	18.500	10.000	19.750
Autocar/regular	5.000	8.000	11.000		
Autocar/discrecional	2.950	5.300	8.000		
Tren	6.400	10.000	12.600		
Barco				7.500	

Pregunta 5: GASTOS DE BOLSILLO		
	En porcentaje	**En cifras absolutas**
Por persona/día	25,00%	
Por persona		25.525

ENGLISH SUMMARY

SUPPLYING OF SUBJECTIVE INFORMATION TO QUANTITATIVE MODELS: THE DELPHI METHOD APPLIED TO A MODEL OF ATTRIBUTION OF THE INDIVIDUAL TOURIST EXPENSE IN CATALUNYA

J. LANDETA RODRÍGUEZ
J. MATEY DE ANTONIO
V. RUÍZ HERRÁN
O. VILLARREAL LARRINAGA
Universidad del País Vasco*

This article shows a study for the Institut d'Estadistica de Catalunya (Idescat), where a method that works with subjective information (Delphi Method) has been used to obtain some data that can feed partially a mathematical model which is mainly provided with statistical data (objective data), in order to increase the global utility of the model. The article justifies the utilization of subjective information under certain circumstances, it describes the Delphi Method, and establishes some proposal with reference to the utility, using and improvement of this method which are approved in the study.

Keywords: Delphi Method, Subjective information, information technologies

AMS Classification (MSC 2000): 93C41

*Universidad del País Vasco/Euskal Herriko Unibertsitatea. Instituto de Economía Aplicada a la Empresa. Facultad de Ciencias Económicas y Empresariales. Avda. Lehendakari Aguirre, 83. 48015 Bilbao

–Received March 2001.
–Accepted January 2002.

In academic and managerial practice is very usual to need the estimation of future (or present) values of variables whose past data are unknown or if they are known, we lack confidence their quality or the maintenance of the conduct patterns in their tendencies noticed before.

The estimated variables in these conditions can be useful by themselves or they can be useful when they joint other variables in a mathematical model that makes up different variables which are supplied with objective or subjective data.

The Delphi Method is a method that works with subjective information, in order to obtain some data that can supply partially a mathematical model which is mainly provided with objective data taken from surveys in order to increase the global utility of the model.

Its principal characteristics are the followings:

- It's a reiterative process. As a minimum, experts must be consulted twice about the same question, so that they can think their answer again by the help of the information that they have got from other experts' opinions.
- It keeps experts anonymity or a least, of their answers, because these answers go straight to the coordinator group.
- Controlled feedback. Information exchange among experts is not free, because it is realized through the coordinator group of the study, in this way we can get rid of all information that is not significant.
- Group's statistic answer. All the opinions take part in the final answer. The questions are framed so that a quantitative and statistical processing of the answers can be made.

Using it involves the creation of a coordinator team of the study made up of a small number of skilful men who are very knowledgeable about the Delphi Method and the subject of study.

This article aim is to test the three following proposals:

- If the group techniques of collecting subjective information can be advantageously used to supply statistical models that are generally provided only with objectives data.
- If the information technologies can reduce the length of Delphi Method helping at the same time to increase its quality.
- If the Delphi Method is a technique of getting information that has a greater utility when it is used in situations of uncertainty and with agents in different spots.

This study has been realized by the professors of Instituto de Economía Aplicada a la Empresa (IEAE) de la Universidad del País Vasco/Euskal Herriko Unibertsitatea

(UPV/EHU) collaborating with the Institut d'Estadística de Catalunya (IDESCAT) and the Departament d'Industria, Comerç i Turisme de la Generalitat de Catalunya and it is the result of a joint initiative that has been carried out since 1.997 by the Departament d'Industria, Comerç i Turisme de la Generalitat de Catalunya and the Institut d'Estadística de Catalunya within the limits of tourism trade statistics production, standing out against the same, the periodic realization of telephonic surveys about the tourist demand, referred to trips of Spaniards to Catalunya and trips of Catalonians inside and outside of their Community.

The kind of information needed by the Institut d'Estadística de Catalunya (IDESCAT) and the Departament d'Industria, Comerç i Turisme de la Generalitat de Catalunya couldn't be often directly obtained by the surveys realized to users. So, it was necessary to seek indirect sources.

Because of this, the study researches into the minimum tourist expense of people who visit Catalunya coming from the rest of Spanish country, by means of the interaction of a group of experts in the tourist field using a Delphi methodology.

The tourist expense per person is analysed separating it into four sources of expense: lodging, maintenance, transport and other expenses.

The reliability of the obtained results will be very related to the way to carry out the research work. The following points will permit us to value positively the results veracity:

- Stability and quality of the panel.
- Lapse of time among rounds.
- Remarks obtained from the experts (qualitative information).
- Stability of the results among rounds.
- Consensus/convergence of opinions.

Despite the positive valuation of the achieved results, we have to admitted that this study has some limitations which mainly stem from working conditions.

Then we'll enumerate the most important limitations:

- The fixed number of rounds: the fact of having limited resources and time limits, determined the necessity of fixing a maximum number of rounds. This limitation has possibly prevent us from reaching a greater consensus in some questions where an important stability in experts' answers hadn't been reached yet.
- The number of experts: the pioneer nature of the study recommended to give it manageable dimensions to the coordinator group and the contracting institution. However, if we had been able to count on a greater number of experts per section and expense area, the result would has been perhaps more accurate.

QÜESTIIÓ, vol. 26, 1-2, p. 197-211, 2002

SMOOTHING THE CATALAN TOURISM MICRO-DATA TIME SERIES*

M. ARTÍS ORTUÑO*
J. L. CARRION I SILVESTRE**
À. COSTA SÁENZ DE SAN PEDRO*
J. SURIÑACH CARALT*

In this paper we propose a method for smoothing the Catalan tourism time series between 1997 and 2000. These time series, built upon a micro database drawn from a survey conducted by the Statistical Institute of Catalonia, are somewhat volatile due, it would seem, to the incomplete nature of the information. The application of a smoothing procedure based on the combination of classical techniques and weighted moving averages allows us to overcome the problems caused by this lack of information and to obtain time series that evolve smoothly over time.

Keywords: Smothing micro-data, tourism time series

AMS Classification (MSC 2000): 62P20

* This paper is a joint undertaking between the Statistical Institute of Catalonia (Idescat) and the Anàlisi Quantitativa Regional Research Group of the University of Barcelona.
* Anàlisi Quantitativa Regional (AQR) Research Group Departament d'Econometria, Estadística i Economia Espanyola Universitat de Barcelona. Av. Diagonal, 690, 08034 Barcelona.
** Institut d'Estadística de Catalunya, Via Laietana, 58, 08003 Barcelona.

– Received October 2001.
– Accepted January 2002.

1. INTRODUCTION

Studies of the tourism sector in Catalonia have traditionally drawn on macro aggregates corresponding to each tourist season, such as private consumption and gross domestic product. In so doing they have tended to rely on one of the main data sources for this sector i.e. the survey of the supply of hotel accommodation in each of the Spanish regions. This survey, undertaken by the Spanish Statistical Institute (INE), provides information about hotel occupancy, in other words, information provided by the supply side of the tourism market.

In 1997, the Statistical Institute of Catalonia (Idescat) introduced a new survey of Catalan tourism, but in contrast to the survey described above this sought to obtain information from the demand side. This survey provides analysts with valuable information about visitors to Catalonia, whose point of origin is one of the other Spanish regions. Given its recent introduction, these statistics offer information for the most recent tourism seasons only and comparative data is only provided with the previous tourism seasons and, consequently, no time series is defined. It should, however, be noted that the time interval of each tourist season has changed since the introduction of the survey which hinders the definition of an appropriate time series. Thus, three tourism seasons were identified for 1997 and 1998: January to May, June to August and September to December; while for 1999 and 2000 four seasons were identified: January to April, May to June, July to August and September to December.

The recent introduction of the survey and the varying time intervals used in defining the tourist season hinder comparisons. Furthermore, although the survey was designed to embrace all the Spanish regions, only a few observations are eventually included within the database, and so the information describing individual characteristics tends to be highly heterogeneous. This heterogeneity becomes even more marked when the data are raised to the entire population.

Consequently using this database to calculate growth rates gives highly volatile time series. Therefore, the aim of this paper is to present a methodology for computing time series from the micro data (the survey) but, in contrast with the original (population-raised) time series, with a smoothed temporal pattern.

It is not, however, our aim to supply the analyst with a specific set of smoothed time series but rather to design a methodology that allows practitioners to obtain smoothed time series automatically whatever the concepts crossed in the database. The successful achievement of this goal depends on the application of simple smoothing methods that can be adequately employed in all cases.

This paper is organised as follows. In Section 2 we describe the database provided by the survey carried out by Idescat. We describe some of the characteristics of this database and define the time series that constitute the focus of this paper. Section 3 outlines

the methodology that is applied in order to smooth these time series. This section contains three sub-sections that offer a detailed description of the transformations involved at each stage of our methodological proposal. Section 4 presents the results obtained. Finally, Section 5 concludes.

2. DESCRIPTION OF THE DATABASE

The availability of a database built upon the conducting of a survey at different points in time allows us to undertake the analysis at different time intervals. As a last resort, the information contained in the database can always be used to define a daily time series. However, problems arise owing to the absence of observations and distortions in the significance of the findings as the time frequency of the analysis increases.

For these two reasons, in this paper, the temporal reference is fixed at monthly intervals and the monthly time series is the basic information to which our methodology is applied. This specification allows us to use classical smoothing techniques including exponential smoothing and Holt-Winters smoothing procedures. In addition, we can obtain time series of varying temporal frequency (quarterly and annual) by aggregating monthly time series.

The micro database used here provides information about individual personal characteristics, including age, profession, marital status and region of residence. It also contains details about their holidays: number and characteristics of the other group members, destination, type of accommodation, amount of expenditure, number of days spent in Catalonia and the number of overnight stays, among others. However, here we focus on just two of these variables: (1) the number of overnight stays and (2) the number of tourists. Both variables are classified by type of accommodation (hotel, family and friends' households, other types of accommodation and total) and by destination (Barcelona, Costa Daurada, other destinations in Catalonia and all destinations in Catalonia). The different combinations give rise to forty time series.

The definition of these time series is strongly conditioned by the quality of the information comprising the tourism micro-database. Thus, firstly, although in aggregating the information we have tried to avoid missing or zero values, this has been unavoidable in certain periods for some time series. This might have a detrimental effect on the quality of the output following the application of the smoothing procedure. Secondly, graphic inspection of the time series indicates that there might be some outliers, the presence of which implies growth rates of doubtful validity. Finally, there would seem to be an Easter Week effect due to the fact it is a moveable feast and as such does not always occur in the same time period.

In this analysis these first two problems with the information are left for future consideration, particularly given that Idescat is planning to modify some of these anomalies. The third problem is discussed below.

3. METHODOLOGICAL PROPOSAL

In this section we present the methodology adopted in smoothing the time series described in the previous section. One of the reasons why these time series are apparently so erratic is that the survey loses precision as the geographical and conceptual range is increased. Our proposal tries to compensate for this absence of observations by increasing the amount of information used when estimating the micro data of one particular month.

The increase in the amount of information is achieved by the joint consideration of the micro-observations referring to the same month in two consecutive years. Thus, we compute the average number of tourists and overnight stays in the same month for two consecutive years and assign this mean value to these months. Hence, we take into account information that refers to two similar periods (month) and, as a consequence, we are able to reduce the volatility of the time series.

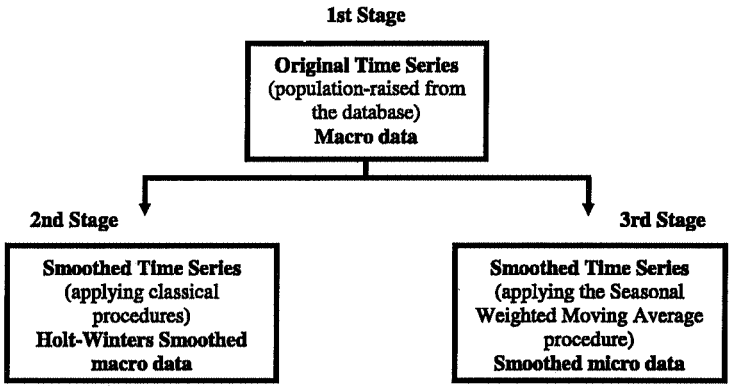


Figure 1. Brief description of the methodological proposal.

This simple method allows us to obtain time series that have a smoother pattern throughout the time period under consideration. The main problem arises, however, when deciding the weightings that should be applied when computing this mean value. One possibility is the specification of equal weights for each time period. Yet, it might be argued that a weighting system that gives greater weighting to more recent information is preferable to a system that attaches the same importance to the two sets of information.

If this is the case, the analyst needs to select these weightings. This is not, however, a straightforward decision, given that different weightings will result in different numbers of tourists and overnight stays. We try to overcome this drawback by suggesting a method by which the weightings can be estimated. The method comprises three stages.

3.1. First stage: The computation of the original time series

The approach relies on the definition of two sets of time series. The first set is the one defined by the original time series, that is, the time series that are derived from raising the data of the survey to the population. As mentioned in the previous section, the forty time series thus obtained are highly volatile over time, which is the problem that this paper seeks to rectify. The large number of observations available for the short period under analysis (forty-eight observations in just four years) means that the application of the stochastic approach to the modelling of these processes is not the most appropriate and that the classical approach should be the one to be adopted.

3.2. Second stage: Definition of the time series of reference

In the second stage of the analysis we obtain the set of forty time series following the application of a classical smoothing procedure to the original time series. This second set of smoothed time series serves as a *referent* for defining the system of weightings to be used when computing the average.

Before applying the classical methodological approach to the modelling of the time series we need to know the type of time series that is being dealt with. Here, the characterisation of the time series was performed using two test statistics. In order to decide the consideration of a time trend we applied the Daniel test, while for the seasonal component we applied the Kruskal-Wallis test. These tests indicated that in most cases the patterns of the time series are given by both components. Nevertheless, it should be noted that these results are not entirely reliable since these statistical tools are more suited to moderate or large sample sizes. Table A.1 in the Appendix shows the results of the application of both tests. The most appropriate smoothing procedure for these data is that of Holt-Winters since there are trend and seasonal components in most of the forty time series. Graphical inspection indicates that the additive model can provide a good fit, although this conclusion might need to be revised as further information comes available.

Before the Holt-Winters smoothing procedure can be applied to the time series under consideration, we need to analyse the effect of Easter Week on these time series. The only period that might have had an influence on the time series was in 1997. In this year Easter fell in the month of March while for the remaining years it fell in April. In order

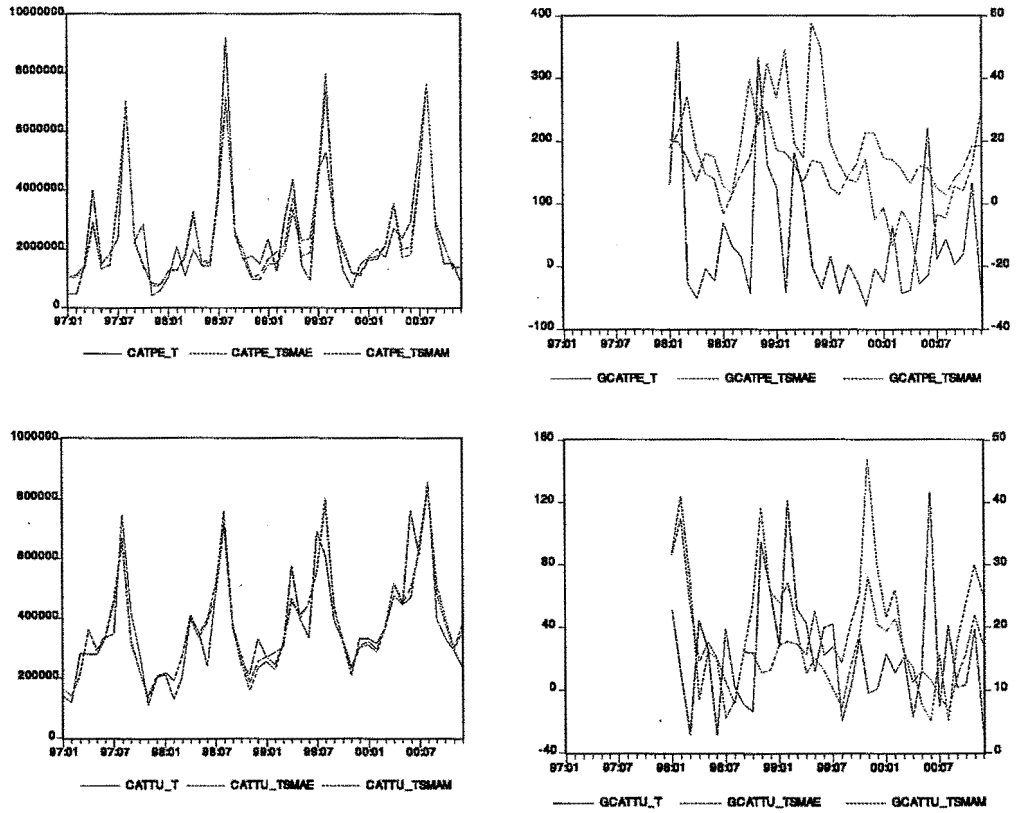
to avoid distortions that might affect the output of the smoothing procedure we decided to test for the presence of a 1997 Easter Week effect and, if there was found to be such an effect to correct the time series to take it into account.

This meant the estimation of a regression model that specifies the time series as a function of an independent term, a time trend, a seasonal dummies set, and an impulse dummy that captures the effect that can be assigned to March 1997. Only in three cases was this impulse dummy found to have a statistical significance of 10%, and the effect was corrected in each case. The three time series were the total numbers of overnight stays in Catalonia (CATPE_T), overnight stays with family or in a friend's household in Catalonia (CATPE_F) and overnight stays with family or in a friends' household using this minimisation criterion in Barcelona (BCNPE_F).

Table 1. Estimated coefficients for the Holt-Winters smoothing procedure.

	Overnight stays			*Tourists*		
	ALFA	BETA	GAMMA	ALFA	BETA	GAMMA
CAT_T	0	0	0	0	0	0
CAT_H	0.02	0.03	0	0.03	0	0
CAT_F	0	0	0	0.01	0	0
CAT_R	0	0	0	0	0	0
CAT_NH	0	0	0	0.01	0	0
BCN_T	0	0	0	0.02	0	0
BCN_H	0.01	0.09	0	0.26	0	0
BCN_F	0	0	0	0	0	0
BCN_R	0	0	0	0.01	0.18	0
BCN_NH	0	0	0	0.01	0.07	0
CD_T	0	0	0	0	0	0
CD_H	0.01	0.08	0	0	0	0
CD_F	0	0	0	0	0	0
CD_R	0	0	0	0	0	0
CD_NH	0	0	0	0	0	0
RD_T	0	0	0	0	0	0
RD_H	0	0	0	0	0	0
RD_F	0	0	0	0	0	0
RD_R	0	0	0	0	0	0
RD_NH	0.1	0	0	0	0	0

Note: CAT_T refers to all types of accommodation used in Catalonia (*CAT*). *CAT_H* indicates those people staying in a hotel. *CAT_F* those staying with family or in a friend's household. *CAT_R* denotes the other types of accommodation used. Finally, *CAT_NH* denotes those staying in accommodation other than a hotel. This notation is repeated for the territorial division considered here: *BCN*-Barcelona, *CD*-Costa Daurada and *RD*-remaining destinations.



Note: CATPE_T denotes the raw time series of the total number of overnight stays in Catalonia, CATPE_TSMAE denotes the smoothed time series using the Holt-Winters procedure with the estimated coefficients and CATPE_TSMAM is the smoothed time series using the Holt-Winters procedure with a fixed value for the coefficients. GCATPE_T, GCATPE_TSMAE and GCATPE_SMAM are the corresponding growth rates. CATTU_T, CATTU_TSMAE and CATTU_TSMAM refer to the tourist series.

Figure 2. Overnight stays and tourists in Catalonia. Levels and growth rates of the original and smoothed time series.

Once the time series affected by the Easter Week had been modified, we applied the Holt-Winters smoothing procedure to estimate the value of the parameters of the independent term (α), the slope (β) and the seasonal parameter (γ) using the criteria of the minimisation of the sum of squared residuals. The estimated coefficients for each time series are presented in Table 1. Note that although it is possible to fix the value of these parameters, the estimation provides a better fit.

As can be seen from Table 1, in most cases the estimated parameters equal zero, indicating that the corresponding component —independent term, trend and seasonality— is stable, that is, it does not vary during the time period under analysis.

The monthly time series for the level and rate of growth for the number of tourists visiting Catalonia and the number of overnight stays following the application of the Holt-Winters smoothing procedure are given in Figure 2. We denote these time series as the smoothed-HW time series. Each figure contains information about the original time series, the smoothed-HW time series with manual selection of the parameters and the smoothed-HW time series with the estimated parameters obtained using the minimisation of the squared sum of errors' criteria.

A number of comments should be made. First of all, it can be seen that the smoothed time series built on the use of the estimated coefficients show a smoother behaviour than those in which the value of such coefficients is imposed (in this case the parameter values were fixed at 0.1). Second, these results indicate that the estimation of the initial values used in obtaining the smoothed-HW time series influences the computation of the growth rates. Thus, for instance, we encounter a contradiction for the time series of tourists coming to Catalonia in which the type of accommodation is not specified (CATTU_T). In this case the growth rates computed using the original time series are negative, while with the smoothed-HW time series they are positive. This is also the case for the time series of tourists coming to Catalonia and staying in other types of accommodation (CATTU_RD) and overnight stays in hotel accommodation in Catalonia (CATPE_T). Third, and in contrast to the smoothed-HW time series, for some time series and periods the original time series present null values which means the growth rates are discontinuous in these cases.

3.3. Third stage: Application of the Seasonal Weighted Moving Average (SWMA) smoothing procedure

In the third stage of the analysis we select the weightings that best fit the time series smoothed in the previous stage. The estimation of these weightings is carried out by specifying the criteria of minimisation of the sum of squared errors, where the errors are given by the difference between the smoothed-HW macro time series and the weighted average time series —hereafter smoothed-micro time series. This section describes the methodology adopted in this optimisation procedure.

Once the macro time series in question has been smoothed using the Holt-Winters procedure, employing an additive specification and by estimating the parameters of the model, we proceeded to select the set of weights used in the procedure applied in this paper to the micro series. This procedure can be understood as the computation of seasonal weighted moving averages (SWMA). For instance, in computing the smoothed time series of tourists for January 1998 using the SWMA procedure we need to take into account the information of the original time series of tourists that corresponds to January 1997 and January 1998. To compute the observation for February 1998 of the smoothed time series we need to look at the observations of the original time series referring to February 1997 and February 1998, and so on.

The important aspect of our proposal is the system of weightings applied in computing the average. As mentioned above, different smoothed time series are obtained depending on the set of weights used. The greater the weight given to more recent values in the time series, the more the smoothed-micro time series tends to resemble the original time series. In the first stage we smoothed the original time series by applying five sets of weights: 50/50, 40/60, 30/70, 20/80 and 10/90. Yet, in order to avoid being subjective when selecting the system of weightings, we estimated this parameter through the minimisation of the sum of squared error of the difference between the smoothed time series using the SWMA procedure —the smoothed-micro time series— and the smoothed-HW time series. This estimation can be outlined as follows.

If we denote the original time series by Y, the smoothed-HW time series by Y^* and the smoothed-micro time series using the SWMA procedure by $\hat{Y}$, the target function to be minimised is the function given by:

$$f(p) = \sum_{i=s+j}^{T} (y_i^* - \hat{y}_i)^2,$$

$j = \{1,2,\ldots,s\}$, where s denotes the order of seasonality —here $s = 12$. The smoothed-micro time series is computed from:

$$\hat{y}_i = p\,y_i + (1-p)\,y_{i-s},$$

where p is the weight (parameter) to be estimated. Therefore, the optimisation program can be represented by:

$$\min \sum_{i=s+j}^{T} (y_i^* - (p\,y_i + (1-p)y_{i-s}))^2$$

$$\text{subject to} \quad 0 \leq p \leq 1.$$

The necessary condition establishes that:

$$\frac{\partial f(p)}{\partial p} = \sum_{i=s+j}^{T} 2\,(y_i^* - p\,y_i - (1-p)\,y_{i-s})\,(-y_i + y_{i-s}) = 0,$$

so that the optimal weights are given by:

$$\hat{p} = \frac{-\sum_{i=s+j}^{T} y_i^* (-y_i + y_{i-s}) + \sum_{i=s+j}^{T} y_{i-s} (-y_i + y_{i-s})}{-\sum_{i=s+j}^{T} y_i (-y_i + y_{i-s}) + \sum_{i=s+j}^{T} y_{i-s} (-y_i + y_{i-s})}. \tag{1}$$

The second derivative indicates that the solution corresponds to a minimum:

$$\frac{\partial^2 f(p)}{\partial p^2} = \sum_{i=s+j}^{T} (-2y_i(-y_i + y_{i-s}) + 2y_{i-s}(-y_i + y_{i-s}))$$

$$= 2\sum_{i=s+j}^{T} (-y_i + y_{i-s})^2 > 0.$$

Table 2. Optimal weight (p) for the SWMA.

	Overnight stays	*Tourists CAT_F 0.53*
CAT_F	0.53	0.53
CAT_H	0.52	0.47
CAT_NH	0.51	0.51
CAT_R	0.49	0.51
CAT_T	0.53	0.57
BCN_F	0.48	0.48
BCN_H	0.53	0.57
BCN_NH	0.50	0.47
BCN_R	0.68	0.48
BCN_T	0.52	0.59
CD_F	0.56	0.51
CD_H	0.51	0.42
CD_NH	0.44	0.46
CD_R	0.43	0.48
CD_T	0.54	0.43
RD_F	0.52	0.49
RD_H	0.47	0.44
RD_NH	0.49	0.50
RD_R	0.49	0.51
RD_T	0.50	0.47

Note: CAT_T refers to all types of accommodation used in Catalonia (CAT). CAT_H indicates those people staying in a hotel. CAT_F those staying with family or in a friend's household. CAT_R denotes the other types of accommodation used. Finally, CAT_NH denotes those staying in accommodation other than a hotel. This notation is repeated for the territorial division considered here: BCN-Barcelona, CD-Costa Daurada and RD-remaining destinations.

As can be shown, the weight can be easily computed from (1) using the original and the smoothed-HW time series. The application of this procedure is undertaken in the next section.

4. RESULTS

The weights computed for the time series in question are shown in Table 2.

These weights were applied after fixing the data corresponding to the year 2000 as the base. Therefore, the values of the original and the smoothed time series using the SWMA procedure for the months of 2000 are the same. The decision to choose this year as the point of reference was made following advice received from experts at Idescat. Tables A.2 and A.3 present the original (Y), the smoothed-HW time series (Y^*) and smoothed-micro quarterly time series ($\hat{Y}$) and their growth rate, for the time series of overnight stays and tourists.

A comparison of the original and the smoothed time series (SWMA) reveals certain discrepancies. These differences are presented in Table 3 for Catalonia.

Table 3. Overnight stays and tourists in Catalonia. Original and smoothed time series.

	Overnight stays	*Overnight stays (smoothed)*	*Tourist*	*Tourist (smoothed)*
Annual time series				
1997	25,148,935	27,583,868	3,538,870	3,773,680
1998	29,746,499	29,491,193	3,947,711	4,388,195
1999	30,991,276	32,897,103	5,045,885	5,451,129
2000	32,862,961	32,862,961	5,420,252	5,420,252
Differences				
1997		−2,434,933		−234,811
1998		255,306		−440,484
1999		−1,905,827		−405,244
2000		0		0

Although these differences are not great, they can be proportionally distributed within each period in order to obtain the same number of overnight stays and tourists for the original and smoothed time series.

Table A.1. Daniel and Kruskal-Wallis Tests.

	Test of Daniel	*Probability*	*Test of Krusal-Wallis*	*Probability*		
CATPE_F	1.56	0.94	20.73	0.96	Trend	Seasonality
CATPE_H	2.22	0.99	30.42	1.00	Trend	Seasonality
CATPE_NH	1.01	0.84	25.46	0.99	Trend	Seasonality
CATPE_R	1.08	0.86	30.25	1.00	Trend	Seasonality
CATPE_T	1.64	0.95	26.70	0.99	Trend	Seasonality
CATTU_F	3.16	1.00	19.09	0.94	Trend	Seasonality
CATTU_H	3.07	1.00	22.34	0.98	Trend	Seasonality
CATTU_NH	2.79	1.00	29.13	1.00	Trend	Seasonality
CATTU_R	1.50	0.93	28.67	1.00	Trend	Seasonality
CATTU_T	3.22	1.00	29.45	1.00	Trend	Seasonality
BCNPE_F	1.47	0.93	20.69	0.96	Trend	Seasonality
BCNPE_H	1.88	0.97	19.27	0.94	Trend	Seasonality
BCNPE_NH	0.65	0.74	18.17	0.92	Trend	Seasonality
BCNPE_R	−0.59	0.28	17.72	0.91	Trend	Seasonality
BCNPE_T	1.55	0.94	21.04	0.97	Trend	Seasonality
BCNTU_F	2.89	1.00	17.30	0.90	Trend	Seasonality
BCNTU_H	3.08	1.00	9.88	0.46	Trend	Seasonality
BCNTU_NH	2.85	1.00	18.88	0.94	Trend	Seasonality
BCNTU_R	−0.23	0.41	11.48	0.60	Trend	Seasonality
BCNTU_T	4.00	1.00	13.00	0.71	Trend	Seasonality
CDPE_F	0.64	0.74	25.35	0.99	Trend	Seasonality
CDPE_H	2.59	1.00	29.41	1.00	Trend	Seasonality
CDPE_NH	0.94	0.83	32.58	1.00	Trend	Seasonality
CDPE_R	0.82	0.79	33.99	1.00	Trend	Seasonality
CDPE_T	1.72	0.96	33.89	1.00	Trend	Seasonality
CDTU_F	0.50	0.69	22.23	0.98	Trend	Seasonality
CDTU_H	2.30	0.99	33.53	1.00	Trend	Seasonality
CDTU_NH	0.78	0.78	35.50	1.00	Trend	Seasonality
CDTU_R	0.99	0.84	34.91	1.00	Trend	Seasonality
CDTU_T	1.87	0.97	36.42	1.00	Trend	Seasonality
RDPE_F	1.34	0.91	13.67	0.75	Trend	Seasonality
RDPE_H	1.49	0.93	24.90	0.99	Trend	Seasonality
RDPE_NH	1.63	0.95	21.46	0.97	Trend	Seasonality
RDPE_R	1.61	0.95	19.01	0.94	Trend	Seasonality
RDPE_T	1.85	0.97	22.12	0.98	Trend	Seasonality
RDTU_F	2.41	0.99	11.90	0.63	Trend	Seasonality
RDTU_H	2.20	0.99	23.06	0.98	Trend	Seasonality
RDTU_NH	2.81	1.00	17.61	0.91	Trend	Seasonality
RDTU_R	1.75	0.96	19.26	0.94	Trend	Seasonality
RDTU_T	2.77	1.00	23.91	0.99	Trend	Seasonality

Note: CAT_T refers to all types of accommodation used in Catalonia (CAT). CAT_H indicates those people staying in a hotel. CAT_F those staying with family or in a friend's household. CAT_R denotes the other types of accommodation used. Finally, CAT_NH denotes those staying in accommodation other than a hotel. This notation is repeated for the territorial division considered here: BCN-Barcelona, CD-Costa Daurada and RD-remaining destinations.

Table A.2. Overnight stays in Catalonia Level and growth rate of the quarterly time series.

	Total	*Total (SWMA)*	*Total (HW)*	*Hotel*	*Hotel (SWMA)*	*Hotel (HW)*	*Family*	*Family (SWMA)*	*Family (HW)*	*Other*	*Other (SWMA)*	*Other (HW)*	*No Hotel*	*No Hotel (SWMA)*	*No Hotel (HW)*
Jan-April97	6,371,284	6,258,097	6,532,288	1,348,147	1,211,671	1,218,876	4,250,680	4,378,042	4,100,855	688,708	636,305	1,199,149	5,006,671	5,040,412	5,316,825
May-Aug97	12,695,217	14,497,545	13,429,469	3,077,329	3,712,300	3,532,048	4,367,604	5,230,961	4,705,140	5,304,046	5,547,960	5,213,216	9,616,045	10,719,327	9,904,455
Sep-Dec97	6,082,434	6,828,226	5,187,177	1,716,735	1,887,964	1,395,263	2,272,053	3,242,448	2,084,342	2,123,632	1,737,762	1,704,021	4,384,007	4,920,523	3,785,444
Jan-April98	6,157,567	8,684,303	7,389,402	1,084,306	1,404,618	1,722,879	4,490,759	5,481,308	4,471,478	582,502	1,709,680	1,157,341	5,073,261	7,176,905	5,645,640
May-Aug98	16,098,317	14,079,534	14,286,583	4,304,880	3,854,960	3,996,419	5,995,048	5,512,106	5,075,763	5,798,388	4,795,953	5,171,408	11,793,437	10,302,730	10,233,270
Sep-Dec98	7,490,615	6,727,355	6,044,291	2,047,762	1,908,446	1,900,333	4,101,267	3,308,045	2,454,965	1,341,586	1,501,159	1,662,214	5,442,853	4,849,087	4,114,259
Jan-April99	10,928,471	9,455,646	8,246,516	1,703,545	2,365,957	2,221,540	6,357,964	5,478,901	4,842,101	2,866,962	1,680,558	1,115,534	9,224,926	7,166,946	5,974,455
May-Aug99	12,286,515	15,372,581	15,143,698	3,435,076	4,531,996	4,467,985	5,084,693	5,978,805	5,446,386	3,766,746	4,763,921	5,129,600	8,851,439	10,729,468	10,562,085
Sep-Dec99	6,049,452	6,342,039	6,901,405	1,778,431	2,077,740	2,258,853	2,606,028	2,685,014	2,825,588	1,664,993	1,578,610	1,620,406	4,271,021	4,257,754	4,443,075
Jan-April00	8,147,529	8,147,529	9,103,630	2,984,146	2,984,146	2,568,869	4,700,913	4,700,913	5,212,724	462,470	462,470	1,073,726	5,163,382	5,163,382	6,303,270
May-Aug00	18,113,527	18,113,527	16,000,812	5,555,687	5,555,687	4,912,999	6,770,112	6,770,112	5,817,009	5,787,728	5,787,728	5,087,792	12,557,840	12,557,840	10,890,900
Sep-Dec00	6,601,905	6,601,905	7,758,519	2,357,068	2,357,068	2,807,356	2,754,919	2,754,919	3,196,211	1,489,919	1,489,919	1,578,598	4,244,838	4,244,838	4,771,890
							INTER-ANNUAL GROWTH RATES								
Jan-April98	−3.35%	38.77%	13.12%	−19.57%	15.92%	41.35%	5.65%	25.20%	9.04%	−15.42%	168.69%	−3.49%	1.33%	42.39%	6.18%
May-Aug98	26.81%	−2.88%	6.38%	39.89%	3.84%	13.15%	37.26%	5.37%	7.88%	9.32%	−13.55%	−0.80%	22.64%	−3.89%	3.32%
Sep-Dec98	23.15%	−1.48%	16.52%	19.28%	1.08%	36.20%	80.51%	2.02%	17.78%	−36.83%	−13.62%	−2.45%	24.15%	−1.45%	8.69%
Jan-April99	77.48%	8.88%	11.60%	57.11%	68.44%	28.94%	41.58%	−0.04%	8.29%	392.18%	−1.70%	−3.61%	81.83%	−0.14%	5.82%
May-Aug99	−23.68%	9.18%	6.00%	−20.21%	17.56%	11.80%	−15.19%	8.47%	7.30%	−35.04%	−0.67%	−0.81%	−24.95%	4.14%	3.21%
Sep-Dec99	−19.24%	−5.73%	14.18%	−13.15%	8.87%	18.87%	−36.46%	−18.83%	15.10%	24.11%	5.16%	−2.52%	−21.53%	−12.19%	7.99%
Jan-April00	−25.45%	−13.83%	10.39%	75.17%	26.13%	15.63%	−26.06%	−14.20%	7.65%	−83.87%	−72.48%	−3.75%	−44.03%	−27.96%	5.50%
May-Aug00	47.43%	17.83%	5.66%	61.73%	22.59%	9.96%	33.15%	13.24%	6.80%	53.65%	21.49%	−0.82%	41.87%	17.04%	3.11%
Sep-Dec00	9.13%	4.10%	12.42%	32.54%	13.44%	24.28%	5.71%	2.60%	13.12%	−10.52%	−5.62%	−2.58%	−0.61%	−0.30%	7.40%

Note: Total indicates all types of accommodation. Hotel indicates those people staying in a hotel. Family denotes those staying with family or in a friend's household. Other denotes all other types of accommodation. Finally, No hotel denotes those that do not use hotel accommodation. (SWMA) indicates that the time series has been smoothed using the SWMA procedure whereas (HW) indicates that the time series has been smoothed using the Holt-Winters procedure.

Table A.3. Tourists in Catalonia Level and growth rate of the quarterly time series.

	Total	*Total (SWMA)*	*Total (HW)*	*Hotel*	*Hotel (SWMA)*	*Hotel (HW)*	*Family*	*Family (SWMA)*	*Family (HW)*	*Other*	*Other (SWMA)*	*Other (HW)*	*No Hotel*	*No Hotel (SWMA)*	*No Hotel (HW)*
Jan-April97	817,174	889,129	882,127	305,527	306,255	336,181	361,723	448,344	390,952	145,427	127,831	152,941	508,722	573,268	544,303
May-Aug97	1,707,462	1,780,283	1,770,938	624,757	639,031	683,393	532,890	599,846	593,934	546,361	533,261	486,036	1,074,600	1,127,987	1,078,180
Sep-Dec97	1,014,234	1,104,269	885,804	479,450	458,087	383,322	428,460	484,085	334,009	114,275	159,325	167,086	545,814	643,047	503,136
Jan-April98	942,458	1,224,179	1,091,170	307,064	358,904	447,206	524,798	623,715	481,041	110,596	209,049	157,426	635,394	829,592	637,263
May-Aug98	1,834,254	1,944,690	1,979,981	654,881	706,447	782,052	658,944	770,472	684,604	520,429	456,523	490,521	1,179,373	1,221,901	1,172,151
Sep-Dec98	1,170,999	1,219,326	1,094,847	434,364	495,208	467,121	533,182	490,147	424,160	203,453	222,022	171,571	736,635	714,082	597,347
Jan-April99	1,432,977	1,466,439	1,300,213	416,467	586,352	529,390	711,022	651,277	572,472	305,488	210,624	161,911	1,016,510	863,509	737,902
May-Aug99	2,026,540	2,396,579	2,189,023	763,706	954,751	864,356	868,911	913,612	784,305	393,923	473,211	495,006	1,262,834	1,385,692	1,284,163
Sep-Dec99	1,255,144	1,256,887	1,303,889	562,770	543,022	556,159	452,162	509,035	524,688	240,212	208,709	176,056	692,374	715,156	703,535
Jan-April00	1,491,239	1,491,239	1,509,255	774,994	774,994	634,810	598,545	598,545	664,605	117,700	117,700	166,396	716,245	716,245	835,520
May-Aug00	2,670,835	2,670,835	2,398,066	1,166,891	1,166,891	1,007,511	953,067	953,067	867,382	550,877	550,877	499,491	1,503,944	1,503,944	1,369,178
Sep-Dec00	1,258,178	1,258,178	1,512,932	521,094	521,094	703,585	559,234	559,234	609,888	177,851	177,851	180,541	737,084	737,084	795,424
							INTER-ANNUAL GROWTH RATES								
Jan-April98	15.33%	37.68%	23.70%	0.50%	17.19%	33.03%	45.08%	39.12%	23.04%	−23.95%	63.53%	2.93%	24.90%	44.71%	17.08%
May-Aug98	7.43%	9.23%	11.80%	4.82%	10.55%	14.44%	23.65%	28.45%	15.27%	−4.75%	−14.39%	0.92%	9.75%	8.33%	8.72%
Sep-Dec98	15.46%	10.42%	23.60%	−9.40%	8.10%	21.86%	24.44%	1.25%	26.99%	78.04%	39.35%	2.68%	34.96%	11.05%	18.72%
Jan-April99	52.05%	19.79%	19.16%	35.63%	63.37%	18.38%	35.48%	4.42%	19.01%	176.22%	0.75%	2.85%	59.98%	4.09%	15.79%
May-Aug99	10.48%	23.24%	10.56%	16.62%	35.15%	10.52%	31.86%	18.58%	14.56%	−24.31%	3.66%	0.91%	7.08%	13.40%	9.56%
Sep-Dec99	7.19%	3.08%	19.09%	29.56%	9.66%	19.06%	−15.20%	3.85%	23.70%	18.07%	−6.00%	2.61%	−6.01%	0.15%	17.78%
Jan-April00	4.07%	1.69%	16.08%	86.09%	32.17%	19.91%	−15.82%	−8.10%	16.09%	−61.47%	−44.12%	2.77%	−29.54%	−17.05%	13.23%
May-Aug00	31.79%	11.44%	9.55%	52.79%	22.22%	16.56%	9.69%	4.32%	10.59%	39.84%	16.41%	0.91%	19.09%	8.53%	6.62%
Sep-Dec00	0.24%	0.10%	16.03%	−7.41%	−4.04%	26.51%	23.68%	9.86%	16.24%	−25.96%	−14.79%	2.55%	6.46%	3.07%	13.06%

Note: Total indicates all types of accommodation. Hotel indicates those people staying in a hotel. Family denotes those staying with family or in a friend's household. Other denotes all other types of accommodation. Finally, No hotel denotes those that do not use hotel accommodation. (SWMA) indicates that the time series has been smoothed using the SWMA procedure whereas (HW) indicates that the time series has been smoothed using the Holt-Winters procedure.

5. CONCLUSIONS

In this paper we propose a new approach to smoothing the Catalan tourism time series. In short, this procedure increases the information for one tourism season by taking into account information obtained by survey for that specific season and information referring to the same season but in the previous year. The application of the Seasonal Weighted Moving Average (SWMA) smoothing method means that, when determining the micro-data for one specific time series, we need to take into account the information for two similar time periods. As a result of this, the smoothed time series are smoother than the original series. The application of this procedure to the Catalan tourism micro-data time series reduces the volatility present in the original time series. As Costa *et al.* (2001) demonstrate elsewhere in this issue, it allows researchers to analyse the evolution between seasons without the distortions caused by an absence of information.

The main advantage of this approach is that once the weighting system has been defined, it can be applied to any micro-data time series that can be built from the database. Moreover, our proposal is not limited to just this database as it can also be applied to other micro databases that contain information of considerable interest to economists. These include the quarterly Active Population Survey (EPA) of the Spanish and Catalan economies, the monthly Overnight Stays in Hotels Survey that is conducted among the Hotels of the different Spanish regions and the quarterly Household Budget Continuous Survey, to name just a few.

QÜESTIIÓ, vol. 26, 1-2, p. 213-243, 2002

ESTIMADORES COMPUESTOS EN ESTADÍSTICA REGIONAL: UNA APLICACIÓN A LA ESTIMACIÓN DE LA TASA DE VARIACIÓN DE LA OCUPACIÓN EN LA INDUSTRIA*

À. COSTA[1]
A. SATORRA[2]
E. VENTURA[2]

Este trabajo es parte de un proyecto que estudia la aplicación de estimadores compuestos (combinación de estimadores directos e indirectos) para áreas pequeñas en estadística regional. Comparamos tres estimadores: uno directo basado en datos muestrales de cada Comunidad Autónoma (CA), otro sintético (indirecto) que combina los datos estatales con información específica de las CCAA, y un tercer estimador, el compuesto, basado en un modelo estadístico que se concreta en una combinación lineal de los dos anteriores. Ilustramos el método de análisis adoptado mediante la estimación de la tasa de variación de la ocupación industrial en las diferentes CCAA de España.

Composite estimators in regional statistics: an application to the estimation of the rate of change in industrial occupation

Palabras clave: Estadística regional, áreas pequeñas, estimadores compuestos

Clasificación AMS (MSC 2000): 62J07, 62J10, 62H12

*Este trabajo es fruto de un proyecto de investigación conjunto entre el Institut d'estadística de Catalunya (IDESCAT) y el Instituto Nacional de Estadística (INE). El INE y el IDESCAT no comparten necesariamente las opiniones expresadas en este trabajo, que son imputables únicamente a los firmantes del mismo. Los autores agradecen los comentarios de N.T. Longford a una versión preliminar de este trabajo.

[1] Institut d'Estadística de Catalunya. Via Laietana 58, 08003 Barcelona.

[2] Departament d'Economia i Empresa. Universitat Pompeu Fabra. Ramon Trias Fargas 25-27, 08005 Barcelona.

–Recibido en diciembre de 2001.
–Aceptado en marzo de 2002.

1. INTRODUCCIÓN

La mayoría de las grandes encuestas que llevan a cabo los organismos estadísticos nacionales han sido diseñadas en base a un tamaño muestral que garantiza la mayor fiabilidad de los estimadores a escala estatal. El tamaño de las muestras y su diseño suelen ser adecuados para proporcionar una precisión aceptable cuando se desea obtener estimadores para algunas áreas más pequeñas, por ejemplo a escala regional o provincial. Sin embargo, en ciertos casos el tamaño de las muestras es insuficiente. Éste es el caso si se quiere estimar no ya los niveles de una variable, sino su tasa de variación interanual, ya que este estadístico tiene una varianza mucho mayor. Esta es una situación normal y muy relevante en el ámbito de la estadística económica. Conocer la evolución de las variables es muchas veces tan o más importante que conocer su nivel.

Si se desea ofrecer información estadística adecuada para esas áreas más pequeñas sólo hay dos alternativas posibles: o se repite la encuesta adecuándola al tamaño del área de interés, o se recurre a la teoría estadística de estimación en pequeñas áreas para mejorar la calidad de los estimadores obtenidos a partir de la encuesta original.

En nuestro país existen encuestas diseñadas y aplicadas a algunas CA, que ofrecen un estimador ideal. Esta situación se da sobre todo en el País Vasco. Sin embargo debemos reconocer que los recursos empleados en la ejecución de estas encuestas son importantes, y la duplicación de cuestionarios en un mismo territorio representa un problema muy significativo. Por esta razón, algunas CCAA se han decantado hacia la utilización de algún tipo de estimador indirecto. Como ejemplo podemos citar el caso de Cataluña, donde el Institut d'Estadística de Catalunya (Idescat) ha elaborado un estimador sintético del Índice de Producción Industrial que ha resultado una alternativa muy adecuada en ese territorio (véase Costa, 1996 o Costa y Galter, 1994). Este estimador es un índice compuesto que utiliza 153 índices primarios estimados a escala nacional, y los pondera en función de su importancia relativa en la economía catalana. El Instituto Nacional de Estadística publicó durante un cierto tiempo unos estimadores regionales del IPI basados en esta metodología; sin embargo, estudios como los de Clar, Ramos y Suriñach (2000) han cuestionado la idoneidad de extender la metodología catalana al resto de territorios del Estado. El problema principal es que, para evitar que los sesgos de los estimadores sintéticos sean relevantes, es necesario que la estructura industrial de la CA sea similar a la del conjunto nacional, y eso no se da en todos los territorios considerados. Por lo tanto el estimador sintético puede ser excesivamente sesgado en algunos casos. En este trabajo pretendemos exponer y aplicar algunas de las soluciones ofrecidas por la teoría estadística para la elaboración de estimadores mejorados.

Existe una variada metodología que versa sobre el desarrollo de estimadores en áreas pequeñas. El lector puede consultar los artículos de Platek, Rao, Särndal y Singh (1987), Isaki (1990), Ghosh y Rao (1994), o Singh, Gambino y Mantel (1994) y obtener una visión de conjunto de las mismas. Una primera clasificación divide los diversos métodos

existentes en dos categorías: métodos tradicionales, y métodos basados en modelos. A su vez, los métodos tradicionales pueden incluir estimadores directos, indirectos, o una combinación de ambos. Los estimadores directos tradicionales utilizan únicamente datos provenientes del área pequeña de interés. Son por lo general no sesgados pero un insuficiente tamaño muestral hace que cuenten con una elevada varianza. Los estimadores indirectos tradicionales y los estimadores basados en modelos obtienen más precisión al utilizar también observaciones muestrales que provienen de áreas relacionadas, o áreas vecinas. Los estimadores sintéticos son estimadores tradicionales indirectos que se obtienen a partir de estimadores no sesgados de un área grande. A partir de éstos últimos es posible derivar estimadores para áreas más pequeñas bajo el supuesto de que éstas poseen la misma estructura (a efectos del fenómeno estudiado) que la gran área inicial. El incumplimiento de esta condición puede resultar en estimadores sesgados. Los estimadores compuestos tradicionales son una combinación lineal de estimadores directos y estimadores sintéticos. Los estimadores basados en modelos pueden interpretarse como estimadores compuestos, pero a diferencia de los tradicionales en los que los pesos se determinan de antemano o simplemente se hacen dependientes del tamaño de la muestra, aquí esas ponderaciones dependen de la estructura de covarianzas del estimador. Para más información sobre este tema se pueden consultar los artículos recientes de Cressie (1995), Datta *et al.* (1999), Farrell, MacGibbon y Tomberlin (1997), Ghosh y Rao (1994), Pfefferman y Barnard (1991), Raghunathan (1993), Singh, Mantel y Thomas (1994), Singh, Stukel y Pfeffermann (1998), o Thomas, Longford y Rolph (1994).

El estimador compuesto propuesto en este artículo minimiza el error cuadrático medio, una vez especificado un modelo de comportamiento de la varianza. En este sentido, representa una mejora respecto de estimadores alternativos, directos o sintéticos (indirectos), en áreas pequeñas. Ilustraremos la técnica con ayuda de los datos de ocupación provenientes de la Encuesta de Población Activa, pero la metodología propuesta puede ser aplicada a la estimación de otros indicadores con muy pocas modificaciones. Nos centraremos en concreto en la variación trimestral del empleo total en el sector industrial. La varianza de los datos tiene múltiples componentes: dependerá del territorio considerado y del período considerado.

En la sección 2 exponemos los fundamentos teóricos de la estimación de áreas pequeñas que proponemos. Los datos son descritos en la sección 3 y en la 4 adaptamos la teoría al caso que nos ocupa. La sección 5 expone los resultados principales de la aplicación del estimador a las 17 CCAA españolas. La sección 6 concluye el presente trabajo, resume las conclusiones y sugiere líneas futuras de desarrollo de estimadores regionales. Los apéndices A, B y C muestran los datos originales, algunos cálculos estadísticos, y desarrollos teóricos que por su complejidad se ha preferido mantener al margen del texto principal.

2. FUNDAMENTOS TEÓRICOS DE LA ESTIMACIÓN EN ÁREAS PEQUEÑAS

Supongamos que disponemos de dos estimadores $\widehat{\theta}_1 \sim (\theta_1, \sigma_1^2)$ y $\widehat{\theta}_2 \sim (\theta_2, \sigma_2^2)$ que son (aproximadamente) incorrelacionados y deseamos estimar el parámetro θ_2 (en el presente trabajo, $\widehat{\theta} \sim (\theta, \sigma^2)$ denota que la media y la varianza del estadístico $\widehat{\theta}$ son respectivamente θ y σ_2). En el caso general en que $\theta_1 \neq \theta_2$, es obvio que el criterio de primar a un estimador centrado conduce a elegir $\widehat{\theta}_2$ y, por tanto, a ignorar por completo el valor de $\widehat{\theta}_1$. Sin embargo, cuando la varianza σ_2^2 es grande en comparación con σ_1^2 y con el sesgo al cuadrado $(\theta_1 - \theta_2)^2$, se puede demostrar que la información que proporciona $\widehat{\theta}_1$ no es despreciable en la estimación de θ_2. Un estimador alternativo al estimador centrado de θ_2, es el estimador compuesto

$$\widehat{\theta}_c = \pi\widehat{\theta}_1 + (1-\pi)\widehat{\theta}_2, \tag{1}$$

con un peso π adecuado para combinar de forma «óptima» la información sobre θ_2 contenida en los estimadores $\widehat{\theta}_1$ y $\widehat{\theta}_2$.

Un posible y comúnmente aceptado criterio de optimalidad consiste en minimizar la media de los errores de estimación al cuadrado, es decir, minimizar el MSE («mean square error») del estimador. Dado que $MSE\left(\widehat{\theta}_c\right) = \left(E(\widehat{\theta}_c) - \theta_2\right)^2 + Var(\widehat{\theta}_c)$, con

$$E(\widehat{\theta}_c) - \theta_2 = E(\pi\widehat{\theta}_1 + (1-\pi)\widehat{\theta}_2) - \theta_2 = \pi\theta_1 + (1-\pi)\theta_2 - \theta_2 = \pi(\theta_1 - \theta_2)$$

y $Var(\widehat{\theta}_c) = \pi^2\sigma_1^2 + (1-\pi)^2\sigma_2^2$ (en el caso de $\widehat{\theta}_1$ y $\widehat{\theta}_2$ incorrelacionados), obtenemos

$$MSE(\widehat{\theta}_c) = \pi^2(\theta_1 - \theta_2)^2 + \pi^2\sigma_1^2 + (1-\pi)^2\sigma_2^2$$

De manera que minimizar el MSE conduce a resolver la ecuación

$$\pi(\theta_1 - \theta_2)^2 + \pi\sigma_1^2 + (1-\pi)\sigma_2^2 = 0,$$

es decir,

$$\pi((\theta_1 - \theta_2)^2 + \sigma_1^2 + \sigma_2^2) = \sigma_2^2,$$

de manera que ell valor óptimo de π es $\pi = \sigma_2^2/\nu$, con $\nu = (\theta_1 - \theta_2)^2 + \sigma_1^2 + \sigma_2^2$.

Nótese que cuando la varianza σ_1^2 es pequeña con relación a los demás términos que componen ν, obtenemos la siguiente expresión

$$\pi = \frac{\sigma_2^2}{(\theta_1 - \theta_2)^2 + \sigma_2^2} \tag{2}$$

Dado que en general el sesgo al cuadrado $(\theta_1 - \theta_2)^2$ es desconocido, aproximaremos dicho valor mediante $\delta^2 = E(\theta_1 - \theta_2)^2$, que a menudo se corresponderá con la denominada varianza *entre* grupos. En general σ_2^2 es una varianza *dentro* de grupos.

Cabe remarcar los casos extremos:

1. cuando la dispersión dentro de los grupos es pequeña en relación al sesgo, $\pi = 0$ y el estimador compuesto coincide con $\widehat{\theta}_2$;
2. cuando la dispersión entre grupos δ^2 es pequeña en relación a la varianza entre grupos, entonces el estimador compuesto coincide con $\widehat{\theta}_1$.

Un caso más general vendría descrito por dos estimadores $\widehat{\theta}_1 \sim (\theta_1, \sigma_1^2)$ y $\widehat{\theta}_2 \sim (\theta_2, \sigma_2^2)$ en donde la varianza σ_1^2 no es ignorable y la covarianza $\gamma = cov(\theta_1, \theta_2)$ es no nula. En este caso se demuestra que el valor óptimo π viene dado por (Longford, 2001)

$$\pi = \frac{\sigma_2^2 - \gamma}{(\theta_1 - \theta_2)^2 + \sigma_1^2 + \sigma_2^2 - 2\gamma} \tag{3}$$

Nótese que en dicho caso será necesaria también información sobre el valor de la covarianza γ. En el caso de muestreo aleatorio simple dicha covarianza será igual a $q \times \sigma_2^2$ con $0 \leq q \leq 1$, donde q es la fracción de muestra aportada por la k-ésima CA, de manera que la expresión anterior se transforma en

$$\pi = \frac{\sigma_2^2 (1-q)}{(\theta_1 - \theta_2)^2 + \sigma_1^2 + \sigma_2^2 (1-2q)} \tag{4}$$

Resultados parecidos se obtienen en un contexto multivariante. Si lo que se desea es estimar $w'\theta_2$, para un vector de pesos w arbitrario, a partir de los estimadores vectoriales $\widehat{\theta}_1 \sim (\theta_1, \Sigma_1)$ y $\widehat{\theta}_2 \sim (\theta_2, \Sigma_2)$, el estimador compuesto se obtendrá a partir de la combinación lineal $w'\widehat{\theta}_c$ con

$$\widehat{\theta}_c = \Pi\widehat{\theta}_1 + (1 - \Pi)\widehat{\theta}_2 \tag{5}$$

La expresión de la matriz de pesos Π que minimiza el MSE en el caso simplificado en que $\widehat{\theta}_1$ y $\widehat{\theta}_2$ no están correlacionados, es $\Pi = \Sigma_2 V^{-1}$ con $V = bb' + \Sigma_1 + \Sigma_2$ y $b = (\theta_1 - \theta_2)$ (Longford, 1999). La expresión de b, en general desconocida, se sustituirá por la siguiente matriz de varianzas y covarianzas *entre* grupos

$$\Delta = E(\theta_1 - \theta_2)(\theta_1 - \theta_2)'$$

Dichos estimadores vectoriales surgen al tratar la estimación de varios parámetros simultáneamente, por ejemplo las tasas de variación de paro de varios sectores económicos.

3. ANÁLISIS ESTADÍSTICO DE LOS DATOS

Los datos utilizados en este artículo provienen de la Encuesta de Población Activa (EPA) y de la Contabilidad Regional de España (CRE). Disponemos de datos trimestrales de la EPA de Ocupación industrial en España y en cada CCAA por ramas de actividad, desde 1987. También contamos con datos anuales de Empleo industrial y por ramas provenientes de la CRE, para España y las CCAA, desde 1995 a 1997.

En este trabajo nos centramos en el sector industrial. El Apéndice B muestra las ramas de actividad del sector industrial consideradas por la CRE, y como hemos construido su equivalencia con las ramas de actividad consideradas por la EPA, algo más desagregadas.

Los datos originales de la EPA representan miles de personas ocupadas en cada Comunidad. A partir de ellos, calculamos las tasas de variación trimestral de la ocupación en cada área geográfica en la forma habitual.

La notación básica utilizada es la siguiente:

- $y_{ij}(k,t)$ es la tasa de variación de la ocupación en la rama j del sector productivo i en el territorio k, y en el trimestre t.
- $y_i(k,t)$ es la tasa de variación de la ocupación en el sector productivo i y la Comunidad Autónoma[1] k, en el trimestre t.
- $y_{ij}(\cdot,t)$ es la tasa de variación de la ocupación en la rama j del sector productivo i en el conjunto del Estado Español, en el trimestre t.
- $y_i(\cdot,t)$ es la tasa de variación de la ocupación en el sector productivo i en el conjunto del Estado Español, en el trimestre t.

Las variables $y_{ij}(\cdot,t)$, $y_i(\cdot,t)$ y $y_i(k,t)$ pueden obtenerse como estimadores directos a partir de los datos disponibles en la EPA aunque el diseño de esta encuesta no permite obtener estimadores suficientemente precisos en el último de los casos.

También podemos crear un estimador sintético para cada territorio inspirado en la metodología del Idescat (véase Costa y Galter, 1994), combinando el estimador directo al nivel nacional y la información suplementaria extraída de la CRE, gracias a la cual podemos adaptar la información nacional a la estructura de la industria de cada una de las diferentes CCAA.

[1]No disponemos de datos detallados trimestrales por ramas de actividad dentro de cada sector productivo y autonomía, sólo anuales.

Calculamos

$$z_i(k,t) = \sum_{j=1}^{J_i} w_{ij}(k) y_{ij}(\cdot,t) \tag{6}$$

donde J_i es el número de ramas en el sector i, y $w_{ij}(k)$ es el peso de la rama j en el sector i de la Comunidad k, en un periodo determinado y fijo (en concreto, el año 1997).

El Apéndice numérico A muestra los estimadores directos y sintéticos industriales de cada Comunidad para cada trimestre, así como las ponderaciones y valores nacionales utilizados en la elaboración del estimador sintético.

En este punto contamos con dos estimadores: el directo que es insesgado pero muy impreciso, y el sintético, que es sesgado aunque con una varianza más reducida.

4. ESTIMACIÓN DE LA TASA DE VARIACIÓN DE LA OCUPACIÓN EN LA INDUSTRIA

Proponemos aquí un estimador compuesto de la tasa de variación de la ocupación de cada CA, de periodicidad trimestral. El estimador compuesto es una combinación lineal de un estimador sintético y uno directo donde los pesos otorgados a cada componente dependen de la composición de la varianza de los indicadores de cada Comunidad.

Los dos estimadores $z_i(k,t)$ y $y_i(k,t)$ representan, respectivamente, los estimadores sintético y directo de la variación de la ocupación para el sector industrial de la Comunidad k en el momento t. Es decir suponemos que $\widehat{\theta}_2(k,t) = y_i(k,t) \sim (\theta_2(k,t)\ ,\ \sigma_2^2(k,t))$ y $\widehat{\theta}_1(k,t) = z_i(k,t) \sim (\theta_1(k,t)\ ,0)$. Si definimos $\delta(k,t) = \theta_2(k,t) - \theta_1(k,t)$, de acuerdo con las expresiones (1) y (2) de la Sección 2 obtenemos el estimador compuesto

$$\widehat{\theta}_c(k,t) = \pi(k,t)\widehat{\theta}_1(k,t) + (1-\pi(k,t))\,\widehat{\theta}_2(k,t) \tag{7}$$

donde

$$\pi(k,t) = \frac{\sigma_2^2(k,t)}{\delta^2(k,t) + \sigma_2^2(k,t)} \tag{8}$$

Dadas las características de muestreo de la EPA, y suponiendo constancia a lo largo del tiempo de los tamaños muestrales, consideramos que las varianzas muestrales $\sigma_2^2(k,t)$ son constantes en el tiempo, es decir $\sigma_2^2(k,t) = \sigma_2^2(k)$, y que los sesgos al cuadrado $\delta(k,t)^2$ se comportan (para cada Comunidad), como un proceso estocástico estacionario en media. Por tanto, aproximaremos $\delta(k,t)^2$ por su valor promedio $(k) =$

$\frac{1}{T}\sum_{t=1}^{T}\delta(k,t)^2$. Se observa que un estimador de la suma $(k)+\sigma\ ^2_2(k)$ es precisamente la media muestral (para cada Comunidad) de la serie de diferencias al cuadrado

$$s_d^2(k)=\frac{1}{T}\sum_{t=1}^{T}d(k,t)^2=\frac{1}{T}\sum_{t=1}^{T}\left(\widehat{\theta}_2(k,t)-\widehat{\theta}_1(k,t)\right)^2 \tag{9}$$

y dado que $\widehat{\theta}_2(k,t)-\widehat{\theta}_1(k,t)=\theta_2(k,t)-\theta_1(k,t)+\varepsilon(k,t)$ donde $\varepsilon(k,t)$ es el error de muestreo, obtenemos (en el caso en que $\widehat{\theta}_1$ y $\widehat{\theta}_2$ no estén correlacionados)

$$s_d^2(k)=\frac{1}{T}\left(\sum_{t=1}^{T}\delta(k,t)^2+\sum_{t=1}^{T}\varepsilon(k,t)^2+2\sum_{t=1}^{T}\delta(k,t)\varepsilon(k,t)\right) \tag{10}$$

que obviamente converge hacia $(k)+\sigma\ ^2_2(k)$, la expresión del denominador de $\pi(k,t)$. De este modo el peso óptimo vendrá dado por

$$\pi(k)=\frac{s_2^2(k)}{s_d^2(k)} \tag{11}$$

donde $s_2^2(k)$ es un estimador de la varianza muestral $\sigma_2^2(k)$. El Apéndice C muestra que el valor de esta varianza puede estimarse mediante la siguiente expresión

$$s_2^2(k)=2(CV(k))^2, \tag{12}$$

donde $CV(k)$ es el error relativo de muestreo medio calculado por el INE para la variable Ocupación en la k-ésima CA. Dicha expresión incorpora ya de forma automática la corrección necesaria para atender el efecto de diseño complejo de la muestra de la EPA.

En el caso en que la varianza del estimador sintético no sea ignorable, y exista una covarianza no nula entre los dos estimadores $\widehat{\theta}_1$ y $\widehat{\theta}_2$, puede verse que $s_d^2(k)$ continúa siendo un estimador consistente ahora del denominador de (4), de manera que en dicho caso la expresión del peso óptimo será

$$\pi(k)=\frac{s_2^2(k)-\widehat{\gamma}(k)}{s_d^2(k)} \tag{13}$$

donde $\widehat{\gamma}(k)$ es un estimador de la covarianza entre estimadores de la k-ésima CA[2].

Cuando la información histórica de cada comunidad no permita una estimación fiable de su sesgo específico (o en el caso extremo en que a priori se asigne una magnitud de sesgo común a todas las comunidades), es fácil ver que el valor común del sesgo será

[2]En general $\widehat{\gamma}(k)=q(k)s_2^2(k)$ donde $q(k)$ corresponde a la fracción de muestra equivalente aportada por la k-ésima CA.

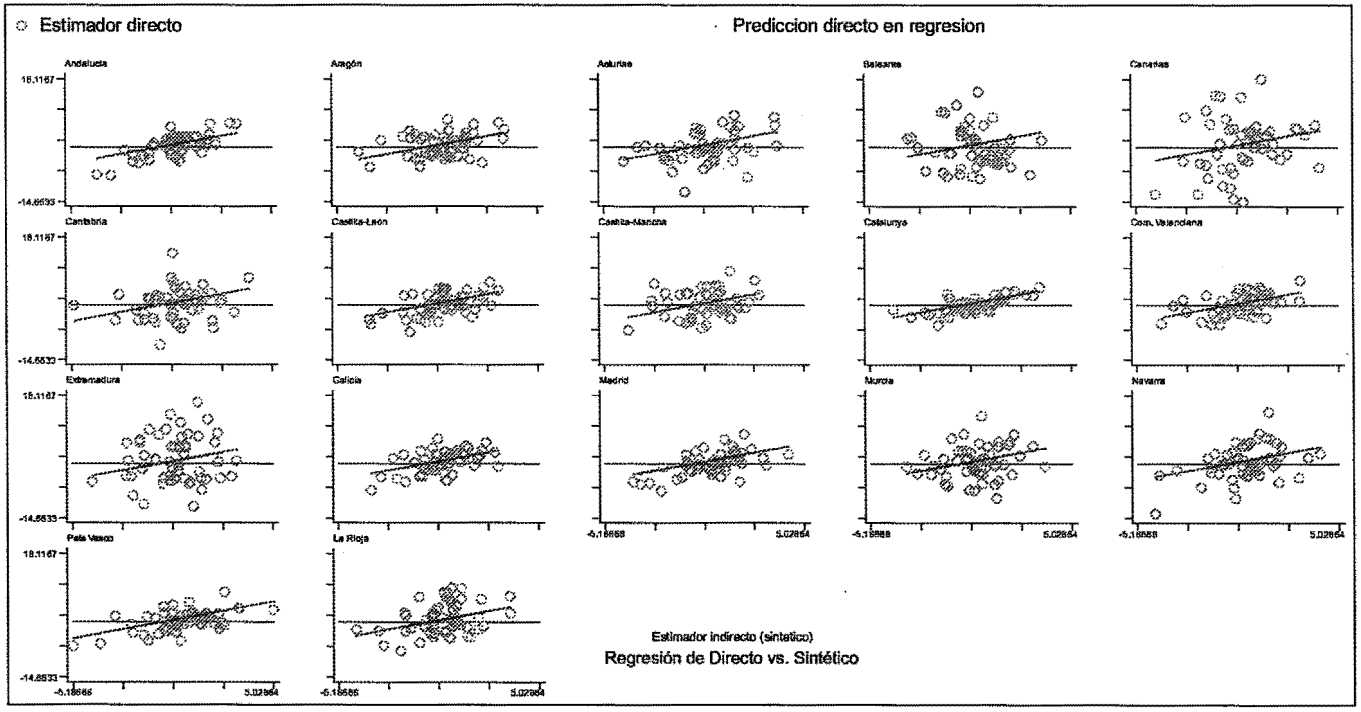


Figura 1. Regresión de Directo sobre Sintético

estimado por $[\frac{1}{K}\sum_k s_d^2(k) - \frac{1}{K}\sum_k s_2^2(k)]$, y la expresión del peso del estimador compuesto será

$$\pi(k) = \frac{s_2^2(k) - \widehat{\gamma}(k)}{[\frac{1}{K}\sum_k s_d^2(k) - \frac{1}{K}\sum_k s_2^2(k)] + s_2^2(k)}, \tag{14}$$

cuyo valor vemos que varía en cada CA a través solamente de la varianza muestral $s_2^2(k)$ y de la covarianza $\widehat{\gamma}(k)$.

5. ILUSTRACIÓN EN EL SECTOR INDUSTRIAL

En este apartado incluimos los cálculos efectuados para obtener el estimador compuesto y los resultados para las diferentes CCAA.

Primeramente hemos obtenido un estimador directo de la tasa de variación de la ocupación en la industria para cada CA. El estimador se obtiene directamente a partir de los datos de la EPA, restringidos a un área geográfica concreta. El cuadro 1 del Apéndice A muestra las tasas de variación calculadas mediante este procedimiento. Puede apreciarse que existen CCAA que presentan una gran variabilidad temporal en sus tasas de ocupación (Canarias, Baleares o La Rioja por ejemplo), mientras que otras como Madrid o Cataluña presentan una evolución más suave de la ocupación.

También hemos calculado el estimador sintético presentado en la ecuación (6). El estimador sintético resulta de ponderar las tasas de variación de las ramas industriales a nivel estatal provenientes de la EPA (véase el cuadro 3 del Apéndice A), con los porcentajes de ocupación industrial por ramas de cada Comunidad[3] que se obtienen de la CRE (véase el cuadro 4 del Apéndice A). Dado que la EPA cuenta con más de las 14 ramas industriales con las que cuenta la CRE, aplicamos las equivalencias que describimos en el Apéndice B con objeto de homogeneizar las series utilizadas en el cálculo del estimador. Los estimadores indirectos obtenidos para cada CA están recogidos en el cuadro 2 del Apéndice A.

La figura 4 permite comprobar la variabilidad del estimador directo en relación al indirecto. La serie de estimadores sintéticos de alguna manera «suaviza» la serie de estimadores directos, en cada CA. Se aprecia no obstante que la diferencia entre ambos estimadores es diferente según consideremos unas u otras CCAA. Por ejemplo cabe destacar el caso de Cataluña, donde ambos estimadores se encuentran bastante próximos; o los de Canarias o Extremadura en que la diferencia se acrecienta. Estas particularidades pueden observarse mejor con ayuda de la figura 6, que muestra mediante diagramas de

[3] Aunque en el presente artículo hemos optado por mantener unas ponderaciones fijas correspondientes al año 1997, el estimador sintético podria mejorarse utilizando ponderaciones variables en el tiempo.

caja la distribución de la diferencia entre los estimadores directo e indirecto trimestrales, para cada una de las 17 CCAA.

Es evidente que el estimador sintético suaviza en exceso la evolución de la tasa de variación de ocupación y dada la relativamente alta fiabilidad de los datos directos estimados en la EPA no resulta una buena alternativa en la mayoría de las CCAA.

Para calcular el estimador compuesto, necesitamos obtener los pesos otorgados a los estimadores directo e indirecto, específicos de cada CA, que hemos mostrado en la ecuación (2). Hemos optado por esta expresión simplificada dada la dificultad, en este momento, de contar con estimadores adecuados de la covarianza entre estimadores directos y sintéticos. El denominador del peso $\pi(k)$ se obtiene calculando el promedio de la diferencia entre los estimadores directo e indirecto, al cuadrado. Esta expresión es un estimador del sesgo al cuadrado más la varianza del error muestral del estimador directo. El numerador, que coincide con la varianza del error muestral del estimador directo, se obtiene multiplicando por dos el cuadrado del error de muestreo relativo que el INE otorga a la variable Ocupación en la EPA. La tabla 1 muestra los resultados intermedios y el valor final de los pesos.

Tabla 1. Cálculo de los pesos del estimador compuesto

	$CV^c(O)$	$s_2^2(k)$	$s_d^2(k)$	$\pi(k)$
Andalucía	1,202	2,89	4,748	0,609
Aragón	1,728	5,97	8,616	0,693
Asturias	2,512	12,62	14,498	0,875
Baleares	2,264	10,25	32,553	0,315
Canarias	2,342	10,97	49,223	0,223
Cantabria	2,738	14,99	17,084	0,878
Castilla-León	1,466	4,30	6,791	0,633
Castilla-La Mancha	1,526	4,66	12,005	0,388
Cataluña	1,382	3,82	2,086	1,00[4]
C. Valenciana	1,454	4,23	5,880	0,719
Extremadura	2,574	13,25	39,589	0,335
Galicia	1,708	5,835	6,591	0,885
Madrid	1,602	5,13	6,293	0,816
Murcia	2,442	11,93	18,664	0,639
Navarra	2,524	12,74	17,326	0,735
País Vasco	1,688	5,70	6,287	0,906
La Rioja	3,002	18,02	18,427	0,978

Nota: C.V. está calculado a partir de los valores medios de 1992.

[4]Este peso ha sido truncado a 1. En el caso de Cataluña se espera una correlación alta entre el estimador directo y el sintético, correlación que induce sesgo en la fórmula utilizada de cálculo del coeficiente para dicha CA. El cálculo del peso óptimo para Cataluña teniendo en cuenta dicha correlación involucraría características del diseño de muestra que creemos exceden el ámbito del presente trabajo.

En general, el estimador compuesto otorga un peso mayor al estimador indirecto en casi todas las CCAA. Se observa que el estimador compuesto otorga menor peso al estimador indirecto en aquellas Comunidades donde la varianza del error de muestreo es menor en relación a su sesgo promedio. Así ocurre en Castilla-La Mancha, Extremadura, Baleares o Canarias, como casos más extremos. En cambio, el estimador sintético recibe mayor peso en el caso de Cataluña, País Vasco o La Rioja entre otras, dado que en estas CCAA el sesgo promedio del estimador sintético es pequeño en relación a la varianza del error de muestreo. Se advierte como algunas de las regiones más industrializadas cuentan con un valor muy alto, mientras que los dos archipiélagos y algunas regiones menos industrializadas muestran valores más bajos. Estos resultados son llamativos ya que el estimador directo pierde peso precisamente en las CCAA mayores, mientras que lo pierde en las CCAA que pueden ser más propiamente consideradas pequeñas áreas.

A continuación examinamos el comportamiento de este estimador en el caso de dos Comunidades concretas que destacan por tener poco y mucho peso respectivamente otorgado al estimador indirecto : Canarias y el País Vasco[5].

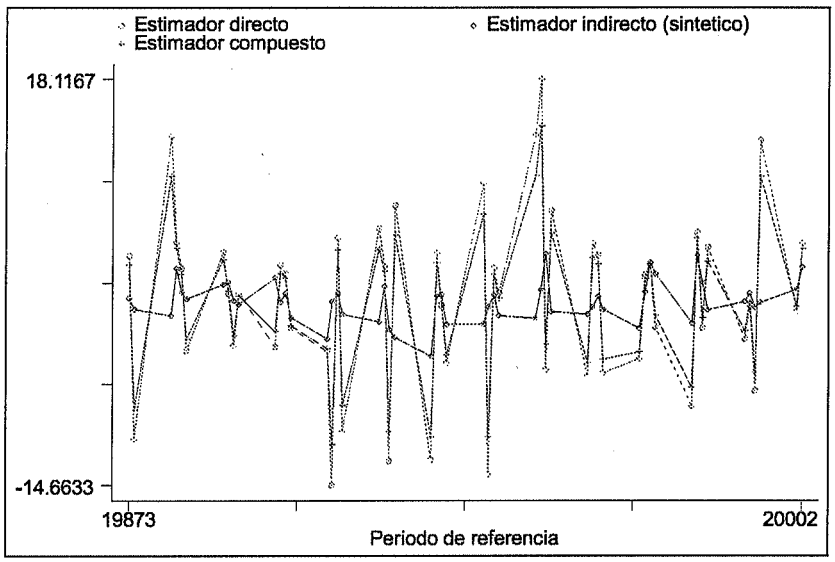


Figura 2. Comunidad Autónoma de Canarias

[5]El estimador compuesto en el caso de Cataluña o La Rioja, otorga un peso mayor todavía al estimador indirecto. En el caso de Cataluña ambos coinciden completamente, y en el caso de La Rioja, las discrepancias són mínimas, por lo que hemos preferido escoger otra Comunidad para mostrar un gráfico más claro de las diferencias entre los tres distintos estimadores.

En el caso de Canarias (figura 2), el estimador compuesto otorga casi el 78% del peso al estimador directo. Dado que el estimador directo de la tasa de variación de la ocupación en esta Comunidad a lo largo del período fluctúa entre los valores del 18,1% y −14,7%, las diferencias en las estimaciones puntuales de unos y otros estimadores pueden ser importantes. Esta última circunstancia se observa mejor en las figuras 6 y 7, que muestra mediante diagramas de caja la diferencia entre los valores del estimador directo e indirecto, y compuesto e indirecto de cada CA.

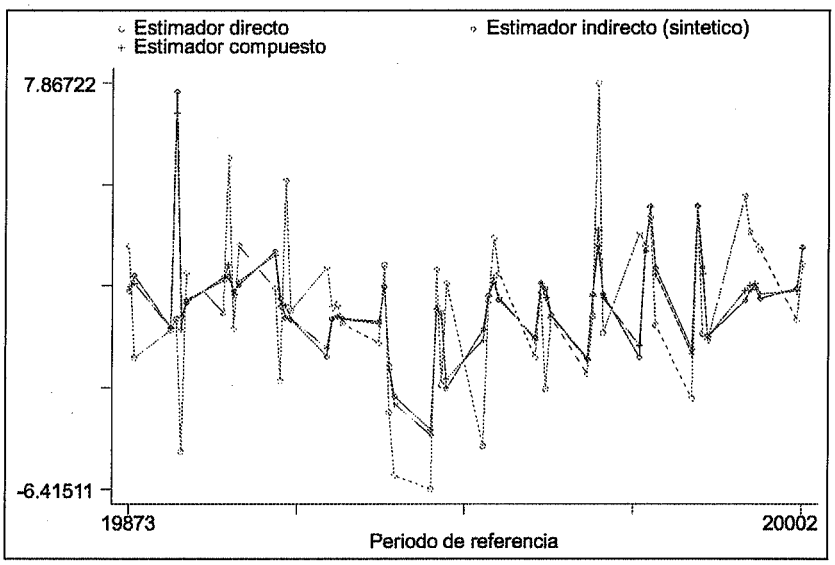


Figura 3. Comunidad Autónoma del País Vasco

En el caso del País Vasco (ver figura 3), donde la diferencia entre la estimación puntual del estimador directo y del sintético es también apreciable, con estimaciones directas de la tasa de variación de la ocupación que oscilan entre el 7,8% y el −6,4%, el estimador compuesto da más peso al estimador indirecto (la ponderación es 0,91). Esto es así porque el error de muestreo en esta Comunidad es bastante acusado en relación al sesgo incorporado en el estimador sintético.

Estos son dos casos extremos. La figura 5 muestra los tres estimadores: directo, indirecto y compuesto en las 17 CCAA.

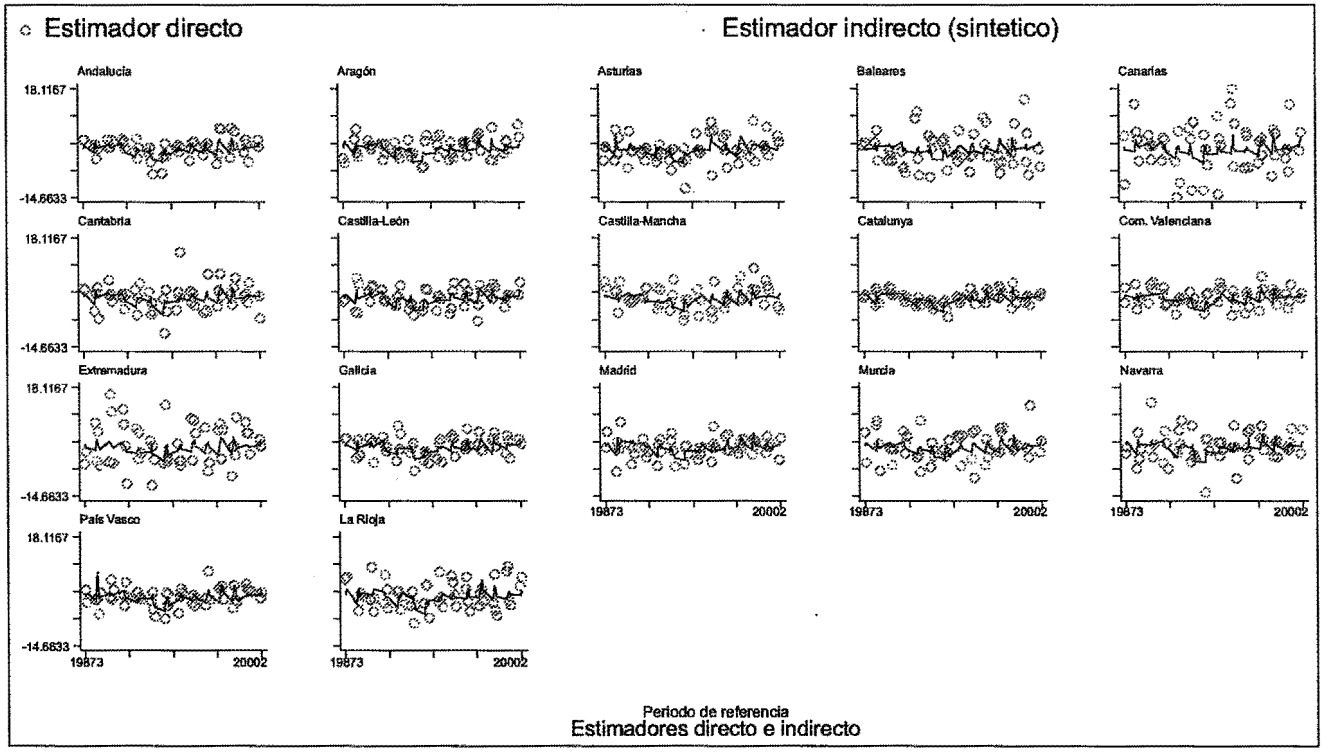


Figura 4. Estimadores directo e indirecto, por Comunidad

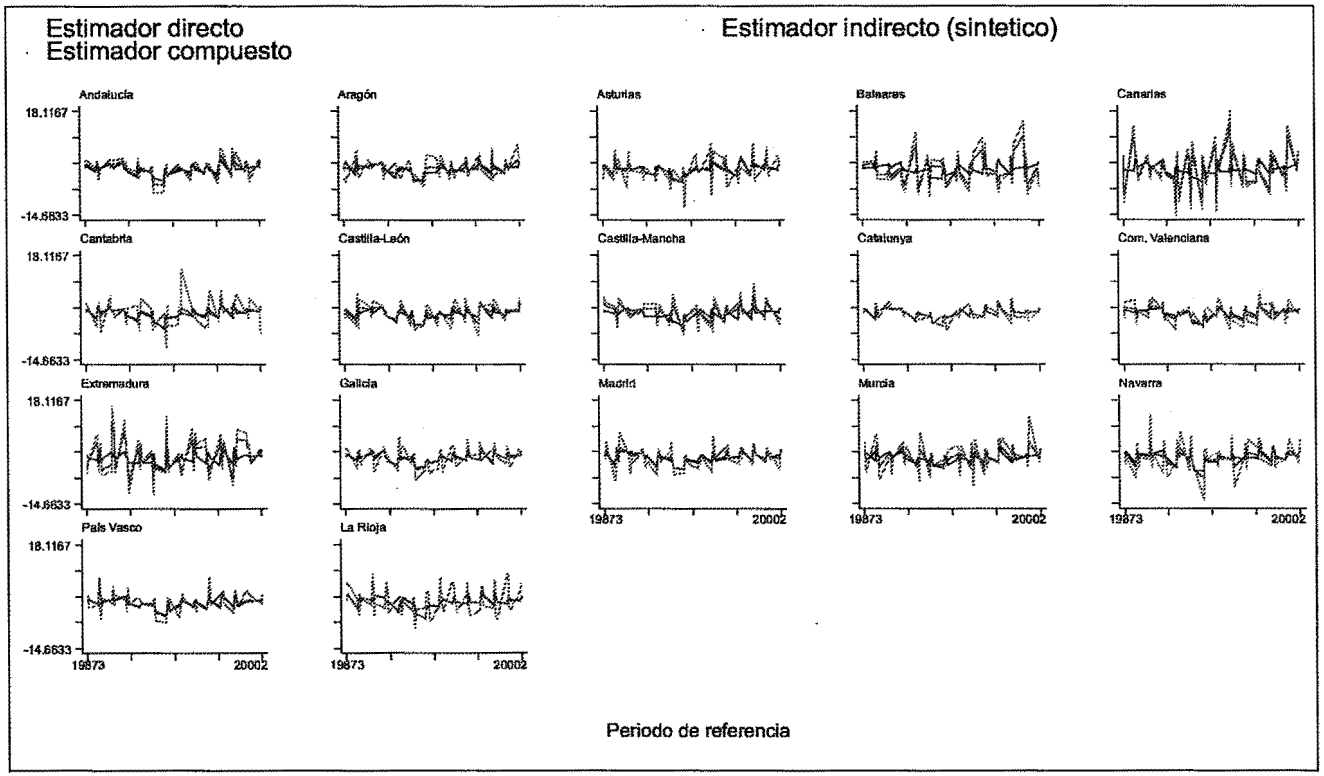


Figura 5. Comparación de estimadores directo, indirecto y compuesto, por Comunidad

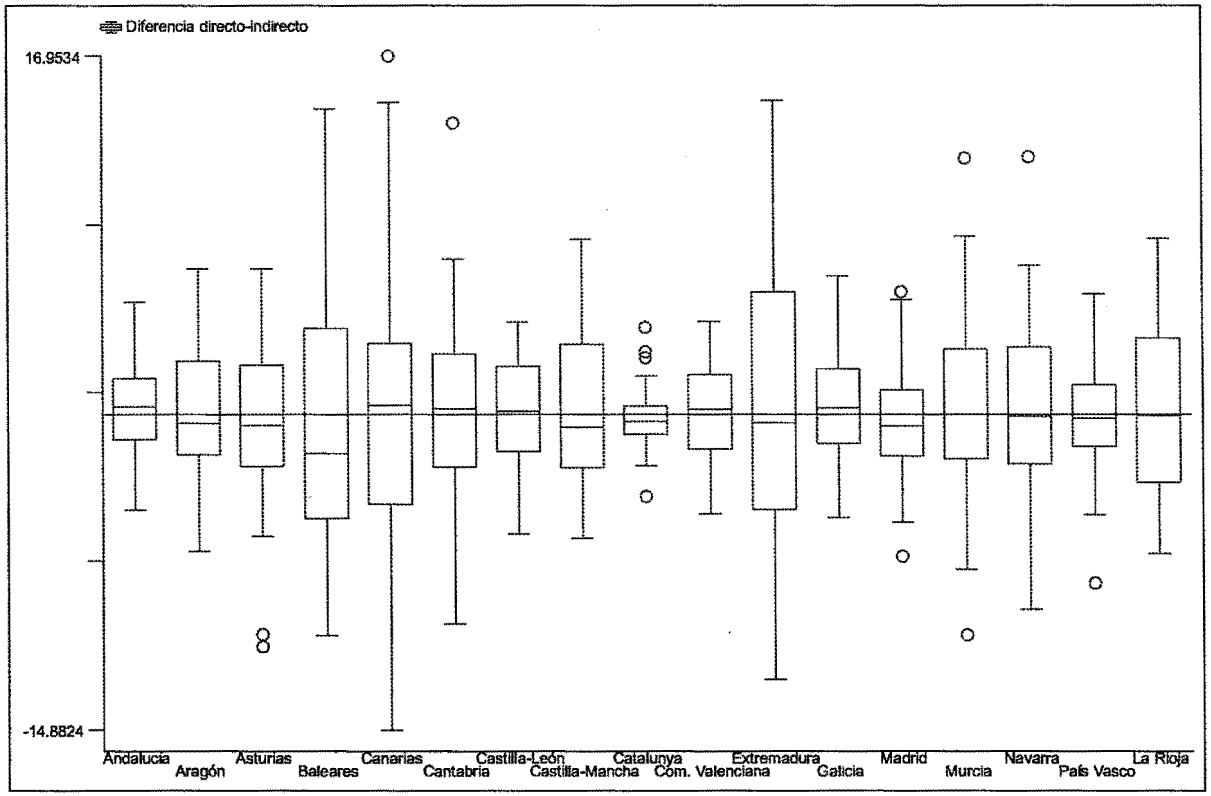


Figura 6. Distribución por Comunidad de la diferencia entre los estimadores directo e indirecto

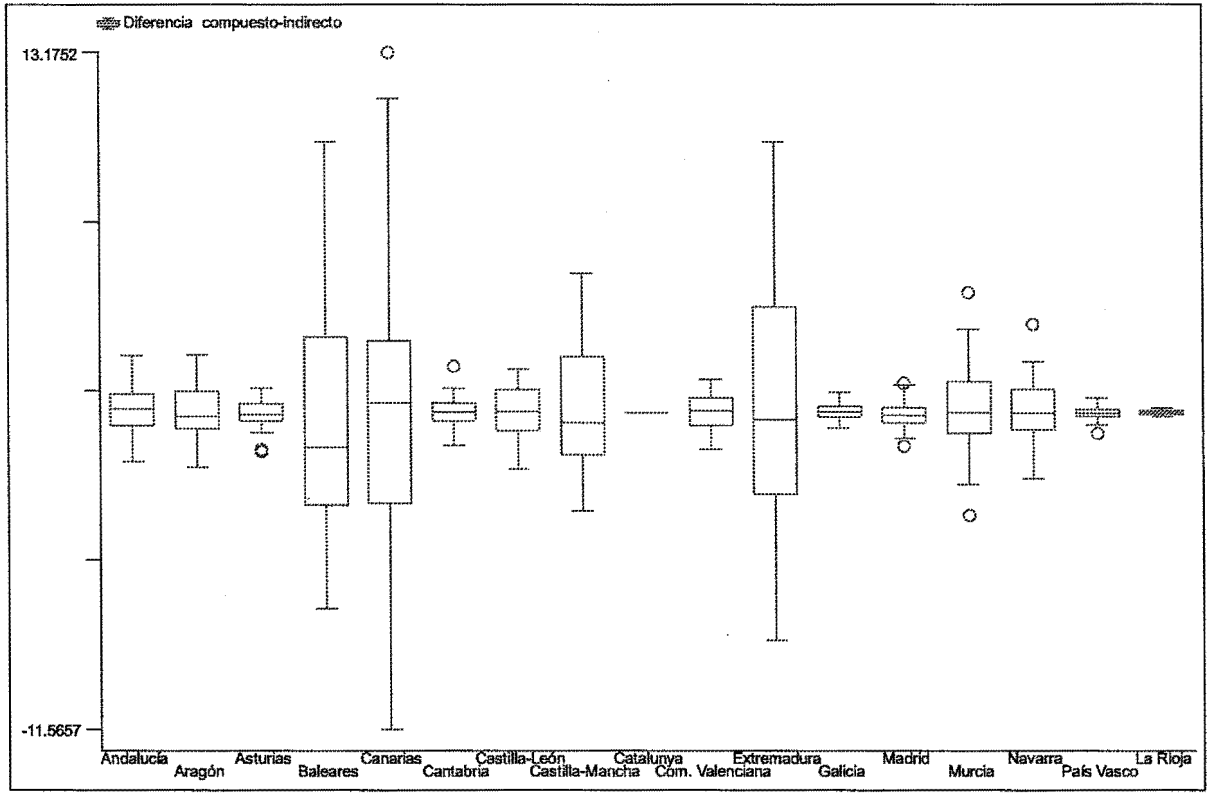


Figura 7. Diferencia por comunidades entre los estimadores compuesto e indirecto

6. CONCLUSIONES

En este trabajo hemos creado un estimador compuesto aplicado a la estimación de la variación de la tasa de ocupación de 17 CCAA españolas. El estimador compuesto es el resultado de combinar de forma lineal otros dos estimadores: el directo y el sintético. El primero es un estimador insesgado pero poco preciso, mientras que el segundo posee una varianza más reducida pero suele ser sesgado. Si asumimos un modelo de comportamiento para el estimador directo en función del indirecto, las ponderaciones utilizadas para obtener el estimador compuesto son función del sesgo del estimador sintético, y de la varianza del error muestral del estimador directo. De esta forma, el estimador compuesto resultante es el que minimiza el error cuadrático medio, dado el modelo adoptado.

El estimador resultante otorga *más peso al estimador directo en aquellas Comunidades donde la varianza del error de muestreo es menor en relación al sesgo promedio del estimador indirecto.* Así ocurre en las Comunidades de Canarias, Baleares, Castilla-La Mancha, o Extremadura, como casos más extremos. En cambio, el estimador sintético gana más protagonismo en el caso de Cataluña, La Rioja, País Vasco, o Madrid, dado que en estas Comunidades el sesgo promedio del estimador sintético es pequeño en relación a la varianza del error de muestreo.

Cabe señalar que la versión del estimador compuesto aquí presentada puede incorporar múltiples refinamientos, desde una modificación adecuada de los pesos en determinados períodos con características especiales, a un tratamiento diferenciado de unas determinadas CCAA. Asimismo la técnica utilizada puede aplicarse a otros conjuntos de datos.

Podemos señalar tres conclusiones principales:

1. Puede verse para Cataluña que el estimador compuesto es equivalente al sintético, lo que avalaría la utilización por parte del Idescat de indicadores sintéticos para aproximar tasas de variación interanual (tal como se está haciendo en los casos del IPI y del IPRI).
2. Por lo que respecta a la adopción por parte del INE (durante la segunda parte de los años noventa) de los estimadores sintéticos para hacer un IPI regional y de algunas CCAA (que lo están utilizando en la actualidad, tanto para el IPI como para el IPRI), habría que coincidir con Clar, Ramos y Suriñach (2000), en que estas estimaciones pueden ser adecuadas en algunos casos, pero no lo son en todos los casos. En algunas CCAA como Cataluña, claramente son aceptables, pero en otras, como en las islas, no parece ser un camino válido.
3. El anterior resultado no deja de ser curioso: en el estimador compuesto gana peso el estimador sintético cuando las CCAA son menos «pequeñas áreas», perdiéndolo en aquellas CCAA que de forma más natural podrían ser tomadas como auténticas pequeñas áreas.

Nos encontramos con un comportamiento paradójico de los estimadores compuestos. En ellos tiene mucho peso el estimador sintético en «pequeñas áreas» muy pequeñas (para las cuales el estimador directo tiene una varianza prácticamente infinita), y también en el caso de las «pequeñas áreas» muy grandes, como en el caso catalán (por tener el estimador sintético un sesgo irrelevante). Esto implicaría un comportamiento no lineal del peso del estimador directo ante incrementos en la dimensión de la pequeña área. También supone un prometedor papel de la estimación indirecta en el terreno de la estadística regional.

El proyecto más inmediato en nuestra agenda de trabajo consiste en estudiar las propiedades muestrales de los estimadores compuestos. Sesgo, varianzas poblacional y muestral del estadístico, y dato a estimar (nivel, proporción o tasa de variación), son aspectos determinantes de los pesos con los que los estimadores compuestos basados en modelos combinan la información disponible. Es por lo tanto fundamental explorar, a través de diversos escenarios, la manera en que las características poblacionales regionales o locales (que pueden estar muy centradas o ser muy diferentes del área general), el tamaño muestral, el tipo de dato a estimar, etc., influyen en su comportamiento. Asimismo será necesario aplicar técnicas específicas de estimación de sesgo y varianza dado que no todos los escenarios permiten derivar expresiones analíticas útiles. Este estudio puede llegar a tener implicaciones en los diseños muestrales y, en concreto, en las distribuciones de la muestra entre los distintos subdominios de las encuestas (afijaciones), en el caso de tener previamente presente que éstos serán estimados no por estimadores directos, sino por medio de estimadores compuestos.

7. REFERENCIAS

Clar, M., Ramos, R. y Suriñach, J. (2000). «Avantatges i inconvenients de la metodologia de l'INE per a elaborar indicadors de la producció industrial per a les regions espanyoles», *Qüestiió*, 24, 1, 151-186.

Costa, A. (1996). «El IPPI, un indicador de la coyuntura industrial catalana», *Fuentes estadísticas*, 17. INE.

Costa, A. y Galter, J. (1994). «L'IPPI, un indicador molt valuós per mesurar l'activitat industrial catalana». *Revista d'indústria*, 3, Generalitat de Catalunya, Departament d'Indústria i Energia, 6-15.

Cressie, N. (1995). «Bayesian Smoothing of Rates in Small Geographic Areas». *Journal of Regional Science*, 35 (4), 659-73.

Datta, G. S., *et al.* (1999). «Hierarchical Bayes Estimation of Unemployment Rates for the States of the U.S». *Journal of the American Statistical Association*, 94 (448), 1074-82.

Farrell, P. J, Macgibbon, B. y Tomberlin, T. J. (1997). «Empirical Bayes Small-Area Estimation Using Logistic Regression Models and Summary Statistics». *Journal of Business & Economic Statistics*, 15 (1), 101-8.

Ghosh, M. y Rao, J. N. K. (1994). «Small Area Estimation: an Appraisal». *Statistical Science*, 9, 1, Canada Statistics, 55-93.

INSTITUTO NACIONAL DE ESTADÍSTICA (1994). *Evaluación de la calidad de los datos de la Encuesta de Población Activa.* Año 1992.

Isaki, C. T. (1990). «Title Small-Area Estimation of Economic Statistics». *Journal of Business & Economic Statistics*, 8 (4), 435-41.

Longford, N. T. (1999). «Multivariate shrinkage estimation of small area means amd proportions», *Journal od the Royal Statistical Society*, A, 162, 227-246.

Longford, N. T. (2001). «Synthetic estimators with moderating influence: the carry-over in cross-over trials revisited». *Statistis in Medicine*, 20, 3189-3203.

Pfeffermann, D. y Barnard, C. H. (1991). «Some New Estimators for Small-Area Means with Application to the Assessment of Farmland Values». *Journal of Business & Economic Statistics*, 9 (1), 73-84.

Platek, R., Rao, J. N. K., Särndal, C. E. y Singh, M. P. (Eds.) (1987). *Small Area Statistics: An International Symposium.* New York; John Wiley and Sons.

Prasad, N. G. N. y Rao, J. N. K. (1990). «The estimation of Mean Squared Error of Small-Area Estimators», *Journal of the American Statistical Association*, 85, 163-171.

Raghunathan, T. E. (1993). «A Quasi-empirical Bayes Method for Small Area Estimation», *Journal of the American Statistical Association*, 88 (424), 1444-48.

Rao, C. R. (1973). *Linear statistical inferece and its applications*, 2nd edn., New York: John Wiley.

Singh, M. P., Gambino, J. y Mantel, H. J. (1994). «Issues and Strategies for Small Area Data», *Survey Methodology*, 20, 1, Statistics Canada, 3-22.

Singh, A. C., Mantel, H. J. y Thomas, B. W. (1994). «Time Series EBLUPs for Small Areas Using Survey Data», *Survey Methodology*, 20, 1, Canada Statistics, 33-43.

Singh, A. C., Stukel, D. M. y Pfeffermann, D. (1998). «Bayesian versus frequentist measures of error in small area estimation», *Journal of the Royal Statistical Society*, B, 60, 377-396.

Thomas, N., Longford, N. T. y Rolph, J. E. (1994). «Empirical Bayes Methods for Estimating Hospital-Specific Mortality Rates», *Statistics in Medicine.*

APÉNDICE

A. TABLAS ESTADÍSTICAS DE LOS ESTIMADORES

Este apéndice muestra las tablas de los estimadores directos e indirectos a partir de los cuales hemos calculado nuestro estimador compuesto.

Tasas de variación. Estimadores DIRECTOS por Comunidad. Sector Industrial

AÑO	TRIM	AND	ARA	AST	BAL	CAN	CANT	CAS-L	CAS-M	CAT	VAL	EXT	GAL	MAD	MUR	NAV	PVAS	RIO	VARIANZA
1987	3	2,496	-2,837	-3,456	0,872	3,928	2,997	-1,161	5,087	0,186	-1,386	-5,048	2,546	-0,179	4,485	-1,966	2,109	5,386	9,901
1987	4	2,928	-4,259	0,524	1,848	-10,920	2,613	-0,348	3,172	0,836	3,287	-1,410	2,674	4,853	-4,591	0,175	-1,804	5,975	16,633
1988	1	0,305	2,787	-3,550	2,136	13,548	-4,043	-4,183	1,982	-2,252	4,327	7,435	-3,541	-7,453	5,896	-8,508	-0,537	0,390	29,806
1988	2	1,284	5,984	5,703	1,981	5,026	1,313	6,197	1,734	2,985	-3,360	-4,068	2,157	1,407	7,774	4,394	-0,442	-4,133	12,935
1988	3	-3,153	-1,285	-1,491	5,589	2,892	3,268	-4,542	-4,454	1,835	-0,683	4,758	2,545	-0,487	0,949	-3,449	-5,109	-1,383	12,641
1988	4	-0,044	-2,506	3,207	-3,496	-3,802	-6,276	4,300	4,984	-1,532	0,165	-5,562	-2,336	7,742	-7,017	-0,368	1,164	1,121	20,000
1989	1	2,515	2,551	-5,758	-3,517	4,204	5,441	2,120	-0,760	3,483	-1,655	-4,327	2,546	0,897	-0,277	-0,644	-0,217	-0,650	45,489
1989	2	0,283	0,886	5,226	-0,395	0,867	0,688	-1,158	-1,299	1,284	4,143	18,122	1,526	-4,998	-0,825	13,857	5,216	5,867	30,989
1989	3	0,498	-0,243	-0,773	-2,026	-3,360	1,051	3,431	-1,029	2,945	2,647	10,927	1,883	-0,102	-5,407	-3,093	-0,777	-4,381	21,925
1989	4	2,832	0,993	-0,623	-3,495	0,911	-1,116	3,763	-0,766	1,012	4,756	-4,746	-4,842	-1,916	-0,591	0,924	2,126	-0,429	24,803
1990	1	3,130	1,530	-1,399	-0,932	-3,466	1,078	2,428	0,151	0,987	1,150	11,604	3,120	2,444	4,829	-4,944	0,809	-1,977	25,287
1990	2	1,834	-3,287	-0,136	-5,812	3,165	0,312	-0,594	0,000	-0,242	3,228	6,972	1,043	-0,197	-1,970	4,886	-2,803	6,837	12,590
1990	3	2,304	1,520	-3,444	-5,421	2,471	-3,320	-0,588	3,093	-1,079	-2,291	0,429	-1,190	-2,991	1,287	1,186	4,419	2,270	10,689
1990	4	0,010	-1,377	0,318	-7,264	-1,929	1,258	-0,913	2,871	0,356	-1,778	-10,959	0,173	-1,348	3,226	-6,386	-0,166	-3,329	28,952
1991	1	-0,841	1,801	-2,382	9,228	-3,852	2,596	-2,159	3,084	-2,551	-5,100	5,563	-4,093	-5,467	-2,973	-1,974	1,375	1,835	177,515
1991	2	3,091	-1,075	-3,807	11,266	-14,663	-5,241	0,838	-3,486	-1,441	-1,144	-3,422	6,442	0,858	-2,052	5,197	-0,066	1,574	35,484
1991	3	-2,934	-1,817	-1,701	-3,094	5,395	0,082	0,298	-1,957	0,131	0,167	4,359	-0,645	2,449	-7,261	1,026	0,052	-3,875	25,636
1991	4	0,267	-2,919	1,107	-7,958	-10,379	4,574	3,758	2,613	-2,195	-0,394	-1,593	4,080	-1,414	7,887	7,986	-0,813	0,105	51,283
1992	1	-1,000	2,486	-3,210	4,048	8,134	1,718	-5,822	-5,824	-1,511	1,848	2,048	-1,696	2,236	-6,088	-4,340	-1,287	-2,681	140,095
1992	2	0,397	-0,987	-0,976	-8,485	3,039	-4,991	-1,512	-1,215	-0,726	-1,488	-11,520	-0,142	4,269	1,284	8,470	1,433	1,619	37,932
1992	3	-4,043	-2,506	-6,560	-0,138	-12,718	-1,401	-1,932	0,884	-2,111	-1,788	0,440	1,253	-1,027	-0,779	-2,338	-8,736	-7,858	113,201
1992	4	-7,664	-1,281	-0,155	2,680	8,006	-4,016	-5,157	5,683	-3,358	-0,107	-3,670	-7,199	-5,700	-2,787	-3,245	-5,936	-2,343	43,085
1993	1	-7,496	-5,523	-3,914	2,234	-12,619	-0,114	-3,872	-6,736	-1,310	-4,842	-4,928	-3,829	-4,959	-0,824	-13,514	-6,415	-1,928	16,628
1993	2	-2,141	-5,270	0,352	3,027	4,139	-10,573	-1,915	-4,944	-0,726	-2,146	12,904	-1,176	0,181	-2,827	2,428	1,258	3,850	112,039
1993	3	1,351	2,417	-12,018	0,230	-0,113	2,516	2,850	-0,455	-3,324	1,152	-4,103	-1,578	0,826	-1,446	2,500	-2,754	3,135	66,895
1993	4	-2,178	4,269	-3,386	-8,398	-4,797	-4,134	2,416	-0,336	-5,591	1,527	-3,328	-4,986	-1,800	2,201	-1,078	0,777	-6,304	19,499
1994	1	0,852	3,421	4,587	-1,716	9,810	-3,729	-3,900	2,631	0,408	-1,181	1,054	-3,569	-3,786	3,508	1,450	-4,898	-1,241	20,663
1994	2	1,259	1,303	-0,345	-3,796	-13,792	1,819	0,393	-5,510	-0,353	2,674	-4,206	1,234	-0,333	2,774	-2,680	0,170	7,583	23,629
1994	3	-0,874	-2,331	-0,412	5,816	2,943	0,132	0,777	0,077	-1,280	4,580	-6,515	2,467	-2,122	-5,537	1,852	2,398	-0,038	63,332
1994	4	0,496	4,197	-1,374	-2,318	0,812	13,842	-0,218	-0,772	0,123	3,217	-4,071	-4,506	0,191	3,475	-0,722	1,174	-2,903	54,483
1995	1	-3,601	-3,208	4,987	-6,840	13,623	2,147	-4,772	-4,248	2,887	-5,207	8,887	3,171	-4,449	-3,546	-0,109	-1,797	6,524	33,269
1995	2	0,195	2,203	7,930	-2,397	16,117	-1,278	-3,245	-3,544	0,529	2,188	-9,934	3,741	1,163	4,490	6,829	0,482	4,448	31,017
1995	3	1,794	-1,427	-8,206	-1,934	-5,345	1,466	2,315	-6,074	0,887	0,410	7,998	2,912	6,351	-9,284	0,170	-2,901	0,583	49,270
1995	4	2,166	-2,852	5,567	4,508	7,568	-2,240	4,569	4,897	-1,788	-2,787	4,148	-2,725	-0,017	5,216	-9,309	-0,405	-2,982	18,715
1996	1	-1,864	1,955	3,317	9,467	-5,867	-4,642	-2,573	-3,130	-1,449	-4,527	6,022	-0,621	3,297	-5,365	-0,131	-2,374	2,252	16,997
1996	2	0,983	1,561	-0,473	1,478	4,897	-3,661	4,406	-1,288	-1,408	1,845	-0,642	0,508	-0,837	2,464	4,585	-0,358	5,942	43,953
1996	3	0,618	2,159	4,182	7,929	3,988	-4,294	0,491	0,547	2,771	0,756	-7,074	5,507	3,875	-2,977	-3,755	7,887	-1,254	16,774
1996	4	1,808	-2,143	-5,778	-2,374	-5,808	7,258	1,940	-0,555	1,818	-2,883	-4,519	0,308	-0,886	-0,824	5,800	-0,950	-4,210	180,002
1997	1	-4,383	3,104	-0,379	-1,638	-4,437	-2,399	-7,134	3,106	-0,008	-3,879	5,242	0,990	-1,928	-1,552	2,889	2,515	-2,816	12,324
1997	2	-1,810	3,586	-4,514	-4,372	2,357	2,175	-2,014	-0,678	3,661	0,541	7,160	4,632	1,382	5,811	8,597	2,123	1,003	50,921
1997	3	6,286	4,704	-1,289	-7,483	3,303	7,187	3,942	6,570	2,673	6,883	-3,561	-0,989	-0,496	4,805	2,859	3,132	2,270	29,284
1997	4	6,000	-0,521	0,371	-3,941	-1,848	0,518	1,977	-0,113	1,266	2,856	5,422	-0,653	2,470	5,987	-0,738	-0,850	-1,387	38,088
1998	1	-2,822	-3,342	-2,378	-1,317	-8,282	1,918	2,563	0,698	-3,219	2,732	-8,887	2,280	-2,188	1,193	-2,720	-3,240	-1,884	17,915
1998	2	6,166	6,495	0,889	1,789	5,778	-2,051	-0,114	2,723	4,805	1,083	0,487	2,750	2,438	-1,000	1,498	3,541	6,843	44,605
1998	3	0,794	-3,466	8,434	-4,604	-1,015	3,011	4,237	9,113	1,252	1,684	-2,421	4,377	1,855	-0,528	1,890	-0,986	-3,924	29,051
1998	4	5,226	-1,536	-3,791	7,670	4,805	6,208	3,440	2,447	-1,508	-2,141	8,898	0,377	0,854	0,727	-3,003	-1,206	-5,550	44,726
1999	1	-0,937	-1,311	0,907	14,812	-2,789	0,472	-2,672	-2,016	-1,285	1,217	7,715	-1,934	-1,590	-0,045	-0,324	3,886	7,797	26,181
1999	2	2,788	2,810	-0,217	-4,085	-0,047	1,383	-0,102	5,186	0,147	-3,093	4,231	3,445	3,297	-0,271	-0,867	2,590	9,050	10,543
1999	3	1,559	2,288	6,647	4,609	-7,046	-1,390	0,544	3,697	0,880	-0,973	4,233	2,185	0,170	-3,034	1,508	2,180	-1,228	51,920
1999	4	-4,019	-0,125	-1,980	-7,885	13,218	4,697	0,975	2,801	-2,084	4,076	-2,114	-3,208	1,258	12,593	5,596	1,982	-2,482	32,536
2000	1	2,844	7,383	3,922	-0,224	-0,563	0,773	1,155	-2,527	0,649	-0,039	2,415	2,469	-3,631	-1,341	-1,863	-0,485	3,264	47,621
2000	2	0,522	3,541	2,087	-5,355	4,918	-8,184	4,797	-5,675	1,818	1,703	0,388	1,144	2,788	1,892	5,384	1,476	5,942	45,204
VARIANZA		8,869	8,592	16,959	28,563	52,785	17,222	8,315	11,976	4,419	8,389	40,388	8,812	8,784	19,522	21,265	7,566	16,691	41,966
PESOS		0,210	0,228	0,404	0,680	1,257	0,410	0,222	0,285	0,105	0,198	0,962	0,210	0,233	0,460	0,505	0,180	0,431	

Cuadro 1. Estimadores directos

Tasas de variación. Estimadores SINTÉTICOS por Comunidad. Sector Industrial

AÑO	TRIM	AND	ARA	AST	BAL	CAN	CANT	CAS-L	CAS-M	CAT	VAL	EXT	GAL	MAD	MUR	NAV	PVAS	RIO
1987	3	0,504	0,525	0,521	0,272	0,516	0,788	0,331	0,882	0,879	0,358	0,827	0,503	0,661	0,734	0,802	0,877	0,528
1987	4	0,770	2,228	1,251	-0,034	-0,418	1,143	1,202	0,745	1,458	0,959	-0,084	0,876	1,370	1,244	1,840	2,284	1,067
1988	1	-0,979	-1,814	-1,876	0,186	-0,917	-1,751	-1,686	-0,307	-1,647	-0,008	-0,738	-0,860	-2,346	-1,079	-1,972	-2,191	-0,758
1988	2	2,195	1,790	3,614	1,987	2,896	2,500	2,588	1,577	2,073	1,356	2,358	2,045	2,676	2,100	2,115	2,453	1,548
1988	3	0,356	-0,032	0,616	-0,364	0,965	0,380	-0,052	-0,858	-1,121	-1,104	0,612	-0,209	-1,069	0,408	0,352	0,343	-0,361
1988	4	0,302	0,973	-0,265	0,724	0,441	0,409	-0,012	0,043	0,893	0,764	-0,527	0,172	1,941	-0,192	1,002	1,280	0,121
1989	1	1,274	0,326	-0,031	1,062	1,615	1,552	1,121	1,661	1,682	0,887	1,679	1,126	1,490	1,552	0,493	0,752	1,015
1989	2	1,472	1,380	1,319	1,503	1,772	1,195	1,853	0,998	1,243	1,079	1,235	1,920	1,727	1,139	1,459	0,775	1,045
1989	3	0,567	1,244	1,557	0,710	0,269	0,760	0,254	0,821	1,126	1,099	0,356	0,321	1,845	0,464	0,980	1,368	0,523
1989	4	0,276	0,957	-0,427	0,361	-0,028	0,961	0,204	0,754	1,321	1,069	-0,149	0,262	0,971	0,407	1,116	2,240	0,736
1990	1	1,888	2,005	0,641	1,949	2,105	1,459	2,046	1,234	1,367	1,414	1,750	2,105	1,327	1,832	1,967	1,321	1,919
1990	2	0,236	0,313	0,076	-0,530	0,168	0,580	0,569	0,436	0,710	-0,143	0,413	0,163	0,460	0,071	0,328	0,065	0,266
1990	3	0,132	-0,233	0,424	0,262	0,851	-0,097	0,216	-0,884	-0,479	-0,728	-0,170	-0,389	0,258	-0,285	-0,220	-0,076	-0,399
1990	4	-0,946	-0,365	-1,315	-0,601	-1,071	-0,901	-0,973	-0,951	0,001	-0,384	-1,519	-1,282	0,660	-0,559	-0,342	0,156	-0,471
1991	1	-2,442	-2,954	-2,091	-1,347	-2,789	-2,729	-2,899	-1,016	-2,205	-0,959	-1,709	-2,205	-2,920	-2,125	-3,206	-3,029	-1,754
1991	2	0,038	0,154	0,463	-0,724	0,219	0,198	0,320	-0,750	-0,191	-0,959	-0,201	-0,136	0,392	-0,312	0,267	-0,142	-0,443
1991	3	0,152	-0,229	-0,417	0,436	0,879	-0,054	-0,125	-0,853	-0,342	-0,246	-0,121	0,024	0,725	0,063	0,159	-0,030	-0,368
1991	4	-0,371	0,632	0,633	-1,212	-0,809	0,148	0,242	-1,302	-0,325	-0,632	-1,604	-0,395	0,241	-0,481	0,951	1,322	-0,431
1992	1	-1,154	-0,706	-1,172	-0,489	-1,417	-1,143	-1,149	-0,714	-0,904	0,152	-1,467	-0,581	-1,051	-1,277	-0,414	-0,457	-0,537
1992	2	0,683	0,319	-0,012	0,445	1,428	0,568	0,648	-0,094	-0,129	-0,079	1,007	0,482	-0,122	0,606	0,829	0,293	0,665
1992	3	-1,999	-1,562	-1,655	-2,709	-2,065	-1,875	-1,865	-2,764	-2,373	-2,688	-2,323	-2,390	-2,211	-2,015	-1,485	-1,531	-2,057
1992	4	-3,117	-4,079	-3,448	-3,075	-2,684	-2,942	-3,510	-2,629	-3,108	-3,355	-2,163	-3,541	-3,340	-2,842	-4,063	-3,757	-3,157
1993	1	-3,847	-3,489	-4,127	-3,250	-4,233	-5,059	-3,574	-3,959	-3,986	-3,928	-4,135	-3,048	-3,745	-3,390	-4,295	-5,187	-4,322
1993	2	-0,053	-0,962	-3,058	-0,539	0,661	-0,657	-0,434	-0,994	-0,804	-1,324	-0,159	-0,885	-0,963	0,265	-0,655	-1,381	-0,134
1993	3	-0,093	-1,436	-1,030	0,080	0,770	-0,833	-0,586	-0,530	-1,188	-0,856	0,583	-0,768	-1,019	0,024	-1,301	-2,124	-0,230
1993	4	-1,979	-1,455	-2,047	-2,304	-1,623	-1,440	-1,762	-2,740	-1,708	-2,838	-2,411	-1,778	-0,886	-2,465	-1,460	-0,988	-2,895
1994	1	-0,891	-1,171	-0,224	-1,026	-1,617	-0,953	-0,940	0,028	-0,586	-0,391	-0,925	-1,039	-1,286	-0,738	-1,614	-1,308	-0,809
1994	2	-0,056	0,347	0,491	0,463	-0,251	-0,265	0,120	0,024	-0,145	0,395	0,000	0,221	-0,424	-0,131	0,131	0,122	0,365
1994	3	0,508	0,150	-0,861	0,623	0,712	-0,160	0,126	0,430	0,146	0,390	0,766	-0,128	0,189	1,115	-0,072	-0,479	0,839
1994	4	-0,210	0,511	-0,158	0,965	-0,923	0,031	-0,567	1,030	1,058	1,502	-0,460	0,349	0,738	-0,351	0,103	1,061	0,216
1995	1	-1,153	-1,137	-0,434	-1,400	-1,168	-0,696	-0,720	-1,406	-1,240	-1,384	-1,257	-0,973	-1,493	-1,389	-0,881	-0,539	-1,104
1995	2	1,235	1,466	3,551	1,071	1,163	1,303	1,698	1,252	1,089	1,187	1,368	1,446	1,410	1,242	1,311	1,497	0,817
1995	3	1,812	-0,194	2,201	1,761	4,068	1,164	1,265	-0,198	-0,331	-0,375	2,249	0,262	0,925	1,181	0,609	0,125	0,612
1995	4	-0,190	-0,322	1,736	-0,539	-0,599	0,025	0,186	-0,219	-0,404	-0,739	0,292	0,238	-0,975	0,325	-0,236	-0,078	-0,307
1996	1	-1,150	-0,867	-0,523	-1,512	-0,804	-0,895	-0,671	-2,097	-1,291	-1,906	-1,719	-0,903	-0,565	-1,395	-0,526	-0,400	-1,849
1996	2	0,382	0,862	-0,168	-0,028	-0,243	0,046	0,346	0,832	0,762	0,572	0,263	0,183	0,975	0,810	0,117	-0,250	0,423
1996	3	1,800	3,314	1,573	1,106	0,628	2,036	1,967	2,236	2,773	2,109	1,446	2,340	2,294	1,978	2,963	2,564	2,144
1996	4	-0,288	-0,473	-0,089	-0,006	-0,413	-0,079	-0,137	0,245	0,118	0,418	-0,127	-0,228	-0,344	-0,056	-0,192	0,117	0,417
1997	1	-1,692	-0,146	-2,308	-1,150	-1,943	-1,444	-1,479	-2,047	-0,929	-1,177	-2,382	-0,789	0,233	-1,638	-0,395	-0,117	-1,791
1997	2	1,522	1,751	0,188	1,151	0,961	1,572	1,339	2,005	2,157	1,773	1,288	1,302	1,624	1,904	1,575	1,636	1,915
1997	3	2,972	3,351	2,504	2,993	3,369	3,878	3,026	2,534	3,113	3,075	2,973	2,934	3,060	3,018	4,077	5,029	3,517
1997	4	1,947	1,285	3,675	1,545	2,465	2,415	2,227	1,674	1,437	1,385	2,088	1,759	1,505	1,289	1,659	2,040	1,324
1998	1	-1,429	-0,960	-1,631	-1,845	-1,560	-1,257	-1,395	-1,950	-1,413	-1,655	-2,088	-1,845	-0,807	-1,490	-1,071	-0,806	-1,527
1998	2	3,258	3,062	0,933	3,593	3,863	3,132	2,677	2,722	3,415	3,061	3,191	2,800	4,192	3,647	3,436	3,291	3,520
1998	3	1,409	1,561	1,533	1,355	1,736	1,974	1,396	1,274	1,605	1,442	1,330	1,409	1,612	1,051	2,020	2,533	1,356
1998	4	-0,085	0,683	1,431	-0,070	-0,443	0,045	0,228	0,075	0,129	0,064	-0,386	1,021	0,597	-0,812	0,362	0,078	-1,060
1999	1	-0,127	-1,058	-0,106	0,346	0,188	0,010	-0,523	0,710	-0,038	0,435	0,704	-0,288	-0,469	0,002	-0,814	-0,542	0,192
1999	2	1,100	1,338	2,224	1,305	0,863	0,647	1,032	0,543	0,689	0,733	0,602	0,876	1,408	1,082	0,742	0,908	0,537
1999	3	0,471	1,405	0,915	-0,079	-0,263	0,989	0,960	0,774	0,948	0,929	0,053	1,181	0,447	0,508	1,524	1,665	0,642
1999	4	0,411	0,238	0,452	0,049	0,142	0,634	0,280	0,593	0,154	0,545	0,258	0,607	-0,365	0,440	0,546	0,631	0,275
2000	1	0,677	0,461	-0,452	1,174	1,191	0,404	0,691	0,078	0,429	0,326	0,538	0,661	0,735	0,628	0,485	0,482	0,627
2000	2	2,200	1,428	1,533	1,990	3,009	2,053	2,295	1,335	1,348	1,414	2,248	1,658	1,755	2,386	1,824	1,748	2,082

Cuadro 2. Estimadores indirectos

Tasas variación España, por RAMA (en porcentaje)

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

Cuadro 3. Estimador directo por ramas de actividad industrial. Ámbito estatal

Ponderaciones Industria 97

	AND	ARA	AST	BAL	CAN	CANT	CAS-L	CAS-M	CAT	VAL	EXT	GAL	MAD	MUR	NAV	PVAS	RIO
Rama 1	2,83%	2,51%	21,15%	0,93%	1,54%	2,35%	2,31%	6,36%	0,50%	0,66%	3,60%	2,80%	1,25%	1,84%	0,77%	0,31%	0,89%
Rama 2	4,74%	1,93%	2,42%	9,01%	8,79%	2,65%	1,45%	2,85%	1,98%	1,79%	4,68%	3,65%	4,69%	2,40%	0,92%	1,26%	2,21%
Rama 3	24,59%	11,90%	13,75%	18,63%	33,19%	20,29%	17,35%	22,58%	12,41%	10,50%	33,09%	18,02%	10,08%	27,86%	15,82%	23,58%	6,48%
Rama 4	10,00%	11,22%	3,32%	18,94%	2,20%	5,29%	26,41%	5,70%	16,89%	25,83%	15,83%	13,09%	6,64%	12,73%	4,76%	19,81%	1,88%
Rama 5	3,55%	2,80%	4,38%	5,59%	5,27%	4,41%	5,38%	5,36%	2,79%	4,83%	7,19%	9,13%	1,64%	3,39%	3,69%	2,52%	2,87%
Rama 6	5,26%	4,55%	4,68%	9,01%	9,89%	4,12%	2,39%	4,77%	8,30%	5,01%	3,96%	3,10%	16,82%	4,10%	6,14%	4,72%	5,35%
Rama 7	3,79%	1,84%	1,96%	0,31%	1,54%	6,47%	6,75%	3,31%	8,59%	1,95%	2,52%	3,16%	6,55%	4,24%	1,38%	0,94%	4,56%
Rama 8	2,23%	3,19%	1,21%	1,55%	1,98%	4,71%	1,79%	6,23%	5,00%	3,11%	0,72%	2,01%	2,89%	3,39%	4,61%	6,29%	8,13%
Rama 9	8,81%	4,55%	7,85%	7,14%	10,99%	7,35%	8,89%	7,75%	4,04%	9,95%	6,83%	7,06%	3,95%	4,95%	6,61%	6,60%	3,62%
Rama 10	11,24%	10,54%	23,11%	12,11%	12,09%	18,82%	10,68%	10,33%	10,81%	13,51%	11,51%	10,59%	9,56%	9,76%	15,82%	12,58%	25,74%
Rama 11	3,39%	9,86%	5,29%	2,48%	1,76%	6,18%	3,25%	3,44%	6,78%	4,27%	3,60%	3,47%	5,17%	6,08%	10,29%	5,68%	14,61%
Rama 12	3,19%	9,19%	1,96%	1,55%	1,98%	5,59%	3,16%	2,98%	7,58%	3,01%	1,80%	2,68%	12,84%	1,70%	5,99%	1,57%	7,61%
Rama 13	8,85%	19,15%	5,14%	2,48%	3,96%	8,24%	2,56%	13,71%	8,90%	4,88%	0,72%	16,80%	10,78%	5,37%	19,20%	5,66%	9,86%
Rama 14	7,53%	6,77%	3,78%	10,25%	4,84%	3,53%	7,61%	4,64%	5,44%	10,69%	3,96%	4,44%	7,16%	12,16%	3,99%	8,49%	6,20%
	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%

Cuadro 4. Ponderaciones de las ramas de actividad industrial por Comunidad

B. EQUIVALENCIAS EN LA CLASIFICACIÓN DE LAS RAMAS INDUSTRIALES

Clasificación CRE	Clasificación CNAE 93
Extracción de productos energéticos, otros minerales y refino de petróleo	B31+B35+B5D
Energía eléctrica, gas y agua	B70
Alimentación, bebidas y tabaco	B41
Textil, confección, cuero y calzado	B44+B47
Madera y corcho	B49
Papel; edición y artes gráficas	B4B
Industria química	B5F
Caucho y plástico	B5H
Otros productos minerales no metálicos	B5J
Metalurgia y productos metálicos	B5L
Maquinaria y equipo mecánico	B5O
Equipo eléctrico, electrónico y óptico	B5Q
Fabricación de material de transporte	B60
Industrias manufactureras diversas	B63

donde, en la clasificación CNAE:

B31: Extracción de productos energéticos
B35: Extracción de otros minerales excepto productos energéticos
B41: Alimentación, Bebidas y Tabaco
B44: Textiles, Confeccion
B47: Cuero y calzado
B49: Madera, Corcho
B4B: Papel, Artes Gráficas y Edicion
B5D: Coquerias, Refino de petroleo
B5F: Industrias Quimicas
B5H: Caucho y Materias Plasticas
B5J: Otros productos minerales no metálicos
B5L: Metalurgia y fabricación de productos metálicos
B5O: Maquinaria y equipo mecánico
B5Q: Material y equipo electrico, electrónico
B60: Material de transporte
B63: Otras Industrias Manufactureras
B70: Energia electrica, gas y agua

C. DERIVACIÓN DE LA FÓRMULA DEL ERROR MUESTRAL

En este apéndice relacionaremos el error de muestreo relativo proporcionado por el INE para la encuesta de la EPA y la varianza muestral $\sigma_2^2(k)$.

De lo reseñado anteriormente en la Sección 4, obtenemos

$$\widehat{\theta}_2(k,t) = 100 \times \frac{o(k,t) - o(k,t-1)}{o(k,t-1)},$$

donde $o(k,t)$ denota el número de ocupados en la muestra de la comunidad k en el instante t (para el sector industrial concreto). La información que obtenemos de los documentos del INE sobre la EPA es la relativa al coeficiente de variación de la ocupación, es decir una estimación del valor C.V. $= \sigma_o/\mu_0$, donde μ_0 y σ_o son respectivamente la media y la desviación estándar de la variable ocupación para la comunidad k en el momento t. Introduciendo la proporción muestral $\widehat{p}(k,t) = o(k,t)/N$, con un tamaño muestral N que suponemos constante, obtenemos

$$\widehat{\theta}_2(k,t) = 100 \times \frac{\widehat{p}(k,t) - \widehat{p}(k,t-1)}{\widehat{p}(k,t-1)}$$

cuya varianza $\sigma_2^2(k,t)$, inducida por la varianza muestral de $\widehat{p}(k,t)$, es necesario determinar.

Dado que $\widehat{p}(k,t)$ es una proporción muestral, obtenemos *bajo muestreo aleatorio simple* que $var(\widehat{p}(k,t)) = p(k,t)(1-p(k,t))/N$, donde $p(k,t)$ indica la proporción poblacional. Dado que $\widehat{p}(k,t)$ converge en probabilidad a $p(k,t)$ cuando el tamaño muestral N tiende a infinito, utilizando el método-δ (Rao, 1973) obtenemos

$$\begin{aligned} avar(\widehat{\theta}_2(k,t)) &= \left(\frac{1}{p(k,t)}, -\frac{p(k,t)}{p(k,t-1)^2}\right) \left[var\begin{pmatrix}\widehat{p}(k,t)\\ \widehat{p}(k,t-1)\end{pmatrix}\right] \left(\frac{1}{p(k,t)}, -\frac{p(k,t)}{p(k,t-1)^2}\right)' \\ &= \frac{1}{N}\left(\frac{1}{p(k,t)}, -\frac{p(k,t)}{p(k,t-1)^2}\right) \begin{pmatrix} p(k,t)(1-p(k,t)) & 0 \\ 0 & p(k,t-1)(1-p(k,t-1)) \end{pmatrix} \\ &\times \left(\frac{1}{p(k,t)}, -\frac{p(k,t)}{p(k,t-1)^2}\right)' = \frac{2}{N}\frac{1-p(k)}{p(k)} \end{aligned}$$

para todo t, donde *avar* denota varianza asintótica. En el cálculo de dicha expresión, hemos sustituido $p(k,t)$ por su valor promedio en cada Comunidad, pongamos $p(k)$ (en este punto utilizamos el supuesto de que $p(k,t)$ es estacionario en media) y se ha considerado que las muestras en dos periodos consecutivos son independientes. En el caso que este último supuesto no sea correcto, entonces la expresión anterior se modifica para obtener:

$$avar(\widehat{\theta}_2(k,t)) = E.D. \times \frac{2}{N} \times \frac{1-p(k)}{p(k)} \times (1-\rho(k))$$

donde $\rho(k)$ es el coeficiente de la correlación muestral en dos periodos sucesivos. La EPA utiliza un diseño de muestra complejo, por lo tanto la expresión derivada bajo el supuesto de muestreo aleatorio simple debe ser modificada multiplicando por el efecto de diseño E.D.

Por otra parte, en el caso de muestreo complejo con efecto de diseño E.D., obtenemos que

$$\text{C.V.}(o(k,t)) = \widehat{\sigma}_o/\widehat{\mu}_0 = \sqrt{\frac{var(\widehat{p}(k,t))}{\widehat{p}(k,t)^2}} = \sqrt{\text{E.D.}} \times \sqrt{\frac{\widehat{p}(k,t)(1-\widehat{p}(k,t))}{N\widehat{p}(k,t)^2}}$$

$$= \sqrt{\text{E.D.}} \times \sqrt{\frac{(1-\widehat{p}(k,t))}{N\widehat{p}(k,t)}}$$

de modo que al comparar las expresiones de la varianza muestral de $\widehat{\theta}_2(k,t)$ y el coeficiente de variación anterior, obtenemos la fórmula simple:

$$avar(\widehat{\theta}_2(k,t)) = 2 \times \text{C.V.}^2(k,t) \times (1-\rho(k))$$

ENGLISH SUMMARY

COMPOSITE ESTIMATORS IN REGIONAL STATISTICS: AN APPLICATION TO THE ESTIMATION OF THE RATE OF CHANGE IN INDUSTRIAL OCCUPATION*

À. COSTA[1]
A. SATORRA[2]
E. VENTURA[2]

This work is part of a project that studies the application of small area composite estimators (a combination of direct and indirect estimators) in regional statistics. We compare three estimators: a direct one based on sample data for each of the Spanish Autonomous Communities, a synthetic (indirect) one that combines Estate wide information with the Community specific information, and a third estimator, the composite one, which is based on a model and results in a linear combination of the two previous estimators. We show the adopted method of analysis by estimating the industrial occupation rate of change in the different Spanish Autonomous Communities.

Keywords: Regional statistics, small areas, composite estimators

AMS Classification (MSC 2000): 62J07, 62J10, 62H12

*This article is a result from a joint research project between the Institut d'Estadística de Catalunya (IDESCAT) and the Instituto Nacional de Estadística (INE). The INE and the IDESCAT do not necessarilly share the opinions expressed in this study, which belong exclussively to the authors. We thank the comments of N. T. Longford to a preliminary version of this article.

[1] Institut d'Estadística de Catalunya. Via Laietana 58, 08003 Barcelona.

[2] Departament d'Economia i Empresa. Universitat Pompeu Fabra. Ramon Trias Fargas 25-27, 08005 Barcelona.

–Received December 2001.
–Accepted March 2002.

Most of the surveys that are carried out by the national statistics offices have been designed to yield estimators with small variance for national wide figures. The size and design of the sample may still allow us to get precise estimators in cases in which the areas under consideration are smaller, such as regions or provinces. But in some cases the size of the sample is just too small for that purpose. It is even worst when one wants to estimate the rate of variation instead of the level of a variable, since rates have larger variances. This is quite a frequent situation in economics statistics.

There are two alternatives which allow us to deal with the issue of providing valid statistical information for the smallest areas. The first one requires to repeat the survey with a new and more suitable sample size. The second one is to use statistical theory on small areas estimation to improve the quality of the estimators obtained from the original survey.

In this article we take the second alternative. If part of the national sample has been extracted from a particular small area, we can obtain a direct estimation of the variable of interest for that area. Such estimator is unbiased but often lacks precision. On the other hand the survey allows us to obtain an estimator which is unbiased and precise at the national level. We can use it to create an indirect estimator for the small area which will exhibit a smaller variance than the direct estimator but will usually be biased.

Here we propose to use a composite estimator. First we specify a model of variance behavior of the direct and synthetic estimators, and then we consider a linear combination of both estimators as our composite estimator. The weights of the linear combination are chosen so as to minimize the mean square error of the composite estimator. In this sense, the composite estimator represents an improvement relative to alternative estimators, either direct or synthetic (indirect) for small areas.

The method proposed is applied to data of the Encuesta de Población Activa, a labor survey conducted at the national Spanish level. The quantity to be estimated is the quarterly variation of the occupation rate in the industrial sector in each of the 17 Autonomous Communities of Spain. This rate of variation has multiple components as it depends on the geographical area and the time period being considered.

Our composite estimator assigns automatically larger weight to the direct estimator in those Communities where the variance of the sampling error is smaller than the bias of the indirect estimator, and larger weight to the synthetic estimator when bias is small relative to the variance of the sampling error. We observe that for most of the Communities, the composite estimator gives more weight to the synthetic estimator, given that the size of the sample in the small area is quite insufficient and therefore the sampling error has a large variance. But there are also some Communities in which the composite estimator assigns higher weight to the synthetic estimator. This is the case of Catalonia, whose industrial web is very similar to the national wide one, and the difference between the synthetic and the direct estimators is very small in this Community.

From this study it is realized that the weights determining the composite estimator depend on the bias, the population and sample variances, the type of variable being estimated (levels or rates), and other elements which are determinant of the way in which the estimator combines the available information. It is therefore important to explore, through different scenarios, the way in which all those elements affect the behavior of the composite estimator. In future work we plan to investigate the sample properties of this composite estimators. The study may have useful implications for the sample design and, particularly, for the sample distribution among the different surveys' subdomains if it is decided that composite instead of direct estimators will be used for estimation in the small areas.

Biometria

QÜESTIIÓ, vol. 26, 1-2, p. 247-258, 2002

MAINTENANCE POLICY UNDER MULTIPLE UNREVEALED FAILURES

F. G. BADÍA BLASCO
M. D. BERRADE URSÚA
CLEMENTE A. CAMPOS
Universidad de Zaragoza*

The unrevealed failures of a system are detected only by inspection. In this work, an inspection policy along with a maintenance procedure for multi-unit systems with dependent times to failure is presented. The existence of an optimum policy is also discussed.

Keywords: Maintenance, reliability, replacement, optimum policy

AMS Classification (MSC 2000): 90B25, 60K10, 62N05

*Departamento de Métodos Estadísticos. Centro Politécnico Superior de Ingenieros. Universidad de Zaragoza. María de Luna, 3. 50015 Zaragoza.
Corresponding author: Clemente A. Campos. Centro Politécnico Superior de Ingenieros. Universidad de Zaragoza. María de Luna, 3. 50015 Zaragoza. SPAIN. E-mail address: C.Campos@posta.unizar.es. Phone: +34 976 761 884. Fax: +34 976 761 861.

–Received November 2001.
–Accepted April 2002.

1. INTRODUCTION

The classical age and block replacement policies are only useful when the failures of a unit are detected as soon as they occur. However, the failures of spare units or in a system during the stand-by mode, may remain undiscovered until the next demand of activity, causing important availability losses. Periodic inspection or testing is the only way to fight against failures of this sort (see Lewis (1987)).

The works due to Nakagawa and Yasui (1991), Vaurio (1995, 97, 99) and Berrade (1999) deal with unrevealed failures; all of them consider maintenance policies for single-unit systems. On the contrary, Ebrahimi (1997) provides an extension of the age replacement policy for multi-unit systems under revealed failures.

In this work we provide a new inspection policy to detect the occurrence of failures in a system, otherwise being unrevealed, along with a maintenance procedure to improve its reliability. This policy can be carried out in multi-unit systems with an only type of failure, or in single-unit systems with multiple «hidden» causes of failure. Many systems of this sort can be found in practice: the fluid levels in a car are checked periodically so that an eventual leakage of oil, water, etc can be observed. The inspection policies are concerned not only with engineering systems: the periodic tests carried out to detect disorders of health in humans constitute the most common example of such procedure.

We assume the possibility of dependent times to failure between units or types of failure which is a realistic assumption when the occurrence of a type of failure is likely to have an effect on the probability of an event of another type. We characterize optimum policies so as to minimize the cost per unit of time for an infinite time span, thus, $\lim_{t\to\infty}\frac{C(t)}{t}$ with $C(t)$ the cost incurred in $[0,t]$.

Under this policy, the unit is periodically checked and replaced, hence, the maintenance process yields a renewal process. In fact, it corresponds to the particular class of the renewal-reward processes with the term cost being used instead of reward. The properties of the renewal reward processes ensure that the previous cost function converges, with probability 1, to the next one

$$Q(T)=\frac{E[C(\tau)]}{E(\tau)}$$

A cycle, denoted τ, is the time span between two consecutive renewals of the system (see Ross (1996)) with $C(\tau)$ denoting its associated (random) cost. From now on $Q(T)$, depending on the time between consecutive inspections, T, will denote the objective function.

In section 2, the maintenance policy is described along with the conditions that guarantee the existence of an optimum policy for a series system (Theorem 1) and a parallel system (Theorem 2). In addition, we consider multivariate distributions which are natu-

ral extensions of the exponential random variable: the bivariate Gumbel distribution, as well as the Marshall-Olkin model. These distributions allow the dependence between components and, provided they hold, we obtain particular conditions for the existence of an optimum policy. Finally, section 3 contains some examples involving the previous distributions which illustrate the theoretical results.

2. THE MODEL

The system consists of n units which are periodically checked at times kT, $k = 1,2,\ldots$. In addition each unit, had it failed or not, is replaced by a new one. Times of inspections and replacements are assumed to be negligible.

The following notation will be handled:

- X_i, is unit i's random time to failure $(i = 1,2,\ldots,n)$
- μ_i, F_i, f_i, R_i and r_i are, respectively, the mean value, cumulative distribution function, density function, reliability function and failure rate of X_i, $(i = 1,2,\ldots,n)$
- $T^\star$: optimum replacement time

Let c_1 be the cost of the replacement when no failure has occurred and c_{2i}, $(i = 1,2,\ldots, n)$ the corresponding cost if the unit i has failed. In practice, it is usual to consider $c_1 < c_{2i}$. In addition, we will take into account the cost derived from the down-time which represents the losses incurred while the failure of the system remains undiscovered. It can be due to defective production, unavailability of the system when it is needed, etc.

Let also consider the ordered times to failure of the units, thus, $X_{(1)},X_{(2)},\ldots,X_{(n)}$. In addition, $\mu_{(i)}$, $F_{(i)}$, $f_{(i)}$, $R_{(i)}$, and $r_{(i)}$ are the mean value, distribution function, density function, reliability function and failure rate corresponding to $X_{(i)}$. Finally, the multivariate distribution and reliability functions of the system are, respectively

$$\begin{aligned} F(x_1,x_2,\ldots,x_n) &= P(X_1 \leq x_1, X_2 \leq x_2,\ldots,X_n \leq x_n) \\ R(x_1,x_2,\ldots,x_n) &= P(X_1 > x_1, X_2 > x_2,\ldots,X_n > x_n) \end{aligned}$$

In what follows, we will analyze the maintenance policy for series and parallel systems, and we will begin with the former.

CASE I. The system works if all the component work.

In this case, the time to failure of the system is given by

$$X_{(1)} = \min(X_1,X_2,\ldots,X_n)$$

being $R_{(1)}(t) = R(t,\ldots,t)$ its corresponding reliability function. Now, c_{di} is the cost derived from the down-time when component i fails.

The next result provides the expression of the cost function.

Proposition 1 *Consider a n-out-of-n system whose failures are only detected by inspection. Under the maintenance policy previously described, the following results hold*

$$(1) \qquad E[C(\tau)] = c_1 R_{(1)}(T) + \sum_{i=1}^{n} c_{2i} F_i(T) - \sum_{i=1}^{n} c_{di} \int_0^T R_i(u)du + \sum_{i=1}^{n} c_{di} T$$

$$(2) \qquad Q(T) = \sum_{i=1}^{n} c_{di} + \frac{c_1 R_{(1)}(T) + \sum_{i=1}^{n} c_{2i} F_i(T) - \sum_{i=1}^{n} c_{di} \int_0^T R_i(u)du}{T}$$

Proof

At times kT, $(k = 1,2,\ldots,)$ the system is renewed, hence, the mean length of a cycle is $E[\tau] = T$. The cost of a cycle under this maintenance procedure is

$$C(\tau) = c_1 I(X_1 > T, X_2 > T, \ldots, X_n > T) + \sum_{i=1}^{n} c_{2i} I(X_i \leq T) + \sum_{i=1}^{n} c_{di} [T - \min(X_i, T)]$$

where $I(.)$ represents the indicator function. The expectation of the expression above, leads to (1).

Both $E[\tau]$ and (1) provide $Q(T)$ given in (2). □

The following theorem gives a sufficient condition which guarantees the existence of an optimum policy.

Theorem 2 *Consider the same conditions given in Proposition 1, under the next inequality*

$$(3) \qquad \sum_{i=1}^{n} c_{2i} < \sum_{i=1}^{n} c_{di} \mu_i, \qquad i = 1,2,\ldots,n$$

there exists T^* in $(0,\infty)$ minimizing $Q(T)$ in (2). Moreover, T^* is one of the roots of the following equation

$$T\left[-c_1 f_{(1)}(T) + \sum_{i=1}^{n} c_{2i} f_i(T) - \sum_{i=1}^{n} c_{di} R_i(T)\right] -$$

$$- \left[c_1 R_{(1)}(T) + \sum_{i=1}^{n} c_{2i} F_i(T) - \sum_{i=1}^{n} c_{di} \int_0^T R_i(u)du\right] = 0$$

(4)

Proof

$Q(T)$ can be expressed as

$$Q(T) = \sum_{i=1}^{n} c_{di} + \frac{a(T)}{T}$$

where

$$a(T) = c_1 R_{(1)}(T) + \sum_{i=1}^{n} c_{2i} F_i(T) - \sum_{i=1}^{n} c_{di} \int_0^T R_i(u)du$$

$a(T)$ is a continuous function in $[0,\infty]$ verifying $a(0) = c_1 > 0$. From (3), it is derived that $a(\infty) = \sum_{i=1}^{n} c_{2i} - \sum_{i=1}^{n} c_{di}\mu_i < 0$. Note that the two limiting cases, $T = 0$ and $T = \infty$, correspond, respectively, to a continuous inspection and no inspection at all. Therefore, there exists T_0 in $(0,\infty)$ such that $a(T_0) < 0$, hence, $Q(T_0) < \sum_{i=1}^{n} c_{di} = Q(\infty)$. In addition, $Q(0) = \infty$, so, there exists T_1, $0 < T_1 < T_0$, satisfying $Q(T) \leq \inf_{x\in[0,T_1]} Q(x)$ for all $T > T_1$.

$Q(T)$ is a continuous function, therefore it has a minimum, T^*, in $[T_1,\infty]$. Moreover, next inequality

$$Q(T_1) \geq Q(\infty) > Q(T_0)$$

leads to $T_1 < T^* < \infty$. Therefore, T^* is also the minimum in $[0,\infty]$, and the result holds. Finally, by differentiation of $Q(T)$, (4) is obtained. □

The particular case of time to failure exponentially distributed is analyzed in the following result.

Proposition 3 *Consider the same conditions given in Proposition 1, and X_i, exponentially distributed with failure rate λ_i, $i = 1,2,\ldots,n$. If the next two conditions hold*

$$\lambda_i < \frac{c_{di}}{c_{2i}}, \quad i = 1,2,\ldots,n \tag{5}$$

$$r_{(1)}^2(T) > r_{(1)}'(T) \tag{6}$$

with $r_{(1)}'(T)$ being the derivative of $r_{(1)}(T)$. Then, T^* is the only root of (4).

Proof

Denote $A(T)$ the left-hand side in equation (4). In this case, the derivative of $A(T)$ is expressed as

$$\frac{dA(T)}{dT} = T\left[c_1\left(r_{(1)}^2(T) - r_{(1)}'(T)\right)R_{(1)}(T) - \sum_{i=i}^{n} \lambda_i^2 e^{-\lambda_i T}\left(c_{2i} - \frac{c_{di}}{\lambda_i}\right)\right]$$

The hypotheses (5) and (6) assure that the expression above is positive and, therefore, $A(T)$ has an only root. □

The multivariate extensions of exponential distribution arise easily to model the time to failure when the maintenance of multi-unit systems is studied. We consider the following bivariate extensions

1. The bivariate Gumbel distribution (Gumbel (1960))

The corresponding bivariate reliability function is given by

$$R(x,y) = e^{-x(\lambda_1+\lambda_3 y)-\lambda_2 y},\ 0 \le \lambda_3 \le \lambda_1\lambda_2,\ \lambda_1,\lambda_2 > 0,\ x,y \ge 0$$

This distribution has exponential marginal distributions, however, the failure rate defined as

$$r(x,y) = \frac{f(x,y)}{R(x,y)}$$

is not a constant, unlike the univariate exponential.

2. The Marshall-Olkin model (Marshall-Olkin (1966))

Marshall and Olkin derived the joint reliability distribution

$$R(x,y) = e^{-\lambda_1 x-\lambda_2 y-\lambda_3 \max(x,y)},\quad \lambda_1,\lambda_2 > 0,\ \lambda_3 \ge 0,\ x,y \ge 0$$

so as to describe a two-component system subject to shocks following a Poisson process. An important property of this distribution is that $X = \min(T_1,T_2)$, $Y = \min(T_2,T_3)$ where T_1, T_2 and T_3 are independent exponential random variables. It follows that $\min(X,Y)$ is also exponentially distributed. In addition, (X,Y) is the only distribution to have the so-called no-aging property

$$P(X \ge x+t, Y \ge y+t \mid X \ge t, Y \ge t) = P(X \ge x, Y \ge y)\quad x,y,t \ge 0$$

which means that the remaining lifetime, given that both components have survived up age t, does not depend on t.

In the former models, both X and Y are exponentially distributed. Hence, by direct application of Proposition 2, the following corollaries are derived

Corollary 4 *Consider a two-out-of-two system whose time to failure is distributed as the bivariate Gumbel model. Under the following conditions*

$$\lambda_i < \frac{c_{di}}{c_{2i}},\quad i = 1,2$$
$$(\lambda_1+\lambda_2)^2 > 2\lambda_3$$

T^* is the only root in equation (4).

Proof

The reliability function and the failure rate corresponding to $X_{(1)} = \min(X,Y)$ are, respectively

$$\begin{aligned} R_{(1)}(T) &= e^{-(\lambda_1+\lambda_2)T-\lambda_3 T^2} \\ r_{(1)}(T) &= (\lambda_1+\lambda_2)+2\lambda_3 T \end{aligned}$$

The hypotheses of Proposition 2 are satisfied and the result follows. □

Corollary 5 *Consider a two-out-of-two system whose time to failure is distributed as the Marshall-Olkin model. T^*, is the only root of equation (4), provided*

$$\lambda_i+\lambda_3 < \frac{c_{di}}{c_{2i}}, \quad i=1,2$$

Proof

X and Y are exponentially distributed with parameters $\lambda_1+\lambda_3$ and $\lambda_2+\lambda_3$, respectively. Besides, the reliability function of $X_{(1)} = \min(X,Y)$ and its failure rate are respectively given by

$$\begin{aligned} R_{(1)}(T) &= e^{-(\lambda_1+\lambda_2+\lambda_3)T} \\ r_{(1)}(T) &= \lambda_1+\lambda_2+\lambda_3 \end{aligned}$$

again, Proposition 2 leads to the result. □

Next, the particular case of independence between times to failure of the components is considered.

Proposition 6 *Let $X_1, X_2, \ldots, X_n$ be independent random variables, and such that*

$$\alpha_i \leq \frac{c_{di}}{c_{2i}}, \quad i=1,2,\ldots,n \tag{7}$$

with $\alpha_i = \sup_{0\leq t<\infty} r_i(t)$. Then, T^* is the only root of (4).

Proof

Denoting $A(T)$ the left-hand side of (4), its derivative can also be expressed

$$\begin{aligned} \frac{dA(T)}{dT} &= T\left[c_1 r_{(1)}^2(T)R_{(1)}(T)+\sum_{i=1}^{n} c_{2i}r_i'(T)R_i(T)\right]- \\ &- T\left[c_1 r_{(1)}'(T)R_{(1)}(T)+\sum_{i=1}^{n} r_i(T)f_i(T)\left(c_{2i}-\frac{c_{di}}{r_i(T)}\right)\right] \end{aligned}$$

The independence of $(X_1, X_2, \ldots, X_n)$ leads to

$$R_{(1)}(T) = \prod_{i=1}^{n} R_i(T), \quad r_{(1)}(T) = \sum_{i=1}^{n} r_i(T), \quad r'_{(1)}(T) = \sum_{i=1}^{n} r'_i(T)$$

hence

$$\begin{aligned} \frac{dA(T)}{dT} &= T\left[c_1\left(\sum_{i=0}^{n} r_i(T)\right)^2 \prod_{i=1}^{n} R_i(T) + \sum_{i=1}^{n} r'_i(T) R_i(T)\left(c_{2i} - c_1 \prod_{j\neq i} R_j(T)\right)\right] - \\ &- T\left[r_i(T) f_i(T)\left(c_{2i} - \frac{c_{di}}{r_i(T)}\right)\right] \end{aligned}$$

Usually, it is assumed $c_1 < c_{2i}$, $i = 1, 2, \ldots, n$. This fact along with the hypotheses given in (7), prove that the derivative of $A(T)$ is positive and the result holds. □

CASE II. The system works if at least one component works.

The time to failure in a parallel system is given by the longer lifetime, thus

$$X_{(n)} = \max(X_1, X_2, \ldots, X_n)$$

with $\mu_{(n)} = E\left[X_{(n)}\right]$. Its reliability function is

$$R_{(n)}(x) = 1 - P(X_1 \leq x, X_2 \leq x, \ldots, X_n \leq x) = 1 - F(x, x, \ldots, x)$$

Now, we consider a unique c_d representing the cost derived while the failure of the whole components is not detected. First, the cost function is obtained.

Proposition 7 *Consider a 1-out-of-n system whose failures are only detected by inspection. Under the maintenance policy previously described, the following results hold*

$$E[C(\tau)] = c_1 R_{(1)}(T) + \sum_{i=1}^{n} c_{2i} F_i(T) + c_d T - c_d \int_0^T R_{(n)}(u)\,du \tag{8}$$

$$Q(T) = c_d + \frac{c_1 R_{(1)}(T) + \sum_{i=1}^{n} c_{2i} F_i(T) - c_d \int_0^T R_{(n)}(u)\,du}{T} \tag{9}$$

Proof

In Proposition 1, it was proved that $E[\tau] = T$. The cost of a cycle is

$$\begin{aligned} C(\tau) &= c_1 I(X_1 > T, X_2 > T, \ldots, X_n > T) + \sum_{i=1}^{n} c_{2i} I(X_i \leq T) + \\ &+ c_d\left[T - \min\left(X_{(n)}, T\right)\right] \end{aligned}$$

The expected value of the expression above is given by (8), therefore, (9) holds. □

The following result provides a sufficient condition for the existence of an optimum policy.

Theorem 8 *Consider the same conditions of Proposition 4, if*

$$\mu_{(n)} > \frac{\sum_{i=1}^{n} c_{2i}}{c_d} \tag{10}$$

then, there exists T^*, $0 < T^* < \infty$ *minimizing* $Q(T)$ *in (9). Moreover,* T^* *is one of the roots of the following equation*

$$T\left[-c_1 f_{(1)}(T) + \sum_{i=1}^{n} c_{2i} f_i(T) - c_d R_{(n)}(T)\right] - \left[c_1 R_{(1)}(T) + \sum_{i=1}^{n} c_{2i} F_i(T) - c_d \int_0^T R_{(n)}(u)du\right] = 0 \tag{11}$$

Proof

$Q(T)$ can be also expressed as

$$Q(T) = c_d + \frac{a(T)}{T}$$

where

$$a(T) = c_1 R_{(1)}(T) + \sum_{i=1}^{n} c_{2i} F_i(T) - c_d \int_0^T R_{(n)}(u)du$$

Clearly, $a(0) = c_1 > 0$ and, from hypothesis (10), it is obtained $a(\infty) = \sum_{i=1}^{n} c_{2i} - c_d \mu_{(n)} < 0$. Hence, there exists T_0, $0 < T_0 < \infty$ such that $Q(T_0) < c_d = Q(\infty)$. In addition $Q(0) = \infty$, and the same strategy used in Theorem 1 shows the existence of T^*, finite, minimizing $Q(T)$.

On the other hand, the roots of the derivative of $Q(T)$ should verify (11). □

Remark 1 *Conditions (3) and (10) mean that an optimum policy exists whenever the mean cost incurred for the down-time is greater than the cost derived from the preventive maintenance.* □

3. EXAMPLES

We present two examples that aim at illustrating the results presented in this work. The first of them refers to a series system following a bivariate Gumbel distribution and the second deals with a parallel system whose time to failure is distributed as the Marshall-Olkin model.

The periodic tests which are carried out so as to detect disorders in the blood pressure and the cholesterol level, correspond to the case of a series system. No sooner does one of the causes appear than the health disorder arises. In general if the cholesterol level is high so does the blood pressure, that is to say, both turn out to be related and may indicate the occurrence of a more serious illness. The disorders in the blood pressure and the cholesterol level are unrevealed failures as, apart from the corresponding tests, there is no other way to detect them.

Consider now an engineering system with two spare units being available. Both can be modeled by a parallel system and should be tested periodically so as to detect its failures which may occur even while they are not in use. The failures may be dependent if the units are stored under the same conditions, hence a failed unit makes more likely the failure of the second spare component.

Table 1

$E[X_{(1)}]$	λ_1	λ_2	λ_3	T_1^*	$Q(T_1^*)$	T_2^*	$Q(T_2^*)$
2849.98	10^{-4}	2×10^{-4}	10^{-8}	12.92	1.56	15.82	1.28
2263.39	10^{-4}	3×10^{-4}	10^{-8}	11.19	1.80	14.16	1.46
1868	10^{-4}	4×10^{-4}	10^{-8}	10.01	2.02	12.93	1.56
1586.3	10^{-4}	5×10^{-4}	10^{-8}	9.14	2.21	11.97	1.57
1376	10^{-4}	6×10^{-4}	10^{-8}	8.46	2.39	11.20	1.81
1332.84	10^{-4}	6×10^{-4}	2×10^{-8}	8.46	2.39	11.20	1.81
1295.53	10^{-4}	6×10^{-4}	3×10^{-8}	8.46	2.39	11.20	1.81
787.44	5×10^{-4}	5×10^{-4}	2×10^{-7}	7.08	2.87	8.18	2.49
285	10^{-3}	2×10^{-3}	10^{-6}	4.09	5.01	5.01	4.11
158.63	10^{-3}	5×10^{-3}	10^{-6}	2.90	7.10	3.80	5.47
152.30	10^{-3}	5×10^{-3}	2×10^{-6}	2.90	7.10	3.80	5.47
147.02	10^{-3}	5×10^{-3}	3×10^{-6}	2.90	7.10	3.80	5.47
140.71	1.1×10^{-3}	5×10^{-3}	4×10^{-6}	2.88	7.17	3.75	5.55
47.81	0.01	0.01	10^{-5}	1.59	13.51	1.84	11.81
46.04	0.01	0.01	2×10^{-5}	1.59	13.51	1.84	11.81
42.14	0.01	0.01	5×10^{-5}	1.59	13.51	1.84	11.81
28.50	0.01	0.02	10^{-4}	1.30	16.58	1.60	13.74
22.63	0.02	0.02	10^{-4}	1.13	19.60	1.30	17.21

1. Series system with Gumbel distribution

Suppose that $c_1 = 10$, $c_{21} = 75$, $c_{22} = 35$, and $c_{d1} = 400$. We assume two different costs for the down-time due to failure in the second component: $c_{d2} = 400$ or $c_{d2} = 200$, being $T_1^\star$ and $T_2^\star$ their corresponding optimum inspection times.

Table 1 contains the optimum inspection times as well as the corresponding optimum cost for different mean times to failure of the system. The smaller $E[X_{(1)}]$ is, the smaller $T^\star$; on the contrary, $Q(T^\star)$ increases. In addition $T_1^\star < T_2^\star$ and $Q(T_1^\star) > Q(T_2^\star)$, which means that the higher the down-time cost is, the more frequent the inspection and, therefore, the more expensive the total cost.

2. Parallel System with the Marshall-Olkin model

Costs: $c_1 = 10$, $c_{21} = 75$, $c_{22} = 35$ and $c_d = 400$ ($T_3^\star$) or 200 ($T_4^\star$).

Table 2 shows that the optimum inspection time, T^*, is non-monotonic when the mean time to failure decreases. Moreover, $T_3^\star < T_4^\star$ and $Q(T_3^\star) > Q(T_4^\star)$; therefore, the conclusions derived are similar to those in the previous example.

Table 2

$E[X_{(n)}]$	λ_1	λ_2	λ_3	T_3^*	$Q(T_3^*)$	T_4^*	$Q(T_4^*)$
12150	10^{-4}	10^{-5}	7446×10^{-8}	25.92	0.79	36.67	0.56
11070	10^{-4}	10^{-4}	2967×10^{-8}	40.71	0.50	57.37	0.36
10632	10^{-4}	10^{-4}	3422×10^{-8}	37.97	0.54	53.55	0.38
10414	10^{-4}	2×10^{-4}	1229×10^{-8}	60.05	0.34	83.16	0.24
10371	10^{-4}	2×10^{-4}	1276×10^{-8}	59.11	0.34	81.93	0.25
10334	10^{-4}	2×10^{-4}	1317×10^{-8}	58.32	0.35	80.90	0.25
10290	10^{-4}	2×10^{-4}	1366×10^{-8}	57.42	0.35	79.71	0.26
3213	5×10^{-4}	2×10^{-4}	1507×10^{-7}	18.09	1.15	25.52	0.83
1215	10^{-3}	10^{-3}	1975×10^{-7}	15.19	1.40	21.14	1.03
1058	10^{-3}	2×10^{-3}	1055×10^{-7}	18.15	1.14	24.47	0.87
1053	10^{-3}	2×10^{-3}	1103×10^{-7}	17.90	1.16	24.18	0.88
1048	10^{-3}	2×10^{-3}	1160×10^{-7}	17.62	1.18	23.84	0.89
1041	10^{-3}	2×10^{-3}	123×10^{-6}	17.30	1.21	23.45	0.91
968	1.1×10^{-3}	2×10^{-3}	1358×10^{-7}	16.56	1.27	22.48	0.96
158	0.01	5×10^{-3}	2421×10^{-6}	4.34	5.51	6.06	4.21
154	0.01	5×10^{-3}	2628×10^{-6}	4.19	5.71	5.86	4.35
152	0.01	6×10^{-3}	2030×10^{-6}	4.62	5.18	6.40	3.98
122	0.01	0.01	1976×10^{-6}	4.47	5.33	6.13	4.14
77	0.02	0.01	5142×10^{-6}	2.96	8.63	4.13	6.72

ACKNOWLEDGEMENT

We would like to thank an associate editor of *Qüestiió* and the two anonymous referees, for helpful comments that led to improvement of the paper. We also thank César Berrade for his helpful assistance.

REFERENCES

Berrade, M. D. (1999). «Técnicas Estadísticas en Fiabilidad: Propiedades de Edad bajo Mixturas y Mantenimiento Óptimo de Sistemas». *PhD-Thesis*, Universidad de Zaragoza.

Ebrahimi, N. (1997). «Multivariate Age Replacement». *Journal of Applied Probability*, 34, 1032-1040.

Gumbel, E. J. (1960)«Bivariate Exponential Distribution». *Journal of American Statistical Association*, 55, 698-707.

Lewis, E. E. (1987). *Introduction to Reliability Engineering*, Wiley, New York.

Marshall, A. W. and Olkin, I. (1966). «A Multivariate Exponential Distribution». *Journal of American Statistical Association*, 62, 30-44.

Nakagawa, T. and Yasui, K. (1991). «Periodic-Replacement Models with Threshold Levels». *IEEE Transactions on Reliability*, 40, 395-397.

Ross, S. (1996). *Stochastic Processes*, 2nd edition, John Wiley & Sons, New York.

Vaurio, J. K. (1995). «Optimization of Test and Maintenance Intervals Based on Risk and Cost». *Reliability Engineering and System Safety*, 49, 23-26.

Vaurio, J. K. (1997). «On Time-Dependent Availability and Maintenance Optimization of Standby Units under Various Maintenance Policies». *Reliability Engineering and System Safety*, 56, 79-89.

Vaurio, J. K. (1999). «Availability and Cost Functions for Periodically Inspected Preventively Maintained Units». *Reliability Engineering and System Safety*, 63, 133-140.

QÜESTIIÓ, vol. 26, 1-2, p. 259-271, 2002

INVERSE SAMPLING AND TRIANGULAR SEQUENTIAL DESIGNS TO COMPARE A SMALL PROPORTION WITH A REFERENCE VALUE*

VÍCTOR MORENO[1,2]
ISAAC MARTÍN[3]
FERRAN TORRES[2]
MANUEL HORAS[2]
JOSÉ RIOS[2]
JUAN R. GONZÁLEZ[1]

Inverse sampling and formal sequential designs may prove useful in reducing the sample size in studies where a small population proportion p is compared with a hypothesized reference proportion p_0. These methods are applied to the design of a cytogenetic study about chromosomal abnormalities in men with a daughter affected by Turner's syndrome. First it is shown how the calculated sample size for a classical design depends on the parameterization used. Later this sample size is compared with the required sample size in an inverse sampling design and a triangular sequential design using four different parameterizations (absolute differences, log-odds ratio, angular transform and Sprott's transform). The expected savings in sample size, when the alternative hypothesis is true, are 20% of the fixed sample size for the inverse sampling design and 40% for the triangular sequential design.

Keywords: Sample size, inverse sampling, sequential methods, triangular sequential test

AMS Classification (MSC 2000): 62L05 Statistics, Sequential methods

*Please address all correspondence to: Victor Moreno. Servicio de Epidemiología y Registro del Cáncer. Instituto Catalán de Oncología. Gran Vía km 2.7.Hospitalet, 08907 Barcelona. Spain.
E-mail: V.Moreno@ico.scs.es. Tel: +34-93 260 78 12. Fax: +34-93 260 77 87.

[1] Instituto Catalán de Oncología. Gran Vía km 2.7, Hospitalet, 08907 Barcelona, Spain.

[2] Laboratorio de Bioestadística y Epidemiología. Facultad de Medicina, Universidad Autónoma de Barcelona. Barcelona, Spain.

[3] Departamento de Estadística y Econometría. Universidad Carlos III, Madrid, Spain.

–Received February 2000.
–Accepted June 2001.

1. INTRODUCTION

The sample size needed in many biological experiments is so large when the characteristic of interest is a rare event that it is appealing to explore different sampling schemes oriented to reduce the number of observations needed. In this paper it is shown how inverse sampling and formal sequential designs may prove useful in reducing the sample size in some specific situations.

The methods described were motivated by the design of a cytogenetic study where the aim was to compare the proportion of chromosomal abnormalities observed in an affected individual with a known reference value for healthy individuals. Similar situations occur in preclinical studies of toxicity or while monitoring rare adverse events in clinical trials, especially in phase IV studies.

The hypothesis of interest was that the proportion of spermatozoa carrying chromosomal abnormalities in men with a daughter affected by Turner's syndrome was higher than that observed in control men (without daughters affected by Turner's syndrome). Turner's syndrome appears when one sexual chromosome is lost and the karyotype results in 22 pairs of somatic chromosomes but only one sexual chromosome X. The affected individual is a female with some specific phenotypic characteristics. The chromosomal studies to determine abnormalities involve complex experiments in which hamster oocytes are fused with human spermatozoa.

The proportion of spermatozoa with the sexual chromosome missing in normal men has been accurately estimated in several studies to be around 0.003. In the present study, due to the technical difficulties in determining the abnormalities, it was thought sensible that the results from a sample of cases (men with a daughter affected by Turner's syndrome) could be compared to the already known proportion in control men.

The aim was to determine whether a population proportion p differs from a hypothesized reference proportion p_0. In our study, doubling the reference proportion was considered an increase interesting to detect. Thus, the interest was to reject the null hypothesis that $p = p_0 = 0.003$ when $p \geqslant 0.006$ with power 0.80. Only the one sided alternative that the proportion was higher in men with an affected daughter than controls was considered and a conservative significance level of 0.025 was adopted.

In the classical fixed sample size design, a sample size should be calculated to satisfy the power requirements. Data should be collected but not examined and a decision taken until the complete sample size had been achieved. In our situation, because the proportion of abnormal spermatozoa was very small, the sample size required was large and further, as will be shown later, the parameterization used to compare the proportions affected the required sample size in a significant amount.

Table 1. Formulas for Z and V and sample size according to the parameterization of θ, the difference between p and p_0.

θ	Z	V	n
Log-odds ratio $\log(p(1-p_0)/(p_0(1-p)))$	$r-np_0$	$np_0(1-p_0)$	$\dfrac{(z_\alpha+z_\beta)^2}{\log^2\left(\dfrac{p(1-p_0)}{p_0(1-p)}\right)p_0(1-p_0)}$
Probability difference $p-p_0$	$\dfrac{r-np_0}{p_0(1-p_0)}$	$\dfrac{n}{p_0(1-p_0)}$	$\dfrac{(z_\alpha+z_\beta)^2\,p_0(1-p_0)}{(p-p_0)^2}$
Angular transformation $\arcsin\sqrt{p}-\arcsin\sqrt{p_0}$	$2\dfrac{r-np_0}{\sqrt{p_0(1-p_0)}}$	$4n$	$\dfrac{(z_\alpha+z_\beta)^2}{4\left(\arcsin\sqrt{p}-\arcsin\sqrt{p_0}\right)^2}$

Alternatives to the fixed sample size design may prove interesting in reducing the sample size needed while preserving the statistical errors at the predefined levels. Two such alternatives will be revised in this paper. First, inverse sampling, that consists on sampling until the first r occurrences of an event are seen in a sample. Secondly, a formal sequential analysis based in successive examinations of accumulating data with a prespecified stopping rule.

2. SAMPLING PROCEDURES

2.1. Fixed sample size design

Several formulas may be used to calculate the sample size needed to compare a proportion with a reference value. The notation used by Whitehead (1992), that allows deriving a variety of these formulas from a unified theory, will be followed. This notation will also be used to design and analyze formal sequential tests.

A sequence $x_1, x_2, \ldots, x_n$ of independent binary observations with event probability p is observed and each one is coded 1 if the event appears or 0 otherwise. The null hypothesis to test is $H_0: p = p_0$. This is equivalent to testing whether a parameter $\theta = g(p, p_0)$ is equal to zero, where g is a function that parameterizes the difference between p and p_0 in an appropriate measurement scale.

As a Bernoulli experiment, the likelihood of p based on n observations is

$$L(p) = p^r (1-p)^{n-r}$$

where $r = x_1 + x_2 + \cdots + x_n$ is the sum of responses. The log-likelihood is

$$l(p) = r \log(p/(1-p)) + n \log(1-p).$$

Following Whitehead's notation (Whitehead, 1992), in this log-likelihood p can be substituted by θ from $g(p, p_0)$. The log-likelihood can be approximated by a Taylor's series expansion of second order and two statistics derived, namely Z and V.

$$l(\theta) = \text{const} + \theta Z - \tfrac{1}{2}\theta^2 V + O(\theta^3)$$

From the series expansion we obtain that $Z = l_\theta(0)$ and $V = -l_{\theta\theta}(0)$, where $l_\theta(0)$ and $l_{\theta\theta}(0)$ denote respectively the first and second derivatives of $l(\theta)$ with respect to θ, evaluated at $\theta = 0$. Z is the efficient score for θ, a cumulative measure of the difference between p and p_0, and V is Fisher's information about θ contained in Z. The actual formula for Z and V are different depending on the choice of the function g. Table 1 shows three possible forms, the first is based on log-odds ratio scale, which measures relative differences, the second is based on absolute differences, and the third uses the angular transformation, which stabilizes the variance of the proportion. Note that Z, the cumulative difference, is always a function of $r - np_0$, that is, the observed minus expected number of responses. Also V, the information, is always proportional to the sample size, n.

For large sample sizes and small values of θ, the distribution of Z is approximately normal with mean θV and variance V. This normal approximation can be used to calculate the required sample size for a fixed sample design. Z, as an efficient score, may be used as the test statistic with working significance level α, and power $1 - \beta$. If Z is greater than some value $k = k(\alpha, \beta)$, then the null hypothesis is rejected at the level of significance α and it is concluded that the proportion in experimental group p is superior to the hypothesized p_0. The requirements for the one sided test are

$$\begin{aligned} P(Z \geqslant k/\theta = 0) &= \alpha \\ P(Z \geqslant k/\theta = \theta_R) &= 1 - \beta \end{aligned}$$

where θ_R is the difference which, if present, should be detected. A fixed sample study will satisfy these requirements if the information V and k are given by

$$\begin{aligned} V &= \left\{ (z_\alpha + z_\beta) / \theta_R \right\}^2 \\ k &= (z_\alpha + z_\beta)\, z_\alpha / \theta_R \end{aligned}$$

where z_γ denotes the upper $100(1-\gamma)$ percentage point of the normal distribution. Formulas for V are used to translate a requirement value for V into a required total sample

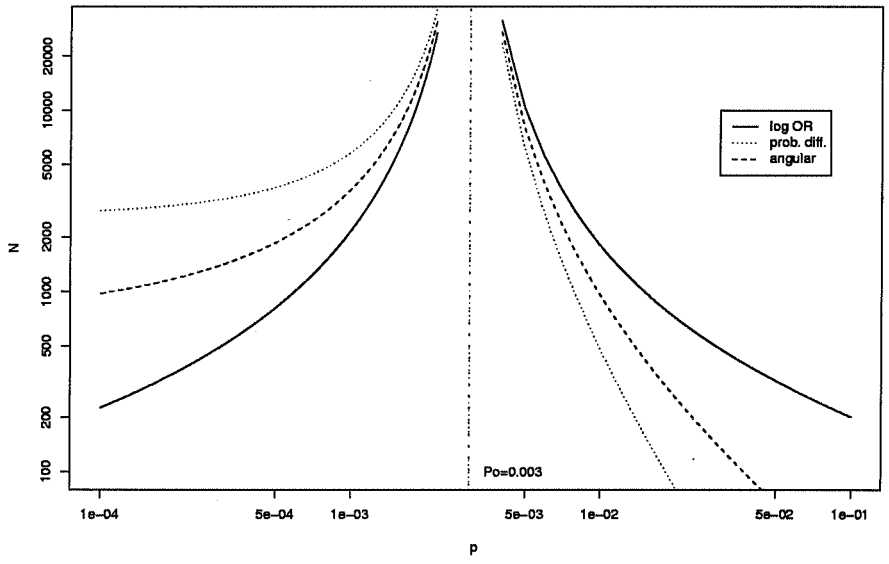


Figure 1. Sample size needed for various parameterizations.

size n. The formulas for Z, V and n are shown in table 1 for each parameterization studied.

The three parameterizations give different values for the sample size n, depending on the values of p and p_0. We can see in figure 1 that

$$\text{If } p > p_0 \qquad n_{\log OR} > n_{\text{angular}} > n_{\text{prob. diff.}}$$
$$\text{If } p < p_0 \qquad n_{\log OR} < n_{\text{angular}} < n_{\text{prob. diff.}}$$

For the final statistical analysis of the difference between p and p_0 the chosen test also affects the results. An exact test will be used since the proportions compared are small, though exact tests are more conservative due to the discreteness of the response variable. An equivalent form to the one sided exact test, will be the calculation of the exact lower limit of the $100(1-2\alpha)\%$ confidence interval for the estimated proportion p assuming a binomial distribution. This limit can be calculated easily using the relation between the F and the binomial distributions (Jowett, 1963):

$$p_{\text{lower}}(r,t,\alpha) = r/\{r+(t+1)f_\alpha(2(t+1),2r)\},$$

Table 2. Sample size and performance of fixed sample design with different parameterizations and tests for $\alpha = 0.025$ and power $= 0.80$.

	Sample	Exact test		χ^2test	
Method	Size	α	Power	α	Power
Log-odds ratio	5415	0.0159	0.8939	0.0280	0.9272
Probability difference	2608	0.0154	0.6010	0.0309	0.7003
Angular transformation	3795	0.0249	0.8157	0.0452	0.8704

where r is the number of cases with the characteristic, t is the number of cases without it and $f_\alpha(a,b)$ is the upper $100(\alpha)$ percentile of the F distribution with a and b degrees of freedom. If a bilateral test was used, the upper limit of the confidence interval could be calculated with the formula:

$$p_{\text{upper}}(r,t,\alpha) = r f_\alpha(2r,2t)/\{r f_\alpha(2r,2t)+t\}.$$

In our study, p_0 was 0.003 and the value of p we wished to detect was 0.006. We can see in table 2 the sample size calculated using the different definitions of θ with significance level $\alpha = 0.025$ and power 0.80. The parameterization of θ results in differences in the sample size needed, especially when the proportion is small. The statistical test used for the analysis is also important. In table 2 we can see the observed type I error rate and power after 50000 simulations for the fixed sample size design with a classical chi-square test without continuity correction, and for the exact test. As expected, for each parameterization of θ, the exact test gives more conservative results. The most accurate results, with respect to the predefined error rates, correspond to the angular transformation, which is in concordance with known results about comparison of two binomial proportions (Haseman, 1978).

2.2. Inverse sampling

The inverse sampling design is an old method (Haldane, 1945; Finney, 1949) to estimate a proportion p. The method consists on sampling until exactly r occurrences of an event appear in a study and counting the needed sample size n. In a classical fixed sample design sampling continues until the complete sample size n is attained and the number of occurrences r are counted. For both designs, the proportion p is estimated as r/n. The differences appear in the way the variance of this proportion is calculated. Methods to calculate the confidence interval for p when the inverse sampling design is used have been described by George and Elston (1993), for the special case when sampling continues until the first occurrence of an event of interest. They use the geometric

Table 3. Sample size needed with an inverse sampling design to detect $p = 0.006$ for given r and $\alpha = 0.025$.

r	T_{max}	Mean n ($p = 0.006$)	Observed α	Expected power	Observed power
10	1591	1426	0.0259	0.4921	0.4967
11	1822	1612	0.0271	0.5402	0.5416
12	2058	1803	0.0256	0.5859	0.5814
13	2297	1985	0.0257	0.6281	0.6263
14	2540	2167	0.0266	0.6677	0.6673
15	2787	2355	0.0228	0.7045	0.7137
16	3036	2533	0.0258	0.7379	0.7384
17	3288	2721	0.0244	0.7685	0.7671
18	3542	2899	0.0264	0.7960	0.8005
19	3798	3084	0.0278	0.8208	0.8142
20	4056	3249	0.0283	0.8431	0.8499
21	4316	3432	0.0282	0.8630	0.8658
22	4578	3595	0.0292	0.8807	0.8858
23	4841	3780	0.0237	0.8963	0.8979
24	5106	3954	0.0253	0.9102	0.9122
25	5372	4112	0.0269	0.9223	0.9265

T_{max}: number of cases without the event needed to observe before r events so that the lower 95% confidence interval around p doesn't include $p_0 = 0.003$.

distribution to calculate the confidence interval and demonstrate that the length is shorter than the one calculated by use of direct binomial sampling under certain situations. This is because of the fact that no occurrences in the first $t = n - 1$ trials is more informative than 1 occurrence in n trials. Nevertheless, the length of the confidence interval calculated on the basis of the first single case ($r = 1$) may be too wide for general utility. Lui (1995*a*; 1995*b*) describes the extension of this procedure to accommodate any finite number of cases ($r > 1$), and calculates the confidence interval using the exact method based on the relations between the negative binomial, the binomial and the F distributions as previously described.

For the comparison between p and p_0, an r large enough should be chosen so that, with probability $1 - \beta$, the lower bound of the lower $100(1 - 2\alpha)\%$ confidence limit around p will exceed the hypothesized proportion p_0. Classical inverse sampling should continue including subjects until r events appear. However, in the one sided design a

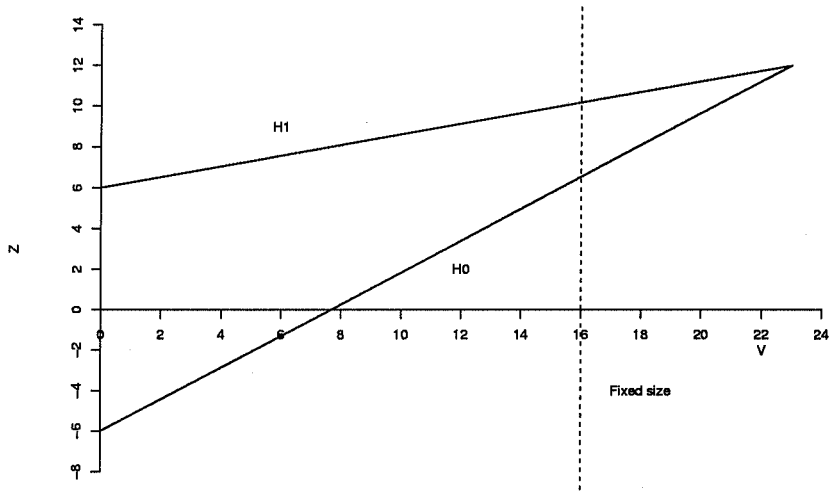


Figure 2. Continuation region for the triangular sequential test.

maximum value of t ($t_{\max}$) exists so that if the r events have not been observed before $t_{\max}$, the confidence interval will always include p_0. The design can be modified to stop sampling either when the r cases have been found or when $t_{\max}$ have been reached. We have checked, through simulations, that this truncation of the inverse sampling design did not alter the theoretic operating characteristic of test, and calculated the average sample size needed to finish the study, which was significantly reduced when the null hypothesis was not true.

Table 3 summarizes the results of simulations done to evaluate the power to detect a significant difference when $p = 0.006$ with the inverse sampling design using different values of r. $T_{\max}$, the maximum number of cases required for each value of r, can be calculated searching the quintile of the negative binomial distribution with p_0. The power to detect a given difference $p_R > p - p_0$ can be calculated from the negative binomial distribution function for those r and $t_{\max}$. In our case, with $r = 18$, the power to detect $p > p_0$ when $p = 0.006$ is 0.80 and the maximum number on cases without event needed to monitor, $t_{\max}$, is 3542. Note that, in this situation the average sample size is 2899, which corresponds to an 18.8% reduction of the maximum sample size for this number of cases and a 24% reduction relative to the fixed sample size design using the angular transform formula. In conclusion, if we were to use the inverse sampling method, we would continue sampling until either 18 cases had been found or until we reached 3542 subjects without the event.

2.3. Sequential methods

Formal sequential methods have proven useful in reducing the sample size needed to test hypotheses in some situations. For the current problem, among the different types of sequential methods available, the triangular test as defined by Whitehead (1992) has been chosen for comparison with inverse sampling. This sequential design is simple to implement and is attractive for practical situations.

In a sequential method the sample size needed is a random variable. The implementation of the method has two phases, the design and the analysis. In the design phase the sequential rule is defined given the following values: the difference of interest to be detected (θ_R), the test characteristics in terms of power ($1-\beta$) and significance level (α) and the shape of the boundaries chosen for the sequential rule. In the analysis phase, as data accumulates, repeated evaluations of the sequential rule are made. The values of Z, the cumulative difference between p and p_0, and V, the information that depends on n, are calculated at each inspection. These two values are plotted against each other, that is, Z against V, in a graph where the sequential rule has been drawn. For the triangular sequential test used here, the sequential rule consists on two straight line boundaries making the shape of a triangle as it is illustrated in Figure 2. The area inside these boundaries is called the continuation region. As data accumulates, the path from (Z,V) is drawn and sampling continues while the path is whithin the continuation region. Whenever this path crosses one of the boundaries, a decision is taken. If the upper boundary is crossed, the null hypothesis is rejected. Otherwise, if the lower boundary is crossed, the null hypothesis is accepted. The position of the boundaries are computed to assure a given power and significance level. Approximate p-values can be calculated depending on the position of the point where the boundaries were crossed in the last inspection. The computer program PEST (Brunier & Whitehead, 1993) performs all necessary calculations for the design and analysis of this sequential method.

As with the fixed sample design, the parameterization used for θ is important in our study because the proportions compared are small and then the normal approximation used is not as good as desired. We have compared the results of inverse sampling with all three parameterizations described in the fixed sample size paragraph and a new one, proposed by Sprott (1973) which was also explored by Whitehead (1981). The Sprott's method has the property that the expected value of the third derivative of the likelihood function is zero and gives a very good normal approximation even for extreme situations. For this parameterization,

$$\theta = \{\eta(p) - \eta(p_0)\}\{p_0(1-p_0)\}^{-1/3},$$

where

$$\eta(p) = \int_0^p \frac{dt}{\{t(1-t)\}^{2/3}}$$

Table 4. Theoretic and simulated sample sizes for the situation $p_0 = 0.003$, $p = 0.006$, $\alpha = 0.025$ and power = 0.80.

	Theoretic values			Simulated values		
Method	$E(n)$	Median (n)	Max (n)	α	Power	Mean (n)
Log-odds ratio	3608	3470	8365	0.0354	0.9127	2553
Probability difference	1738	1672	4029	0.0406	0.6758	1477
Angular transformation	2530	2433	5890	0.0376	0.8122	2018
Sprott's transformation	2147	1977	3795	0.0244	0.7905	2301
Inverse Sampling $r = 18$			3559	0.0264	0.8005	2899

All sample sizes calculated when $p = 0.006$, equivalent fixed sample size $n = 3795$.

This integral can be calculated numerically. The comparison of sequential methods and inverse sampling is shown in table 4. Although the sequential trials were designed with a type I error equal to 0.025 and a power of 0.80, simulations show that the triangular sequential test results in a type I error slightly greater than the specified for all the parameterizations studied, except for the Sprott's transformation. This parameterization adjusts very finely to the design characteristics; the angular transformation maintains the desired power of 0.8, while the log-odds ratio parameterization results in an excess power reaching 0.9 and the probability difference only 0.7. The average final sample size parallels the results of the attained power. The log-odds ratio parameterization needs more patients than the angular transformation and the lower number is seen for the probability difference, though with this parameterization the desired power is not attained. The Sprott's transformation needs a sample size intermediate between the log-odds ratio and the angular transformation. Note that, even for the log-odds ratio parameterization, the average sample size is smaller than the inverse sampling design. The Sprott's transformation, that gives best results in power and type I error, needs on average about 20% less sample size than the equivalent inverse sampling design and about 40% less sample size than the equivalent fixed sample design.

3. OTHER ALTERNATIVES

Alternative hypothesis different from $p = 0.006$ have been explored. We have chosen smaller values for the difference of interest ($p = 0.0035$, that corresponds to a 17% relative increase, $p = 0.004$ that corresponds to a 25% relative increase) and greater values ($p = 0.009$ that corresponds to a three-fold relative increase). Table 5 shows the summary results for the simulations using these alternatives. When the difference of interest is very small ($p = 0.0035$ or $p = 0.004$), the test characteristics are preserved better for

Table 5. Sample size and performance for triangular sequential tests with different parameterizations and inverse sampling for other alternative hypothesis p ($\alpha = 0.025$ and power $= 0.80$).

Method	p	α	Power	Average sample size (n)
Log-odds ratio	0.0035	0.0271	0.8364	67587
	0.004	0.0290	0.8545	18136
	0.009	0.0403	0.9524	853
Probability difference	0.0035	0.0280	0.7744	60274
	0.004	0.0300	0.7513	14553
	0.009	0.0600	0.6743	394
Angular transformation	0.0035	0.0266	0.8044	63984
	0.004	0.0307	0.8067	16287
	0.009	0.0451	0.8079	581
Sprott's transformation	0.0035	0.0239	0.7992	65992
	0.004	0.0250	0.7961	17316
	0.009	0.0219	0.7728	718
Inverse sampling				
r	t_{max}			
337	100339	0.0035	0.0217	0.7819	95619
99	26731	0.004	0.0223	0.7883	24432
8	1146	0.009	0.0242	0.8076	846
Fixed sample size for angular transform formula:				
$p = 0.0035$	101542			
$p = 0.004$	27232			
$p = 0.009$	1213			

all parameterizations. The triangular sequential test using the Sprott's parameterization needs, on average, a 35% smaller sample than the fixed size design. However, the sample sizes needed to detect these small differences are prohibitive for practical purposes. For alternatives of greater magnitude ($p = 0.009$), the sample size is reduced but with low performance of the tests characteristics.

Inverse sampling design keeps the test characteristics in all cases, but average sample size needed is greater than sequential tests.

4. DISCUSSION

We have explored alternative designs to reduce the sample size needed when the interest is to compare a small proportion with a reference value. Inverse sampling design, which stops sampling when a predefined number of events have been observed is an easy procedure to implement, and in our study could save 24% of the fixed sample size design in optimal situations, when the real proportion doubled the reference value of 0.003. The maximum sample size needed with this design is always inferior (6%) to the equivalent with fixed sample.

Formal sequential designs, based on continuous boundaries as the triangular sequential test, can reduce even more the sample size needed in optimal situations, up to 40%. However this method also has some limitations. As Whitehead (1992) stresses and we have checked for the design of our experiment, it is important to choose an adequate parameterization for the difference to be tested. The transformation proposed by Sprott performs well with respect to error rates, but other parameterizations explored have type I error rates greater than the specified and should be used with caution in situations similar to our case, where the reference proportion is small. Exact sequential methods have been developed for special designs, including this one (Stallard & Todd, 2000), but are not easy to implement.

The boundaries of these sequential methods impose the risk of needing more observations than the fixed sample case in some situations. For the triangular test, the maximum sample size in our experiment would be up to 30% more in the extreme case, which might occur if the real proportion is about half the difference between the reference proportion and the population proportion p_0. Though the situation where the true proportion is not as high as expected might not be so rare in practice, the median sample size of the triangular test is always smaller than the equivalent fixed design. For example, in our study the reference proportion p_R was 0.006 and the population proportion p_0 was 0.003. In the case that the real proportion was about 0.0045, the expected sample size would be 2147 and the 90^{th} percentile 3340, still 12% less than the 3795 needed with a fixed design as calculated by the angular transform formula.

The observed gain in sample size for the triangular test can be compared with the expected ones shown in table 4, derived from sequential theory (Whitehead, 1992). These theoretical sample sizes show some disagreement with the simulated values that parallel the observed and expected type I error rate. For the log-odds ratio, probability difference and angular transformation parameterizations, the observed mean sample size is smaller than the expected. These parameterizations show a lower type I error coverage than expected. For the Sprott's parameterization the observed mean sample size is slightly greater than the expected and type I error coverage is correct.

In conclusion, to compare a small proportion with a reference value, the inverse sampling design and formal sequential methods like the triangular test may prove useful to save sample size. The use of the triangular sequential method, that is based on approximations to the likelihood function should be cautious since the type I error rate is slightly increased and power varies unless the Sprott's parameterization is used. With these sequential methods the researcher must accept a small risk of needing to exceed the sample size that would be used with a classical design.

5. REFERENCES

Brunier, H. & Whitehead, J. (1993). *PEST Version 3, Operating Manual.* Reading: The University of Reading.

Finney, D. J. (1949). «On a method of estimating frequencies». *Biometrika*, 36, 233-234.

George, V. T. & Elston R. C. (1993). «Confidence limits based on first occurrence of an event». *Statist. Med.*, 12, 686-690.

Haldane, J. B. S. (1945). «On a method of estimating frequencies». *Biometrika*, 33, 222-225.

Haseman, J. K. (1978). «Exact sample sizes for use with Fisher-Irwin test for 2 ×2 tables». *Biometrics*, 34, 106-109.

Jowett, G. H. (1963). «The relationship between the binomial and *F* distributions». *The Statistician*, 13, 55-57.

Lui, K.-J. (1995). «Confidence limits for the population prevalence rate based on the negative binomial distribution». *Statist. Med.*, 14, 1471-1477.

Lui, K.-J. (1995). «Notes on conditional confidence limits under inverse sampling». *Statist. Med.*, 14, 2051-2056.

Sprott, D. A. (1973). «Normal likelihoods and their relation to large sample theory of estimation». *Biometrika*, 60, 457-465.

Stallard, N. & Todd, S. (2000). «Exact sequential test for single samples of discrete responses using spending functions». *Statist. Med.*, 19, 3051-64.

Whitehead, J. (1981). «Use of the sequential probability ratio test for monitoring the percentage germination of accessions in seed banks». *Biometrics*, 37, 129-136.

Whitehead, J. (1992). *The design and analysis of sequential clinical trials.* London: Ellis Horwood.

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.

FE D'ERRADES

En el número 3 del volum 25 (2001), a la solució del problema 89 (pàgina 585) hi ha dues errades:

On diu:	**Ha de dir:**
$0(n-1)$	$0(n^{-1})$
$2(p\Pi)\operatorname{vec}V^{-1}$	$2(p+1)\operatorname{vec}V^{-1}$

SOLUCIÓ ALS PROBLEMES PROPOSATS AL VOLUM 25 N. 3

PROBLEMA N. 90

1) The median M satisfies $F(M) = {}^1/_2$. Then $M = 1$ satisfies $F(1) = 1 - F(1)$, so $F(1) = {}^1/_2$.

2) The covariance between X and ${}^1/_X$ must be negative. Hence

$$\operatorname{cov}\left(X, \frac{1}{X}\right) = E\left(X \cdot \frac{1}{X}\right) - E(X) \cdot E\left(\frac{1}{X}\right)$$
$$= 1 - \mu^2 < 0$$

Thus $\mu > 1$.

3) The correlation coefficient between X and ${}^{-1}/_X$ is positive:

$$\rho\left(X, \frac{-1}{X}\right) = \frac{\mu^2 - 1}{\sigma^2} \leqslant 1.$$

Thus $\sigma^2 \geqslant \mu^2 - 1$.

4) Define the change $t = F(x)$. Then $F^{-1}(t) = x$ and, from $F(x) = 1 - F\left({}^1/_x\right)$, $F^{-1}(1-t) = {}^1/_x$. Thus

$$\int_0^1 \frac{F^{-1}(t)}{F^{-1}(1-t)}\, dt = \int_R \frac{x}{{}^1/_x}\, dF(x)$$
$$= \int_R x^2\, dF(x) = \sigma^2 + \mu^2.$$

C. M. Cuadras
Universitat de Barcelona

PROBLEMA N. 91

Let us find the covariance between $G(X)$ and $f'(X)/f(X)$ where

$$f(x) = F'(x) = \frac{e^{-x}}{(1+e^{-x})^2}$$

Note that $f(x) = F(x)\,(1-F(x))$. We have

$$E\,[f'(X)/f(X)] = \int_R \frac{f'(x)}{f(x)} f(x)\,dx =$$

$$= \int_R f'(x)\,dx = f(+\infty) - f(-\infty) = 0.$$

Thus

$$\operatorname{cov}(G(X), f'(X)/f(X)) = \int_R G(x)\,f'(x)\,dx =$$

$$= \Big[G(x)\,f(x)\Big]_{-\infty}^{+\infty} - \int_R G'(x)\,f(x)\,dx =$$

$$= -E\,G'(X).$$

Using the notation $f = F(1-F)$, $f' = F(1-F)(1-2F)$, the variance of $f'(X)/f(X)$ is

$$\int_R \left(\frac{f'(x)}{f(x)}\right)^2 f(x)\,dx = \int_R (1-2F)^2\,dF =$$

$$= \int_0^1 (1-2t)^2\,dt = \frac{1}{3}.$$

Hence

$$\big(EG'(X)\big)^2 \leqslant \operatorname{var}(G(X))\,\frac{1}{3}.$$

If $F = G$ then $G' = f$, $\operatorname{var}(F(X)) = {}^1/_{12}$ and

$$E\,f(X) = \int_R F(1-F)dF = \frac{1}{6}.$$

Thus we obtain an equality.

C. M. Cuadras
Universitat de Barcelona

PROBLEMA PROPOSAT

PROBLEMA N. 92

Let $\mathbf{H} = \mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}'$ the $n \times n$ centring matrix, $\mathbf{D} = (d_{ij})$ a $n \times n$ Euclidean distance matrix, and $\mathbf{A} = -{}^1/_2\left(d_{ij}^2\right)$. Define $\mathbf{B} = \mathbf{HAH}$. Then $\mathbf{B}$ is the centred inner product matrix for $\mathbf{D}$.

1) Find the eigenvalues of $\mathbf{H}$.
2) Suppose $\mathbf{B} = \mathbf{H}$. Find the distance matrix $\mathbf{D}$.
3) If $\mathbf{D}$ contains the distances between n objects, represent these objects by means of a dendrogram.

C. M. Cuadras
Universitat de Barcelona

QÜESTIIÓ, vol. 26, 1-2, p. 283-287, 2002

GEOMETRICAL UNDERSTANDING OF THE CAUCHY DISTRIBUTION

C. M. CUADRAS
Universitat de Barcelona

Advanced calculus is necessary to prove rigorously the main properties of the Cauchy distribution. It is well known that the Cauchy distribution can be generated by a tangent transformation of the uniform distribution. By interpreting this transformation on a circle, it is possible to present elementary and intuitive proofs of some important and useful properties of the distribution.

Keywords: Cauchy distribution, distribution on a circle, central limit theorem

AMS Classification (MSC 2000): 60-01; 60E05

– Received December 2001.
– Accepted April 2002.

The Cauchy distribution is a good example of a continuous stable distribution for which mean, variance and higher order moments do not exist. Despite the opinion that this distribution is a source of counterexamples, having little connection with statistical practice, it provides a useful illustration of a distribution for which the law of large numbers and the central limit theorem do not hold. Jolliffe (1995) illustrated this theorem with Poisson and binomial distributions, and he pointed out the difficulties in handling the Cauchy (see also Lienhard, 1996). In fact, one would need the help of the characteristic function or to compute suitable double integrals. It is shown in this article that these properties can be proved in an informal and intuitive way, which may be useful in an intermediate course.

If U is a uniform random variable on the interval $I = (-\pi/2, \pi/2)$, it is well known that $Y = \tan(U)$ follows the standard Cauchy distribution, i.e., with probability density

$$f(y) = \frac{1}{\pi}\frac{1}{1+y^2} \qquad -\infty < y < \infty.$$

Also $Z = \tan(nU)$, where n is any positive integer, has this distribution.

The tangent transformation can be described using the geometric analogy of a rotating diameter of a circle (Figure 1). Suppose, using usual rectangular co-ordinates, that the circle has center O and radius 1, and suppose a diameter POP' of the circle is a needle which rotates uniformly round the circle. Suppose P is the endpoint with positive value for x, and let OP make angle U with the x-axis. Then U lies in the interval $I = (-\pi/2, \pi/2)$. Let $Y = \tan(U)$; Y has the Cauchy distribution. Note also that $Y = \tan(U + \pi)$, so we only need to look at the half-right part of the final position of the needle. In the following, we will take all angles in the interval I modulo π and use the property that if V is any arbitrary random value, then $U + V \pmod{\pi}$ is also uniform on the interval I and is independent of U.

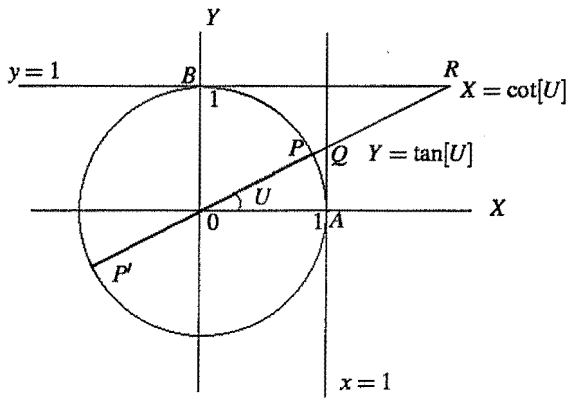


Figure 1.

In Figure 1, Y is the vertical coordinate of the point Q, where Q is the intersection of the line through the half-needle OP and the line $x = 1$. An interesting physical illustration is obtained by taking the origin as a radioactive source of α-particles impacting on a fixed line (Rao, 1973, p. 169).

The use of the circle and the rotating needle can give intuitive demonstrations of several features of the Cauchy distribution. Some examples are:

1) *Y has heavy tails, i.e., Y takes extreme values with high probability.*

   This feature empirically distinguishes Y from the standard normal and many other distributions. It is easily seen by observing that an angle U close to $\pi/2$ (or $-\pi/2$), which will give a large tangent value Y, has the same probability density ($= 1/\pi$) as an angle close to 0.

2) *$X = Y^{-1}$ is also distributed as a standard Cauchy distribution.*

   By symmetry, using the rotating needle analogy, the Cauchy variable could as well be generated by using the angle between the needle and the y-axis and projecting on the line $y = 1$ (giving the value BR in Figure 1). But this is the same as taking $X = \cot(U) = 1/Y$.

3) *The mean does not exist.*

   Actually, it is easy to give an analytic proof of this result, as taking the expectation of Y gives an indeterminate integral. Using the analogy of the circle, we should consider the position of the needle after several successive rotations. It is clear that the mean position might be anywhere around the circle, and a formal «mean» is not clearly defined.

4) *The distribution of the mean of any number of independent observations has the same distribution as Cauchy. Consequently, the law of the large numbers describing the convergent behavior of the mean does not hold for this distribution.*

   The analogy in 3) can be used again. Rotating the needle several times, we obtain axes uniformly distributed around the circle, and the intuitive average axis is also uniformly distributed. Taking U as its angle with the x-axis and constructing $Y = \tan(U)$ gives the standard Cauchy distribution. A rigorous definition of the mean direction and a proof of its uniform distribution needs, of course, more advanced arguments.

   The above analogy is clearly not a proof of 4), because the tangent of the angle of the mean axis is not the mean of the tangents. Instead, the Cauchy distribution for the mean of the Cauchy sample can be illustrated using the following simulation. Choose an integer m, generate independent uniform angles $u_1, \ldots, u_m$ and compute

$$\bar{u} = \operatorname{atan}\left(\frac{1}{m}\sum_{i=1}^{m}\tan(u_i)\right)$$

Repeat this operation n times and plot a histogram of the obtained sample $\bar{u}_1, \ldots, \bar{u}_n$. The uniform distribution of $\bar{u}$ will be quite apparent (Figure 2), showing that $\tan(\bar{u})$ is Cauchy. Note that to recognize the Cauchy distribution by a histogram of a Cauchy sample is less evident, due to the distortion produced by the very large values.

This geometrical approach, and the associated trigonometry, can give the proof of other properties, for example:

5) $Z = (Y^{-1} - Y)/2$ *has the standard Cauchy distribution.*

This is a consequence of the trigonometric identity $[\cot(u) - \tan(u)]/2 = \cot(2u)$.

6) *If W is any random variable then $T = (Y + W)/(1 - YW)$ has the standard Cauchy distribution.*

Using the formula $\tan(u+v) = [\tan(u) + \tan(v)]/[1 - \tan(u)\tan(v)]$, write $W = \tan(V)$, where V is any angle in I. Then $U + V \pmod{\pi}$ is uniformly distributed and $T = \tan(U + V)$.

7) *If U_1, U_2 are independently uniform on I, and if $Y_1 = \tan(U_1)$, $Y_2 = \tan(U_2)$, and $Y_3 = -\tan(U_1 + U_2)$, then Y_1, Y_2, Y_3 are pairwise independent with standard Cauchy distributions but are jointly dependent.*

From the trigonometric relation in 6) above, we have $Y_3 = -(Y_1 + Y_2)/(1 - Y_1 Y_2)$. Thus the relation $Y_1 Y_2 Y_3 = Y_1 + Y_3$ exists and the Y-values are not independent.

Further formulas were proved in a similar way by Jones (1999) for the normal and Cauchy distribution. See also Cuadras (2000). For example:

8) *Using that* $\tan(U)$ *is distributed as* $\tan(nU)$ for $n = 2,3,4$, *then if Z is standard Cauchy so is*

$$2Z/(1-Z^2), \quad Z/(3-Z^2)/(1-3Z^2) \quad \text{and} \quad 4Z/(1-Z^2)/(1-6Z^2+Z^4).$$

9) *Using* $\tan(2U + c)$ *or equivalently combining* $2Z/(1 - Z^2)$ *and* $(Z + B)/(1 - BZ)$, *where B is independent of Z (see above), then*

$$\frac{2Z + B(1 - Z^2)}{1 - B^2 - 2BZ}$$

is also standard Cauchy.

Finally, while it is relatively easy to prove, using only geometry, that the sum of independent N(0,1) is also normal (see Mantel, 1972), it is an open question to give a geometric but conscientious proof, elementary enough for teaching purposes, (i.e., without using double integrals or the characteristic function), that the mean of the tangents of uniform angles is the tangent of an angle also uniformly distributed in I, i.e, following the Cauchy distribution. See Cohen (2000).

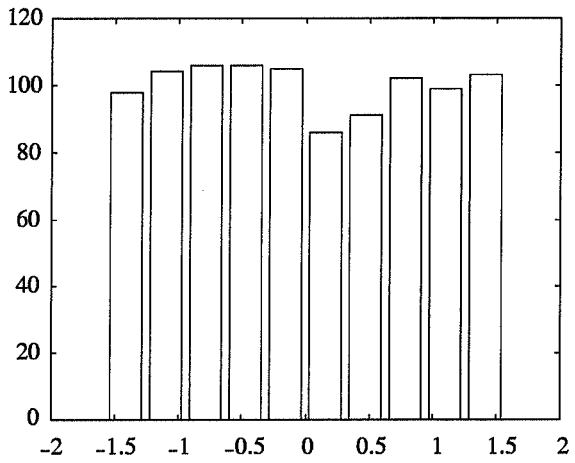


Figure 2. Histogram of a sample of $\bar{u}$ revealing a uniform distribution, indicating that $\tan(\bar{u})$ has the standard Cauchy distribution.

APPENDIX

MATLAB code to generate samples giving the histogram of Figure 2.

$m = 100$; $n = 1000$; rand('seed',2002);
for $i = 1 : n$, $u = \text{rand}(m, 1) * pi - pi/2$; $c = \tan(u)$; $mc(i) = \text{atan}(\text{mean}(c))$; end
hist(mc)

REFERENCES

Cohen, M. P. (2000). «Cuadras, C. M. (2000) Letter to the Editor, *The American Statistician*, 54, 87», *The American Statistician*, 54, 326.

Cuadras, C. M. (2000). «Distributional relationships arising from simple trigonometric formulas» (Revisited), *The American Statistician*, 54, 87.

Jolliffe, I. T. (1995). «Sample sizes and the central limit theorem: The Poisson distribution as an illustration», *The American Statistician*, 49, 269.

Jones, M. C. (1999). «Distributional relationships arising from simple trigonometric formulas», *The American Statistician*, 53, 99-102.

Lienahrd, C. W. (1996). «Sample sizes and the central limit theorem: The Poisson distribution as an illustration» (Revisited). *The American Statistician*, 50, 282.

Mantel, N. (1972). «A property of certain distributions», *The American Statistician*, 26, 29-30.

Rao, C. R. (1973). *Linear Statistical Inference and its Applications*, New York: Wiley.

QÜESTIIÓ, vol. 26, 1-2, p. 289-305, 2002

HOJAS DE CÁLCULO PARA LA SIMULACIÓN DE REDES DE NEURONAS ARTIFICIALES (RNA)

J. GARCÍA[1], A. M. LÓPEZ[2,6]
J. E. ROMERO[3], A. R. GARCÍA[4]
C. CAMACHO[2], J. L. CANTERO[5]
M. ATIENZA[5] y R. SALAS[5]

La utilización de Redes de Neuronas Artificiales (RNA) en problemas de predicción de series de tiempo, clasificación y reconocimiento de patrones ha aumentado considerablemente en los últimos años. Programas informáticos de matemáticas de propósito general tales como MATLAB, MATHCAD y aplicaciones estadísticas como SPSS y S-PLUS incorporan herramientas que permiten implementar RNAs. A esta oferta de software hay que añadir programas específicos como NeuralWare, EasyNN o Neuron. Desde un punto de vista educativo, el acceso de los estudiantes a estos programas puede ser difícil dado que no están pensadas como herramientas didácticas. Por otro lado, las hojas de cálculo como Excel y Gnumeric incorporan utilidades que permiten implementar RNAs y son de fácil acceso para los estudiantes. El objetivo de este trabajo es proporcionar un pequeño tutorial sobre la utilización de Excel para implementar una RNA que nos permita ajustar los valores de una serie de tiempo correspondiente a actividad cerebral alfa y que permita al alumno entender el funcionamiento de estos dispositivos de cálculo.

Spreadsmeet for the simulation of artificial neural networks (ANNs)

Palabras clave: Redes de neuronas artificiales, hoja de cálculo, aprendizaje supervisado

Clasificación AMS (MSC 2000): 97U70

[1] Departamento de Ingeniería del Diseño. Universidad de Sevilla
[2] Departamento de Psicología Experimental. Universidad de Sevilla
[3] Departamento de Economía Aplicada I. Universidad de Sevilla
[4] I.E.S. Los Viveros. Sevilla
[5] Laboratorio de Sueño y Cognición
[6] La correspondencia debe dirigirse a Ana María López Jiménez, Departamento de Psicología Experimental, Universidad de Sevilla, Avda. Camilo José Cela s/n. 41005 Sevilla, España. Telf: (34) 954557812. Fax: (34) 954551784. E-mail: analopez@us.es

–Recibido en marzo de 2001.
–Aceptado en diciembre de 2001.

1. INTRODUCCIÓN

El uso cada vez más frecuente de redes de neuronas artificiales (en adelante RNA) en distintos campos es la razón de que nos planteemos la necesidad de incluir esta temática en asignaturas optativas de análisis de datos, de estadística y/o en los programas de doctorado de nuestros departamentos. Sin pretender ser exhaustivos, pensamos que las razones de este incremento hay que buscarlas en que

- son procedimientos muy flexibles de optimización no lineal,
- son, o pueden ser, una alternativa no paramétrica a las técnicas de análisis multivariante, y
- el número de aplicaciones informáticas que permiten implementarlas es cada vez mayor.

No obstante en el contexto de aprendizaje en el que insertamos este trabajo nos encontramos con que tanto los programas informáticos de matemáticas de propósito general (MATLAB, MATHEMATICA, IMSL, etc.) como los estadísticos (SPSS, S-PLUS, SAS, etc.) que tienen implementados módulos de RNA, así como las aplicaciones específicas (EasyNN, Neuron, NeuroShell, NeuralWare, etc.), son, a veces, de difícil acceso para los alumnos y que, además, funcionan como una «caja negra» lo que no facilita la comprensión de las RNA. Este trabajo pretende solventar, en parte, esos problemas mostrando la implementación de una RNA en una aplicación informática muy difundida como es Excel. El uso de hojas de cálculo va a permitir al alumno entender mejor el funcionamiento de una RNA y le va a facilitar el acceso a una metodología cada vez más extendida.

En los apartados que siguen desarrollaremos los conceptos básicos de las RNA y un pequeño tutorial para mostrar la implementación de una RNA en Excel.

2. REDES DE NEURONAS ARTIFICIALES (RNA)

Las RNA son dispositivos de cálculo inspirados en las redes de neuronas biológicas. Como estas últimas, están constituidas por elementos simples denominados *nodos* o *neuronas* organizados en capas y altamente interconectados. A la forma particular de organizarse y conectarse las neuronas se le denomina *arquitectura* o *topología de red.* A cada conexión se le asigna un peso numérico que va a constituir el principal recurso de memoria a largo plazo de las RNA. El aprendizaje se realiza, usualmente, con la actualización de los pesos mediante una determinada regla de aprendizaje.

Vamos a describir, brevemente, las características más importantes de las RNA. Para mayor información sobre RNA pueden consultarse: Bishop, 1995; Cheng y Tittering-

ton, 1994; Hilera y Martínez, 1995; Kohonen, 1989; Michie, Spiegelhalter y Taylor, 1994; Ripley, 1996; Smith, 1993; van der Smagt, 1994.

2.1. Neuronas artificiales: funcionamiento y tipos

Una neurona es la unidad básica de procesamiento. Recibe y emite información de otras neuronas y del/hacia el mundo exterior. En cada neurona se realiza un cálculo local y sencillo con las entradas que le proporcionan sus vecinas sin que sea necesario un control global en el conjunto de unidades. La neurona artificial procesa la información que le llega mediante la obtención de dos componentes. El primero es un componente lineal denominado *función de entrada* (in_i), que calcula la suma ponderada de los valores de entrada a la neurona ($in_i = \Sigma w_{ji} a_j$). El segundo es un componente, generalmente, no lineal conocido como *función de activación* o *función de transferencia*, f, que transforma la suma ponderada en el valor de salida de la neurona ($a_i = f(in_i)$). La Figura 1 es un esquema del funcionamiento de una neurona artificial. En dicho esquema a_j representan las entradas a la neurona y w_{ij} los pesos de cada conexión.

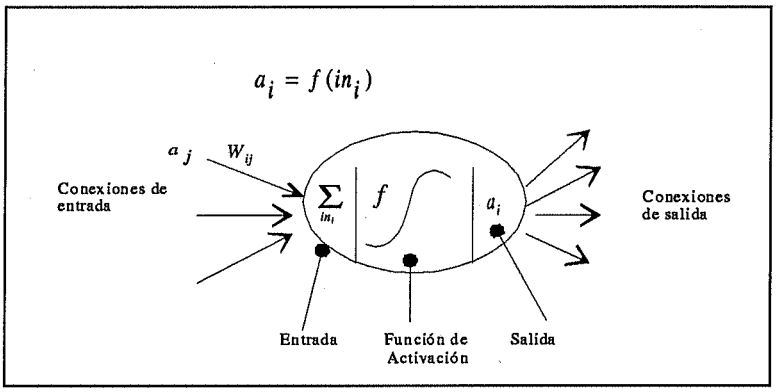


Figura 1. Esquema de una neurona artificial.

La utilización de diferentes funciones matemáticas para f da lugar a distintos tipos de neuronas. Cuatro de las funciones de activación más comunes son: escalón, signo, sigmoidal o logística y lineal.

La *función escalón* o *umbral* tiene un límite, u, de manera que produce un 1 cuando la entrada es mayor que u y en el caso contrario produce 0. La *función signo* produce $+1$ si la entrada es mayor que u y -1 en caso contrario. La *función lineal* o *identidad* deja invariante la entrada. La *función sigmoidal* o *logística* dada por

(1) $$f(in_i) = \frac{1}{1+e^{-in_i}}$$

es la función de activación más usada en las aplicaciones estadísticas de las RNA.

2.2. Arquitecturas de Red

Para caracterizar a una RNA, además del funcionamiento de los elementos que la constituyen, es necesario describir como se organizan y conectan dichos elementos.

Las neuronas se suelen organizar en columnas paralelas denominadas *capas*. Es útil distinguir entre tres tipos de capas: entrada, salida y ocultas. Las neuronas de la *capa de entrada* reciben señales del entorno y no realizan más operación que la de distribuir estas señales a las neuronas de la capa siguiente. Las neuronas de la *capa de salida* procesan la información y envían la señal fuera de la red. Las *unidades ocultas* son aquellas cuyas entradas y salidas se encuentran dentro de la red. Las neuronas de capas distintas pueden estar conectadas unidireccionalmente o bidireccionalmente. Las redes que sólo contienen relaciones unidireccionales se denominan *estáticas* o *redes acíclicas*. Las redes con relaciones bidireccionales se denominan redes *dinámicas* o *cíclicas*. En las redes estáticas la información fluye en un solo sentido (ni lateralmente ni hacia atrás). Entre las redes dinámicas se pueden distinguir las que contienen relaciones bidireccionales entre los nodos de la misma capa, denominadas redes *competitivas*, y redes con relaciones bidireccionales entre neuronas de distintas capas, denominadas redes *recurrentes*. A su vez, las neuronas de distintas capas pueden estar total o parcialmente conectadas. A las redes con conexiones unidireccionales y totalmente conectadas se les denomina *perceptrones multicapa* o redes *feedforward* y son, junto a las redes competitivas, las más utilizadas en análisis de datos. La Figura 2 muestra tres redes con distintos tipos de conexiones.

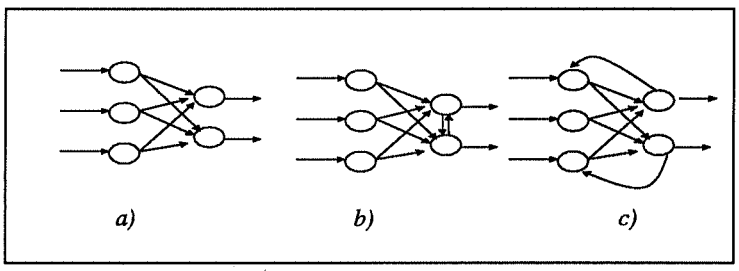


Figura 2. Tres arquitecturas de redes de neuronas artificiales: *a)* Red feedforward, *b)* red competitiva y *c)* red recurrente.

2.3. Aprendizaje de una RNA: tipos y reglas

La función que realiza una RNA depende de la función de activación y de las conexiones entre nodos. Al procedimiento mediante el que se modifica el valor de los pesos de conexión para obtener la salida deseada se le denomina *aprendizaje* o *entrenamiento*. Desde un punto de vista muy global se distinguen dos tipos de aprendizaje: *supervisado* y *no supervisado*. El aprendizaje supervisado (o aprendizaje con maestro) trata de conseguir que la red sea capaz de predecir, a partir de un conjunto de características suministradas como entradas, el valor que tomarán otras características, llamadas *objetivo*, habiendo sido observadas ambos tipos de características en cada uno de los patrones de entrenamiento. El aprendizaje se lleva a cabo modificando los pesos de la red según alguna regla que trata de minimizar la discrepancia entre la salida de la red y la salida correcta proporcionada por los objetivos. El proceso de aprendizaje comienza una vez definida la arquitectura de la red. A los datos utilizados en el proceso de aprendizaje se les denomina *conjunto de entrenamiento*. Durante el aprendizaje se procesan varias veces los pares de vectores del conjunto de entrenamiento y en cada paso del aprendizaje se modifican los pesos de las conexiones. Terminado el proceso de aprendizaje, la red se evalúa sobre un conjunto de datos no utilizados en el entrenamiento y denominado *conjunto de validación*.

En el aprendizaje no supervisado existen características de entrada pero no objetivo (aprendizaje sin maestro). No existe información que indique si la salida que produce la red es o no correcta. Las técnicas de aprendizaje no supervisado se utilizan sobre todo para obtener estructuras, relaciones o clasificaciones (*cluster analysis*).

Existen muchos algoritmos de aprendizaje pero los más extendidos son la regla delta generalizada (*backpropagation*) para el aprendizaje supervisado y la regla estándar del aprendizaje competitivo en el caso no supervisado.

3. USANDO EXCEL PARA IMPLEMENTAR UNA RNA

3.1. Planteamiento del problema

Las posibilidades de aplicación de las RNA a problemas de investigación abordados tradicionalmente desde las técnicas estadísticas clásicas son múltiples. El problema que pretendemos abordar, en este trabajo, con la implementación de una RNA en Excel es el ajuste a una serie de tiempo empírica.

La serie seleccionada corresponde a ritmo cerebral alfa (7.5-12.8 Hz) extraída del EGG de un estudiante de psicología. Se utilizó una frecuencia de muestreo de 200 Hz (período de muestreo de 5 milisegundos) y la señal original ($N = 1024$) se multiplicó por 20000 (ganancia) para que fuera observada sin problemas.

3.2. Topología de la RNA utilizada

Para ajustar los valores de la serie hemos utilizado un perceptron multicapa con tres capas (entrada-oculta-salida). La capa de entrada está constituida por tres neuronas correspondientes a los valores de la serie y sus retardos de lag 1 y 2. La capa oculta está constituida por dos neuronas. Cada neurona de esta capa recibe el vector $\mathbf{y} = (y_{t-2}, y_{t-1}, y_t)$ lo procesa y el resultado lo envía a la neurona que constituye la capa de salida. En la Figura 3 hemos representado la arquitectura de la red utilizada para el ajuste.

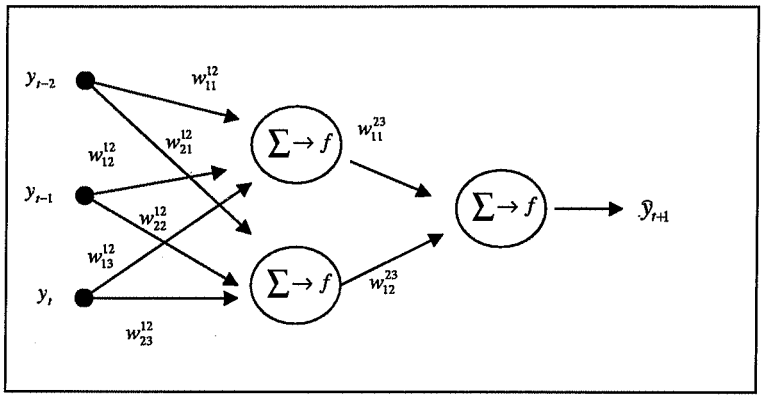


Figura 3. Perceptron multicapa.

El número de capas de la RNA utilizada es el mínimo compatible con la propiedad del aproximador universal.

Kolmogorov en 1957, demostró que toda función continua de varias variables definida sobre un compacto (cerrado y acotado) puede ser representada como la superposición de funciones de una variable. Este teorema, en términos de RNA, se traduce en que toda función $f: C \subseteq \mathbb{R}^p \longrightarrow \mathbb{R}$, C compacto, puede ser representada exactamente por una RNA tipo perceptron multicapa con dos capas ocultas, la primera constituida por $p(2p+1)$ nodos y la segunda por $2p+1$ nodos. Posteriormente, se ha demostrado que los perceptrones multicapa con una sola capa oculta verifican interesantes propiedades teóricas, como la del aproximador universal, según la cual toda función $f: C \subseteq \mathbb{R}^p \longrightarrow \mathbb{R}^q$, C compacto, puede ser aproximada uniformemente mediante un perceptron multinivel con una capa oculta, con función de activación logística en dicha capa y función de activación identidad en la capa de salida (ver Cybenko, 1989). Esta última propiedad, junto con su simplicidad, justifica que el perceptron multinivel con una sola capa oculta y función de transferencia logística sea el modelo utilizado para el desarrollo de nuestra red.

La composición de la capa de entrada obedece a resultados de estudios exploratorios previos acerca del alcance de la dependencia temporal en la serie.

3.3. Implementación de la RNA en Excel

En este apartado describiremos la secuencia de operaciones realizadas en Excel para implementar la RNA de la Figura 3. Partiremos de la Tabla 1 que contiene una porción de la serie empírica y el resultado de la secuencia de operaciones necesarias para resolver mediante la RNA el problema planteado.

Tabla 1. Una porción de Excel mostrando las operaciones necesarias para implementar la red de la Figura 3.

	A	B	C	D	E	F	G	H	I	J
1	ALFA	PREP	IN1A	IN2A	ON1A	ON2A	OR	DIF	DIFCUA	POSP
2	516		−0,0565855	−0,308024						
3	508		0,24836293	3,7925034						
4	478		0,02437906	−5,179935						
5	446		2,62826963	−2,61636						
6	416	−1,35								
7	400	−1,12								
8	386	−1,06	−0,2280374	1,6489059	0,4432364	0,8387431	−1,029509	−0,154135	0,0237576	387,10539
9	384	−0,88	−0,2209722	0,8622244	0,4449806	0,7031252	−0,670099	0,0441864	0,0019524	389,25279
10	384	−0,71	−0,1748784	0,7064098	0,4563915	0,6696074	−0,552414	−0,492653	0,2427065	386,2662
11	388	−0,06	−0,1293256	−2,129734	0,4677136	0,1062402	0,9513148	1,6221352	2,6313227	394,76742
12	384	−0,67	0,00922171	3,468178	0,5023054	0,9697686	−1,21707	0,4702534	0,2211382	383,02284
13	374	−1,69	−0,2043606	6,2145431	0,4490869	0,9980039	−1,430816	0,2087081	0,0435591	375,33778
14	356	−1,64	−0,4210798	2,3000779	0,3962584	0,9088835	−1,336492	0,1292379	0,0167024	359,91837
15	336	−1,47	−0,3474521	1,8942239	0,4140004	0,8692364	−1,18613	0,2798095	0,0782933	342,02795
16	304	−1,47	−0,306998	2,5396997	0,4238477	0,9268785	−1,311062	0,113659	0,0129184	308,94447
17	266	−1,42	−0,3158794	2,2718589	0,4216803	0,9065194	−1,263491	0,0856594	0,0073375	272,92573
18	228	−1,35	−0,3037879	2,0368013	0,4246318	0,8846072	−1,198404	0,1121967	0,0125881	235,82143
19	186	−1,31	−0,2864118	2,1110148	0,4288826	0,8919692	−1,206493	0,1068235	0,0114113	192,19591
20	140	−1,31	−0,2811796	2,24801	0,4301646	0,9044787	−1,235853	0,0162587	0,0002643	145,00177
21	102	−1,25	−0,2825435	1,9087969	0,4298303	0,8708839	−1,148836	0,0376325	0,0014162	108,79648
22	70	−1,19	−0,2655885	1,8017232	0,4339904	0,8583586	−1,105131	0,0711351	0,0050602	75,15527
23	40	−1,18	−0,2524996	1,9789778	0,4372084	0,8785722	−1,149559	0,1904805	0,0362828	41,53483
24	−6	−1,34	−0,2576729	2,8457877	0,4359358	0,9451005	−1,326966	0,0614478	0,0037758	−5,266328
25	−58	−1,39	−0,3001049	2,4721063	0,4255318	0,9221631	−1,294298	0,0414517	0,0017182	−52,06189
26	−110	−1,34	−0,301568	2,0662974	0,4251742	0,8875841	−1,204767	0,0230725	0,0005323	−100,4686
27	−152	−1,23	−0,2831202	1,7219572	0,429689	0,8483808	−1,090331	0,0567748	0,0032234	−141,3831
28	−184	−1,15	−0,257331	1,6967916	0,4360199	0,8451152	−1,065148	0,0645138	0,004162	−178,1413
29	−212	−1,13	−0,2429607	1,8793754	0,4395569	0,8675394	−1,114521	0,2100056	0,0441023	−211,0779
30	−258	−1,32	−0,2479472	2,9300539	0,4383288	0,9493123	−1,331696	0,1101061	0,0121233	−258,4049

En la primera fila de la Tabla 1, hemos escrito etiquetas representativas del contenido y de las operaciones realizadas en cada una de las columnas.

La columna A de la Tabla 1 contiene los valores de la serie (celdas A2 a A1025). Previo al ajuste con la RNA, la serie fue filtrada y estandarizada (Columna B: PREP). El filtrado obedece a la necesidad de eliminar ruido y destacar la posible señal; y, dado el recorrido de la función de activación logística empleada, es también necesario, para no perjudicar la capacidad de aprendizaje de la red, que los valores a estudiar no sean muy elevados por lo que estandarizamos la serie.

La columna B muestra los valores suavizados de la serie obtenidos mediante la expresión (2).

$$PREP = \frac{y_{t+5} - \frac{\sum_{i=1}^{5} y_{t+i}}{5}}{S} \tag{2}$$

En (2), S es la desviación típica para los cinco valores promediados.

Para obtener los valores de la columna B colocamos el cursor en la celda B6 e introducimos la fórmula (2) con la siguiente sintaxis:

= ($A6-PROMEDIO($A2:$A6))/DESVEST($A2:$A6)

Copiando esta fórmula en las celdas B7 a B1025 obtenemos los valores de la serie suavizada (Columna B de la Tabla 1).

Las celdas C2 a C5 y D2 a D5 (Fila 5 de la Tabla 1) contienen los valores de los pesos de la RNA de la Figura 3 tras el entrenamiento de la RNA. Concretamente, en las celdas C2 a C4 (Filas 2 a 4 de la Tabla 1), hemos definido los pesos de las conexiones de los nodos de la capa de entrada con el primer nodo (IN1A) de la capa oculta. Los valores de las celdas D2 a D4 (Filas 2 a 4 de la Tabla 1) corresponden a los pesos de las conexiones de los nodos de la capa de entrada con el segundo nodo (IN2A) de la capa oculta. Las celdas C5 y D5 contienen los pesos de conexión de los nodos de la capa oculta (IN1A e IN2A) con el nodo de la capa de salida. Los pesos iniciales de la red pueden fijarse en cualquier valor. En nuestro caso se fijaron en 0.5.

En las celdas C8 a C1024 y D8 a D1024 hemos calculado el input total a las neuronas de la capa oculta. El input total (in_i) es la suma ponderada de los valores de la serie. En nuestro ejemplo hemos utilizado la serie suavizada. Para calcular la entrada a la primera neurona de la capa oculta (Columna C de la Tabla 1: IN1A), nos situamos en la celda C8 y escribimos la siguiente expresión:

(3) = $C$2*$B6+$C$3*$B7+$C$4*$B8

Copiando (3) en las celdas C9 a C1024 obtenemos los valores de la Columna C correspondientes a la entrada neta a la neurona 1 (IN1A) para el conjunto de patrones de entrenamiento.

De manera análoga, para calcular la entrada neta a la segunda neurona de la capa oculta (IN2A), nos situamos en la celda D8 y escribimos

(4) $$= \$D\$2*\$B6+\$D\$3*\$B7+\$D\$4*\$B8$$

Copiando (4) en las celdas D9 a D1024, obtenemos los valores de la columna D de la Tabla 1.

Las salidas de las neuronas IN1A e IN2A se obtienen aplicando la función de activación seleccionada a los valores de las celdas C8 a C1024 y D8 a D1024. En nuestro ejemplo hemos utilizado la función logística (1) por las razones comentadas en el apartado 3b.

Los valores de las columnas E y F, etiquetadas en la Tabla 1 como ON1A y ON2A respectivamente, corresponden a las salidas de las neuronas de la capa oculta para el conjunto de patrones de entrenamiento. Para obtener los valores de la columna E nos situamos en la celda E8 e insertamos la fórmula:

(5) $$= 1/(1+\mathrm{EXP}(-\$C8))$$

Copiando (5) en las celdas E9 a E1024 se obtienen el resto de los valores.

Para los valores de la columna F se procede de la misma manera. Situándonos en la celda F8 escribimos la fórmula

(6) $$= 1/(1+\mathrm{EXP}(-\$D8))$$

y copiamos (6) en las celdas F9 a F1024.

La columna G (OR) de la Tabla 1 contiene la salida de la red. Este valor es la suma ponderada de las entradas al nodo de la capa de salida. Para obtener los valores de la columna G multiplicamos las columnas E y F de la Tabla 1 por los pesos de las celdas C5 y D5. Situándonos en la celda G8 escribimos la expresión:

(7) $$= \$C\$5*\$E8+\$D\$5*\$F8$$

Copiando (7) en las celdas G9 a G1024 obtenemos la salida de la RNA para el conjunto de patrones de entrenamiento. El nodo de la capa de salida, en esta red, sólo calcula la suma ponderada de las entradas, y el valor resultante es la salida de la red (OR).

Los valores de la columna H (DIF) en la Tabla 1 corresponden al error de predicción esto es, a la diferencia entre la serie empírica suavizada (Columna B: PREP) y la salida

de la RNA (Columna G: OR). Para calcular dichos errores, situándonos en la celda H8, restamos las columnas G (OR) y B(PREP) utilizando la expresión:

(8) $= \$G8-\$B9$

Copiando (8) en las celdas G9 a G1024 se obtienen el resto de los valores de la columna H.

En la columna I (DIFCUA) hemos calculado el cuadrado de los errores de predicción. Para ello en la celda I8 hemos escrito la siguiente fórmula:

(9) $= \$H8*\$H8$

Copiando (9) en las celdas I9 a I1024 se obtienen los cuadrados de los errores de predicción para el conjunto de patrones de entrenamiento.

En la celda J1025 hemos calculado la suma de cuadrados de los errores (Σe^2). La fórmula para calcular esta suma es:

(10) SUMA(\$I8:\$I1024).

Los valores de los pesos de la red se modificarán en el sentido de hacer mínimo el valor de la celda J1025.

3.4. Cálculo de los valores óptimos de los pesos de la RNA

El problema de aprendizaje de la RNA se formula en términos de la minimización de la función de error (10). Este error depende de los parámetros adaptativos de la red. El proceso de aprendizaje se lleva a cabo modificando los pesos de manera que la suma de cuadrados de los errores sea mínima para el conjunto de patrones de entrenamiento. Para resolver este problema Excel dispone de una macro denominada «Solver» que permite obtener el valor máximo o mínimo de una celda modificando los valores de otras celdas.

Para ejecutar «Solver» seleccionamos, en el menú «Herramientas», dicha opción. Cliqueando sobre «Solver» abrimos el cuadro de diálogo: «Parámetros de Solver». En este cuadro, insertamos la celda objetivo (\$J\$1025) y el rango de celdas a modificar (\$C\$2:\$D\$5). Podemos elegir, también, el método de resolución para el problema de mínimos planteado. Cliqueando en «Opciones» del cuadro de diálogo «Parámetros de Solver» abrimos el cuadro «Opciones de Solver» en este cuadro podemos modificar las opciones por defecto de los siguiente parámetros:

Tiempo: Esta opción establece un limite temporal para encontrar la solución al problema. En nuestro ejemplo hemos elegido la opción por defecto: 100 segundos.

Iteraciones: Esta opción establece un límite al número de cálculos provisionales. Este valor es el número de ciclos de aprendizaje (epochs) de la red. El valor por defecto es 100.

Precisión: Esta opción controla la precisión de la solución. La precisión debe indicarse como una fracción entre 0 y 1.

Tolerancia: Es el porcentaje máximo de discrepancia entre el valor de la celda objetivo y las restricciones externas si las hubiese. Valores grandes de tolerancia aceleran el proceso de solución. Nosotros hemos utilizados el valor por defecto: 50.

Convergencia: Si el valor del cambio relativo en la celda objetivo es más pequeño que el valor introducido en esta opción en las cinco últimas iteraciones «Solver» se detendrá. Esta opción sólo se aplica a problemas no lineales y debe indicarse mediante una fracción entre 0 y 1.

Además de las opciones anteriores, hemos seleccionado el método de Newton para la aproximación de raíces, derivadas progresivas y estimación cuadrática. La selección de este conjunto de opciones obedece a que suelen requerir menos tiempo de cálculo. No obstante, en el caso de la selección del método de aproximación de raíces, la elección del método de Newton o del método del gradiente conjugado no es crucial ya que «Solver» es capaz de cambiar automáticamente entre ambos métodos en función de las necesidades de almacenamiento que requiera el problema que se esté resolviendo.

Una vez que hemos dado valores a las opciones anteriores pulsamos en «Aceptar» y volvemos al cuadro «Parámetros de Solver». En este cuadro, pulsamos en «Resolver» y obtenemos la solución al problema planteado.

Los valores finales de los pesos después de 100 iteraciones son los correspondientes a las celdas C2 a C5 y D2 a D5 de la Tabla 1.

Finalmente, en la columna J de la Tabla 1, efectuamos la operación inversa a la realizada para obtener los valores de la serie suavizada (columna B:PREP). Para realizar esta operación, nos situamos en la celda J8 y escribimos la siguiente expresión:

(11) = $G8*DESVEST($A4:$A8)+PROMEDIO($A4:$A8)

La expresión (11) la copiamos en las celdas J9 a J1024 y obtenemos los valores de la columna J.

3.5. Valoración del ajuste obtenido

Para evaluar el grado de ajuste de la red hemos realizado un gráfico de líneas utilizando el asistente para gráficos de Excel. Hemos representado la serie de la columna A (ALPHA) para el rango de valores A9 a A1025 y la serie de la columna J (POSP) para el

rango de valores de J8 a J1024 (ver Figura 4). Como puede observarse en la Figura 4, las dos series (ALFA y POSP) están superpuestas por lo que el ajuste es prácticamente perfecto.

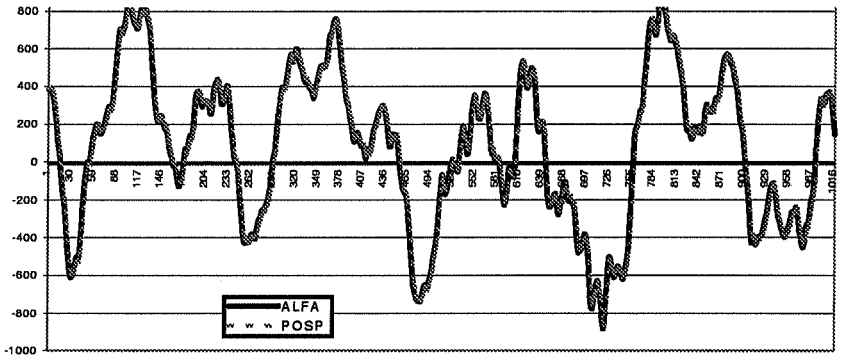


Figura 4. Gráfico de líneas para evaluar el ajuste a la serie empírica.

Hemos calculado también el coeficiente de correlación de Pearson entre las dos series obteniendo un valor de 0,99644869. Para calcular este coeficiente hemos utilizado la siguiente expresión

(12) = COEF.DE.CORREL(A9:A1025;J8:J1024)

3.6. Alternativas

En este trabajo hemos utilizado todos los valores de la serie para entrenar a la RNA. Otra posibilidad, de cara a validar la ejecución de la RNA y a compararla con otros procedimientos de ajuste y predicción de series de tiempo, consiste en utilizar parte de la serie, por ejemplo el rango B8 a B1000 para entrenar a la RNA y los valores restantes utilizarlos como conjunto de validación. La implementación es la expuesta en este trabajo salvo que, una vez terminado el proceso de adaptación de los pesos, las fórmulas correspondientes a las distintas columna de la Tabla 1 hay que copiarlas en las celdas del conjunto de validación.

Para el ajuste y predicción de series de tiempo pueden utilizarse técnicas estadísticas clásicas como modelos ARIMA, análisis espectral, métodos exponenciales, etc. No es el objetivo de este trabajo comparar la ejecución de estos métodos con la ejecución de la metodología basada en redes de neuronas.

4. CONCLUSIONES

En este trabajo hemos pretendido mostrar las posibilidades didácticas del uso de Excel para implementar redes de neuronas artificiales. Hemos elegido un problema de ajuste a una serie empírica; pero, de la misma manera, pueden implementarse redes para resolver problemas de reducción de datos (análisis de componentes principales), clasificación y discriminación. La utilización de hojas de cálculo como herramientas didácticas para la implementación de RNA tiene, frente a programas específicos, una serie de ventajas que podríamos resumir en: 1º) permiten el acceso a una metodología cada vez más utilizada, las RNA, a través de aplicaciones de uso frecuente y de fácil acceso (hojas de cálculo) y 2º) aumentan las posibilidades de aprender y entender estos dispositivos de cálculo.

5. REFERENCIAS

Bishop, C. M. (1995). *Neural networks for pattern recognition.* Oxford: Oxford University Press.

Chen, B. y Titterington, D. M. (1994). «Neural networks: a review from a statistical perspective». *Statistical Science*, 9, 1, 2-59.

Cybenko, G. (1989). «Aproximation by superposition of a sigmoidal function». *Mathematics of Control, Signal and Systems*, 2, 303-314.

Hilera, J. R. y Martínez, V. J. (1995). *Redes Neuronales Artificiales. Fundamentos, Modelos y Aplicaciones.* Madrid: Ra-Ma.

Kohonen, T. (1989). *Self-organization and associative memory.* Berlín: Springer-Verlag.

Michie, D., Spiegelhalter, D. J. y Taylor, C. C. (1994). *Machine Learning, Neural and Statistical Classification.* Disponible en: http://www.amsta.leeds.ac.uk/ charles/statlog/

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks.* Cambridge: Cambridge University Press.

Smith, M. (1995). *Neural networks for statistical modeling.* New York: Van Nostrand Reinhold.

Van der Smagt, P. P. (1994). «Minimization methods for training feedforward neural networks». *Neural Networks*, 7(1), 1-11.

Los archivos de este trabajo están disponibles en analopez@us.es

Queremos agradecer al profesor Antoni Solanas sus acertados comentarios a este manuscrito.

ENGLISH SUMMARY

SPREADSMEET FOR THE SIMULATION OF ARTIFICIAL NEURAL NETWORKS (ANNs)

J. GARCÍA[1], A. M. LÓPEZ[2,6]
J. E. ROMERO[3], A. R. GARCÍA[4]
C. CAMACHO[2], J. L. CANTERO[5]
M. ATIENZA[5] and R. SALAS[5]

In recent years, the use of Artificial Neural Networks or ANNs has increased considerably to solve prediction problems in time series, classification and recognition of patterns. General-purpose mathematical programs such as MATLAB, MATHCAD and mathematical and statistical programs such as SPSS and S-PLUS incorporate tools that allow the implementation of ANNs. In addition to these, specific programs such as NeuralWare, EasyNN, or Neuron, complete the software offer using ANNs.
From an educational point of view, an aspect that concerns the authors of this work, student access to these programs can be expensive or, in some case, unadvisable given the few possibilities they provide as didactic instruments. These programs are usually easy to use but do not facilitate the understanding of the technique used. On the other hand, spreadsheets like Excel or Gnumeric incorporate tools that allow all of the necessary calculations to implement an ANN. These programs are user-friendly to the point that they are used by university laboratories, as well as psychology, economic science, and engineering students, to mention a few. This paper provides a small tutorial on the use of a spreadsheet, specifically Excel, to implement an ANN to adjust the values of a time series corresponding to cerebral alpha activity.

Keywords: Artificial neural network, spreadsheet, supervised learning

AMS Classification (MSC 2000): 97U70

[1] Departamento de Ingeniería del Diseño. Universidad de Sevilla
[2] Departamento de Psicología Experimental. Universidad de Sevilla
[3] Departamento de Economía Aplicada I. Universidad de Sevilla
[4] I.E.S. Los Viveros. Sevilla
[5] Laboratorio de Sueño y Cognición
[6] La correspondencia debe dirigirse a Ana María López Jiménez, Departamento de Psicología Experimental, Universidad de Sevilla, Avda. Camilo José Cela s/n. 41005 Sevilla, España. Telf: (34) 954557812. Fax: (34) 954551784. E-mail: analopez@us.es

–Received March 2001.
–Accepted December 2001.

In recent years, the use of Artificial Neural Networks or ANNs has increased considerably in order to solve prediction problems in time series, classification and recognition of patterns. We believe that this increase is due to the difficulty of solving this type of problem when the behavior of the series is nonlinear or when the classes are not lineally separable. Another reason is the access to numerous relatively user-friendly software applications allowing the implementation of frequent use networks.

General-purpose mathematical programs such as MATLAB, MATHCAD and mathematical and statistical programs such as SPSS and S-PLUS incorporate tools that allow the implementation of ANNs. In addition to these, specific programs such as NeuralWare, EasyNN, or Neuron, complete the software offer using ANNs.

From an educational point of view —an aspect that concerns the authors of this work— student access to these programs can be expensive or, in some case, unadvisable given the few possibilities they provide as teaching instruments. These programs are usually easy to use but do not facilitate the understanding of the technique. On the other hand, spreadsheets like Excel, or Gnumeric incorporate tools that allow all of the necessary calculations to implement an ANN. These programs are user-friendly to the point that they are used by University laboratories, as well as psychology, economic science, and engineering students, to mention a only few.

In the first part of this work, we introduce the basics aspects of ANNs: elements, architectures, and types and rules of learning. In the second part, we describe the necessary operations to implement in Excel an ANN to adjust the values of a time series corresponding to cerebral alpha activity.

The ANN to implement consists of three layers. The input layer is composed of three nodes corresponding to the values of the series (y_{t-2}, y_{t-1}, y_t). The hidden layer is composed of two nodes, each of which applies the logistic function as the activation function and, finally, the output layer with only one node and linear activation function.

To perform the task, the ANN modifies the synaptic weights (w_{ij}) adjust so that the value of the function

$$\sum (y_{t+1} - \hat{y}_{t+1})^2$$

is minimal.

Table 1 denotes a portion of Excel to implement the net.

The column A of Table 1 contains the values of the series (cells A2 to A1025) and the column B shows the softened values of the series.

Columns C and D of Table 1 contain the values of the weights of the net to implement. Specifically in cells C2 to C4, we have defined the weights of the connections of the nodes of the input layer with the first node (IN1A) of the hidden layer. The values of

cells D2 to D4 correspond to the weights of the connections of the nodes of the input layer with the second node (IN2A) of the hidden layer. Cells C5 and D5 contain the values of the weights of the connection of node IN1A and IN2A of the hidden layer with the neuron of the output layer. Initially we can fix the weights at any value. In cells C8 to C1024 and D8 to D1024 we have calculated the total input to nodes 1 and 2 of the hidden layer. The total input is the weight sums of the values of the series.

Table 1. A portion of Excel.

	A	B	C	D	E	F	G	H	I	J
1	ALFA	PREP	IN1A	IN2A	ON1A	ON2A	OR	DIF	DIFCUA	POSP
2	516		−0,0565855	−0,308024						
3	508		0,24836293	3,7925034						
4	478		0,02437906	−5,179935						
5	446		2,62826963	−2,61636						
6	416	−1,35								
7	400	−1,12								
8	386	−1,06	−0,2280374	1,6489059	0,4432364	0,8387431	−1,029509	−0,154135	0,0237576	387,10539
9	384	−0,88	−0,2209722	0,8622244	0,4449806	0,7031252	−0,670099	0,0441864	0,0019524	389,25279
10	384	−0,71	−0,1748784	0,7064098	0,4563915	0,6696074	−0,552414	−0,492653	0,2427065	386,2662
11	388	−0,06	−0,1293256	−2,129734	0,4677136	0,1062402	0,9513148	1,6221352	2,6313227	394,76742
12	384	−0,67	0,00922171	3,468178	0,5023054	0,9697686	−1,21707	0,4702534	0,2211382	383,02284
13	374	−1,69	−0,2043606	6,2145431	0,4490869	0,9980039	−1,430816	0,2087081	0,0435591	375,33778
14	356	−1,64	−0,4210798	2,3000779	0,3962584	0,9088835	−1,336492	0,1292379	0,0167024	359,91837
15	336	−1,47	−0,3474521	1,8942239	0,4140004	0,8692364	−1,18613	0,2798095	0,0782933	342,02795
16	304	−1,47	−0,306998	2,5396997	0,4238477	0,9268785	−1,311062	0,113659	0,0129184	308,94447
17	266	−1,42	−0,3158794	2,2718589	0,4216803	0,9065194	−1,263491	0,0856594	0,0073375	272,92573
18	228	−1,35	−0,3037879	2,0368013	0,4246318	0,8846072	−1,198404	0,1121967	0,0125881	235,82143
19	186	−1,31	−0,2864118	2,1110148	0,4288826	0,8919692	−1,206493	0,1068235	0,0114113	192,19591
20	140	−1,31	−0,2811796	2,24801	0,4301646	0,9044787	−1,235853	0,0162587	0,0002643	145,00177
21	102	−1,25	−0,2825435	1,9087969	0,4298303	0,8708839	−1,148836	0,0376325	0,0014162	108,79648
22	70	−1,19	−0,2655885	1,8017232	0,4339904	0,8583586	−1,105131	0,0711351	0,0050602	75,15527
23	40	−1,18	−0,2524996	1,9789778	0,4372084	0,8785722	−1,149559	0,1904805	0,0362828	41,53483
24	−6	−1,34	−0,2576729	2,8457877	0,4359358	0,9451005	−1,326966	0,0614478	0,0037758	−5,266328
25	−58	−1,39	−0,3001049	2,4721063	0,4255318	0,9221631	−1,294298	0,0414517	0,0017182	−52,06189
26	−110	−1,34	−0,301568	2,0662974	0,4251742	0,8875841	−1,204767	0,0230725	0,0005323	−100,4686
27	−152	−1,23	−0,2831202	1,7219572	0,429689	0,8483808	−1,090331	0,0567748	0,0032234	−141,3831
28	−184	−1,15	−0,257331	1,6967916	0,4360199	0,8451152	−1,065148	0,0645138	0,004162	−178,1413
29	−212	−1,13	−0,2429607	1,8793754	0,4395569	0,8675394	−1,114521	0,2100056	0,0441023	−211,0779
30	−258	−1,32	−0,2479472	2,9300539	0,4383288	0,9493123	−1,331696	0,1101061	0,0121233	−258,4049

The outputs of neurons INI1A and INI2A are obtained by applying the activat ion function selected for the values of columns C and D. Columns E and F are labeled as ON1A and ON2A and correspond to neuron outputs of the hidden layer.

The column G of Table 1 contains the network output. It corresponds to the weighted sum from the inputs to the node of the output layer. To obtain the values of column G, the values of columns E and F are multiplied times the weights of cells C5 and D5. The output layer of the node only calculates the weighted sum of the input and this is the output of the net (OR).

The values of column H in Table 1 correspond to prediction errors (DIF). To calculate these errors we subtract columns G (OR) and B (PREP).

The column I contains the errors to the square (DIFCUA). They are obtained by multiplying the values of column H with the same values. Cell J1025 has calculated the sum of squares of the errors. The formula to calculate this sum is: SUMA ($I8:$I1024). The values of the network weights will be modified in the sense of making the value of this cell minimum.

To solve this problem Excel has a macro denominated: «Solver» that allows the maximum value or minimum of a cell modifying other cells to be determined. To execute this macro in the Tools menu one must click on «Solver». Clicking on «Solver» opens the dialog box: «Parameters to Solver2». In this box, the cell objective ($J$1025) must necessarily be fixed and the range of the cells they modify ($C$2:$D$5). We can also choose the desired resolution method to solve the problem. By clicking on «Options» of the dialog box «Parameters to Solver» we can choose the values to the next options: Maximum time, Iterations, Precision, Tolerance, Convergence. In addition to the previous options, we can choose the method for the approach of roots.

Finally, in column J of Table 1, we have defined the inverse operation of the column B and to evaluate the adjustment of the implemented net, we have carried out a line graph with the Assistant of Excel graphics. We have represented the series of the column A (ALPHA) and the series of the column J (POSP).

Comentaris de llibres

An Introduction to Statistical Modeling of Extreme Values

Stuart Coles

Springer-Verlag, London, 2001
208 pàgines

This is a really interesting, clear, updated and very well documented book. It summarizes an important part of the new research on the extreme value theory and is directly oriented towards real practical application. Most aspects of extreme modeling techniques are covered, including historical and contemporary techniques. A wide range of worked examples using genuine datasets illustrate the various modeling procedures. My personal opinion is clearly positive, I simply like this book.

First of all, I would like to remark the point of view of the author. The first thinks one finds in the book are genuine datasets. Chapter 1 shows the examples that will be deeply studied in the text. A practical point of view is always considered, complemented by the theoretical framework of extreme value models. All the computations are carried out using S-PLUS, and the corresponding datasets and functions are available on Internet Web page: http://www.stats.bris.ac.uk/ masgc/ismev/summary.html

Chapter 2 gives a nice summary of the most updated likelihood methods for inference, including profile likelihood for quantile estimations. This provides a coherent and global method really useful to consider models with parameters which depend on time and covariants. Maximum likelihood methods adjust automatically the complex models, allowing us to quantify the uncertainty of the model and giving us a way to check the goodness of fit. Moreover, Chapter 9, the last one, deals with a brief introduction to more advanced topics as Bayesian inference and Markov chain Monte Carlo methods.

The core of the book is chapter 3 and chapter 4, in which the classical extreme value theory is developed. They make the theory available for statisticians and non-statisticians alike thanks to its elementary treatment, with heuristic proves often replacing more detailed mathematical proves. The material includes the generalized extreme value distributions and the threshold models with a generalized Pareto distribution.

In general, the book is a good complement of a more classical text, because it considers more current statistical inference techniques for using these models in practice.

In chapters 5 and 6, series of dependent observations are studied. The first one develops methods for stationary series, and the other one for non-stationary series. In the first ca-

se, the same methods are used in independent series work. The same limit distributions arise as natural limits, but the degree of accuracy depends on the degree of dependence. Chapter 5 ends with some applications to Dow Jones financial series. In order to study non-stationary series Chapter 6 introduces covariants, order statistics and all kind of information related to the data. The Chapter includes the study of a nice dataset on the annual sea level in Venice.

Chapter 7 shows an elegant formulation of the extreme value behavior from the theory of point process. This is now the newest point of view of the theory. Chapter 8 focuses on the multivariate extremes. The same methods of single theory can be extended, but new problems arise. The dimensionality raises problems for validation and computation. Special attention is payed to the two-dimensional case.

Finally, I would like to say that the extreme value theory is closely related to the heavy tailed distributions theory and this one is now really popular in mathematical finance. In the last years, many new books on extreme values have appeared. I think this one is an essential reference both for students and researchers in statistics and finance, and the book will also appeal to practitioners looking for practical help to solve real problems.

More information about this book can be found in http://www.springer.de/.

Joan del Castillo
Universitat Autònoma de Barcelona

The Elements of Statistical Learning
Data Mining, Inference and Prediction

Trevor Hastie, Robert Tibshirani i Jerome Friedman

Springer-Verlag, 2002
533 pàgines

El libro esta dedicado al apasionante tema de la modelización estadística, en el sentido más general del término, desde la perspectiva de la predicción. Este es un problema de gran actualidad en el entorno de la minería de datos. Esto es, se utilizan los datos para aprender de ellos, sin suponer la existencia de un modelo teórico a confirmar, sino que se pretende utilizar la información disponible para construir modelos que permitan hacer predicciones sobre la(s) variable(s) de respuesta.

La primera característica del libro que sorprende es la alta calidad de la edición, la impresión en color de los gráficos y títulos es sencillamente espectacular. Pero también el contenido esta a la altura. La gran amplitud de los modelos presentados, muchos de ellos sucintamente, pero el nivel conseguido y la claridad en el enfoque, muestran el conocimiento y la experiencia de los autores, fruto de años de investigación y aplicación de los modelos descritos, siendo los apartados de mayor dificultad teórica señalados mediante una tarjeta amarilla (por suerte no aparecen tarjetas rojas). En este sentido y dado el número de modelos diversos que se presentan, el primer problema es ordenarlos mediante un encadenamiento lógico que relacione todos los modelos. Debemos decir que esto lo consiguen en gran parte, siendo este otro de los atractivos del libro, presentar de forma relacionada y comparativa modelos, en principio disconexos.

Otra característica del libro es que intenta unificar la nomenclatura estadística con la utilizada en la comunidad informática de «machine learning» de cara a utilizar un solo léxico para los mismos problemas y sus soluciones. En este sentido es muy útil la presentación en la mayoría de modelos de su correspondiente algoritmo, que facilita su comprensión y permite su implementación mediante una herramienta informática de alto nivel, como S-plus, R o Matlab y a su vez tiende puentes de entendimiento con la comunidad de «machine learning». Señalemos que la solución informática adoptada en el libro es el S-Plus, de la que el libro hace una exuberante demostración de posibilidades estadísticas y gráficas. Los métodos siempre vienen de la mano de su aplicación a problemas reales y actuales, lo cual no es óbice para utilizar datos simulados cuando interesa dilucidar el distinto comportamiento de los modelos sobre unos datos. Utiliza como «datasets» conjuntos de datos disponibles via web, con problemas actuales que

van desde el reconocimiento de la escritura manuscrita, el filtraje de mensajes spam, o la predicción de diferentes tipos de cáncer a partir de información sobre los genes, etc. Lo cual por otro lado, es muy útil desde el punto de vista docente para prácticas de laboratorio, siendo también útiles en este sentido los problemas de final de capítulo. También unifica los problemas de regresión y clasificación bajo un solo marco, puesto que se trata del mismo problema.

A continuación presentamos los «principales» modelos presentados para dar idea de la amplitud de libro, lista que no cubre todos los modelos presentados, pero si da una idea del contenido del libro.

El libro empieza con una presentación de los modelos de predicción más sencillos, los lineales, para hacer a continuación un salto al aproximador universal más flexible, de los k-vecinos más próximos. Este es uno de los mejores momentos del libro, señala las limitaciones de ambos extremos, la simplicidad de los primeros y el problema de la dimensionalidad en los segundos y justifica la necesidad de encontrar modelos situados entre ambos extremos, a lo cual dedica el resto del libro.

La primera generalización que presenta es permitir más flexibilidad en los modelos lineales mediante expansiones del espacio de características, ya sea por funciones polinómicas, por splines de regresión, alisados (smoothing splines), wavelets. Otra forma de superar la rigidez de los modelos lineales es realizando una regresión local en la vecindad de un punto, mediante regresión kernel y su generalización (y simplificación) al problema de clasificación (Naive bayes) y los «radial basis». A continuación los modelos aditivos generalizados y los árboles de decisión. Es interesante la presentación de los árboles de decisión como un modelo aditivo, el cual permitirá más adelante su generalización para el caso multivariante (MARS, Multivariate Adaptative Regression Splines).

El tema de la complejidad de los modelos se trata en el capítulo 7, a mayor complejidad mejor ajuste pero menor poder de generalización, este problema lo soluciona por regularización, esto es, penalizar el criterio a optimizar por la complejidad del modelo, en este sentido señalar la presentación de la «ridge regressión» como un problema de regularización y su generalización en los «smoothing splines». Ligado a la complejidad del modelo está la selección del modelo, al necesario equilibrio entre sesgo y variancia en las predicciones. El criterio a optimizar es en general la suma de cuadrados residuales penalizados, también la entropía en problemas de clasificación, y también la maximización de la verosimilitud y el método bayesiano. El error de predicción se mide mediante métodos heurísticos, de la muestra test, validación cruzada o bootstrap, pero también se explican los criterios AIC, BIC y la dimensión de Vapnik Chernovenkis para la complejidad del modelo.

También se trata los métodos de consenso, tales como el «bagging» (por promedio de las predicciones) como un método minimizar el error de predicción de los modelos, y

el de consenso mejorado, ponderando las observaciones en función de la malclasificación («Boosting»), como una forma de producir un clasificador fuerte a partir de un clasificador débil inicial, el cual se reduce a un modelo aditivo particular.

Finalmente se presentan los modelos de predicción basados en factores construidos a partir de los datos, Projection Pursuit Regression (si bien aquí estaría bien empezar por la regresión sobre componentes principales (PCR) y también la regresión PLS) y su generalización en redes neuronales. También se presentan las generalizaciones del análisis discriminante «Support Vector Machines», el cual construye hiperplanos no lineales separadores y los «flexible discriminants».

Los últimos capítulos se dedican a una serie de técnicas de minería de datos no directamente relacionadas con el problema de predicción, tales como el aprendizaje no supervisado donde se ha incluido las reglas de asociación (Market Basket Analysis) y por supuesto los métodos de «clustering», en particular el «k-means», mapas de Kohonen, «vector quantization» y métodos aglomerativos. También las Componentes Principales no lineales, «Independent Component Análisis», método relacionado con Factor Analysis y Multidimensional Scaling.

Ciertamente es imposible describir en profundidad todos y cada uno de los modelos, pero a menudo es suficiente para empezar en su praxis, si bien un conocimiento previo facilita en gran medida una mejor comprensión.

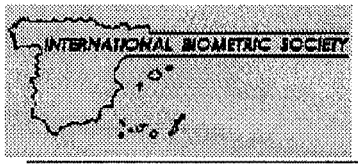
En resumen, se trata de una obra de referencia imprescindible en la biblioteca de cualquier investigador o aplicador de las técnicas de mineria de datos, y que a su vez puede utilizarse como guía docente de varios cursos cuatrimestrales de minería de datos.

Información complementaria sobre este libro se puede encontrar en http://www.springer.de/.

Tomàs Aluja
Universitat Politècnica de Catalunya

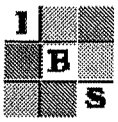
Ressenyes d'activitats institucionals

Sociedad Española de Biometría



Región Española
de la Sociedad Internacional de Biometría

Spanish Region
of the International Biometric Society



http://www.iata.csic.es/ibsresp

La **Sociedad Española de Biometría/Región Española de la Sociedad Internacional de Biometría** (abreviadamente **SEB** o ***REsp***) tiene como objetivos promover, impulsar y difundir el desarrollo y la aplicación de los métodos matemáticos y estadísticos a la biología, medicina, psicología, farmacología, agricultura y otras ciencias afines (ciencias relacionadas con los seres vivos). Cualquier profesional o alumno de estas disciplinas puede ser miembro de la SEB.

Consejo Directivo

Presidente:	María Jesús Bayarri García (Medicina)
Vicepresidente:	Guadalupe Gómez Melis (Biología)
Secretario y Tesorero:	Fernando López Santoveña (Agronomía)
Vocal en calidad de Miembro del Consejo de la IBS:	Amparo Oliver Germes (Psicología)
Vocales:	Antonio López Quílez (Medicina)
	Juan Ferrándiz Ferragud (Biología)
	Purificación Galindo Villardón (Medicina)
	Eduardo García Cueto (Psicología)
	José Luis González Andújar (Agronomía)
	Alex Sánchez Plá (Biología)
Corresponsal de la REsp en el «Biometric Bulletin» de la IBS:	María Luz Calle Rosingana

IX CONFERENCIA ESPAÑOLA DE BIOMETRÍA
se celebrará en A CORUÑA los días 28, 29 y 30 de mayo de 2003
más información en:
http://www.udc.es/dep/mate/biometria2003/index.htm

The TES Institute
Training of European Statisticians

GENERAL INTRODUCTION

The TES Institute is a non-profit making association of 17 European National Statistical Institutes and the University Centre of Luxembourg. It offers a post-graduate vocational training programme for statisticians and other groups working in the statistical environment. The TES Institute today boasts a solid foundation of statistical training experience, built up since the beginning of the nineties.

Our mission is to provide world-wide cutting edge lifelong international training for professional statisticians. The training emphasises the inter country know-how transfer and focuses on the dissemination of best practices in terms of providing statistics compliant with both European and International Statistical System standards.

The training programmes of the TES Institute are designed for both public and private sector statisticians in the broadest sense of the word i.e. university graduates of any discipline involved in statistical work within Statistical Institutes, European Institutions, Government Agencies, National or private Banks, Enterprises, ...

Training methods are problem-oriented and based on a twofold track approach combining both teaching of theoretical concepts and practical sessions using real life cases. Moreover, the «learning-by-doing» process provides an opportunity to respond in an effective way to recent scientific and technological developments as well as new requirements from the labour market.

The training programmes offered by the TES Institute provide both theoretical and practical background but the courses have all a very strong applied character. These programmes also offer participants the opportunity to meet colleagues form all over Europe and other countries since the TES Institute has extended its activities to the Central European, Mediterranean Basin and TACIS countries.

The above characteristics represent the basic conditions to acquire sharper competence in their work environment and highlight the European dimension of their activity.

After ten years of existence, the programme became entire part of the statistical world. For the time being, around 500 participants coming from more than 30 countries are trained every academic year. Such an interest is mainly due to the large number of courses on offer. Indeed, the TES portfolio comprises more than 80 courses of short duration all at post-graduate level.

The TES Institute offers a coherent set of interrelated courses in the following domains:

- Official Statistics: General Issues (OSG)
- Official Economic Statistics (ECO)
- Official Social Statistics (SOC)
- Data Collection and Survey Methodology (DAT)
- Applied Statistical Analysis (ASA)
- Publication and Dissemination of Statistics (PDS)
- Management in a Statistical Institute (MSI)
- Statistical Information Systems (SIS)

To reach the widest audience possible, the TES Institute currently offers courses in English, French, German, Arabic and Russian.

After a few years of co-operation with the Central European countries, the TES Institute has recently extended the co-operation to the MEDSTAT and TACIS region. Such an internationalisation is the direct result of the growing importance of training as a part of the current technological and intellectual development. Therefore, as far as statistics and economics are concerned, it is of the utmost importance to extend the best national practices to an international level. It is obvious that the TES programmes should be considered as a complement and not a substitute to the training provided at national level.

In brief, one may say that by offering training opportunities that are complementary to the ones provided at national level, the TES Institute is offering a new approach of the subsidiarity concept.

TRAINING

The Core Programme of Eurostat

The Core Programme 2002 started in January will end in November. The overview of the remaining courses can be found at the end of the present paper.

The TES Institute is currently preparing the Core Programme 2003 that will start in January 2003 and offer 28 courses. An overview of the Core Programme 2003 will be published in the next issue of Questiio and the final brochure will be available by October 2002.

The Special Courses Programme

Beside the execution of the subsidised annual vocational training programme –which is only one of our many activities– the TES Institute also offers Special Courses that may be repetition of courses in the Core Programme or tailor-made and organised at the request of a country or a group of countries.

In this context, the TES Institute is active through the PHARE programme for Central European Countries, through the MEDSTAT programme in the countries of the Mediterranean Basin and through the TACIS programme in a number of CIS countries.

CONSULTING

In addition to the vocational training activities, we will continue to develop and extend our consulting activities in the above mentioned regions and elsewhere in order to maintain and solidify our position in the market of professional training and staff development for statisticians. These consulting activities consist in either *competence building* by the training of «in-country» trainers with multiplicative effects of the training or *institutional building* by the development of training centres, improvement of vocational training system and curricula. For the time being, the TES Institute has been involved in consulting activities with the Russian Federation, Ukraine and Kazakhstan.

RESEARCH

In the context of the 5th Framework Programme of the Commission, the TES Institute is participating in a consortium which will focus on the creation and the development of on-line training courses. On the basis of possibilities offered by existing tools, the VL-CATS project will provide access over the Internet, to teaching and educational

material with a special focus on Official Statistics. The main objectives of VL-CATS Project are:

- To create a virtual library of all reference material available in Official Statistics;
- To define and implement standards for the development of distance courses in the various fields of Official Statistics;
- To develop a controlled environment that will support the virtual library and give selective access to training courses;
- To develop a service for update and quality assurance of VL-CATS.

The VL-CATS Project foresees the design and the implementation of the following courses:

- Sampling techniques
- Official Statistics and the European Statistical System
- Confidentiality
- Time Series Analysis
- Statistical Modelling
- Strategic Management

The TES Institute is also involved in a consortium which started in January 2001 the implementation of the ASSO project (Analysis System of Official Symbolic Data). The general objective of the ASSO project is to design methods, methodology and software tools for extracting knowledge from multidimensional complex data coming from very large databases in Statistical Offices and Administrations. The aim of the project is to use Symbolic Data Analysis to solve problems of Statistical Offices, in particular problems of confidentiality and missing data.

The previous project SODAS has developed a software prototype for the building and the analysis of symbolic data and has shown the interest of such methods for administrative data. ASSO will improve the first project in order to render it more operational and attractive following users requests. It will also add new innovative methods and demonstrate that these techniques meet needs of Statistical Offices with a particular interest given to business registers, «new economy» data, environment and employment.

PUBLICATIONS

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 (Idescat).

The TES Institute has started the production of TES Manuals on subjects covered by the vocational training programme:

- The first manual available at the TES Institute presents «The role of statistics in a Democracy».
- The second manual to be published soon will cover the Index numbers for spatial and temporary comparisons.

Further manuals will cover topics of sampling techniques, seasonal adjustment methods.

TES NEWSLETTER

The TES Institute will issue its Newsletter «Facts & Visions» on a quarterly basis. This newsletter should be considered as an important information tool between the TES Institute and its partners from all over Europe focusing on matters related to training and staff development in the statistical world. So, may we encourage you to send us any papers or information that might be of interest to be published.

GENERAL INFORMATION

For further details on any of the above-mentioned topics, please contact directly Ms. Valérie Vandewalle (*Head of Curriculum Development*):

Phone: (352) 29.85.85.34
Fax: (352) 29.85.29
E-mail: vvandewalle@tes-institute.lu

European Statistical Training Programme

Core Programme 2002

Course Code	Course Title	Course Leader	Course Location	Date
DAT-102	Survey Non-response: Reduction, Weighting and Imputation	Lynn	London	21-25 January 2002
SOC-109	Systems of Education Statistics	Pilos	Luxembourg	4-6 February 2002
MSI-105	Planning and Management of Statistical Programmes and Structures	Miles	London	18-21 February 2002
PDS-001	Basic Principles of Publication and Dissemination of Statistical Products	Swires-Hennessy	London	4-8 March 2002
ECO-202	The European System of Accounts (ESA95)-Sector Accounts	Newson	Luxembourg	11-13 March 2002
DAT-203	Estimation for Small Areas	Smith & Pfeffermann	Barcelona	18-20 March 2002
ECO-101	Nomenclatures, Classifications and their Harmonisation	Langkjaer	Luxembourg	18-21 March 2002
SIS-203	Geographical Information Systems	Painho	Lisbon	8-11 April 2002
ECO-205	The European System of Accounts (ESA95)-Quarterly Accounts	Barcellan	Luxembourg	15-17 April 2002
PDS-102	Effective Presentation of Official Statistics to News Media	Finn	London	15-19 April 2002
OSG-001	The European Statistical System	Eurostat Experts	Luxembourg	22-24 April 2002
ASA-102	Price Index Theory and Price Statistics	von der Lippe	Düsseldorf	27-31 May 2002
SOC-106	Introduction to Demographic Data and their Analysis	Andersen	Copenhagen	27-31 May 2002
ECO-102	Business Registers: Set-up, Use and Maintenance	Bergström	Oslo	3-7 June 2002
DAT-002	Sampling Techniques and Practice	Smith	Southampton	10-21 June 2002
SOC-001	Systems of Social Statistics	Everaers	Voorburg	17-21 June 2002
SIS-202	Measurement of the Quality of Statistics	Depoutot	Paris	24-26 June 2002
SOC-108	Systems of Health Statistics	Bonte	Luxembourg	16-19 September 2002
PDS-103	Techniques of Electronic Data Dissemination	Argueso	Madrid	23-26 September 2002
OSG-003	Confidentiality and Protection of Privacy	Nanopoulos	Luxembourg	30 September-2 October 2002
ECO-001	National Accounts Statistics in Practice/EN	van Nunspeet	Voorburg	7-18 October 2002
SIS-150	Statistics on the Information Society	Koszerek & Deiss	Luxembourg	21-23 October 2002
DAT-103	New Advanced Technologies for Data Collection	Kunzler	Luxembourg	28-30 October 2002
ECO-152	Economic Accounts for Agriculture	Eidmann	Luxembourg	4-6 November 2002
ECO-153	Structural Business Statistics	Egmose	Copenhagen	18-21 November 2002
SIS-151	New Tools and Methods of Data Transmission for IT Specialists	Knueppel	Luxembourg	19-22 November 2002
ECO-125	Economic Analysis and Flash Estimates: Application for SNA Statistics in Practice	Mazzi	Luxembourg	25-28 November 2002

First Barcelona Workshop on Survival Analysis

Barcelona, Spain: June 12th-14th, 2002

Research and Training in Failure Time Methods in the New Millennium

The Statistics and Operations Research Department of the Universitat Politècnica de Catalunya, together with the Biostatistics Department of Harvard University and the Spanish Region of the International Biometric Society, are very pleased to announce a three-days workshop on Survival Analysis in Barcelona from June 12th to 14th, 2002. The workshop will consist on invited methodological sessions, a round table on how to train future generations of biostatisticians, a short course on optimization methods for statistics and a session of contributed papers in poster form.

The topics to be covered in the invited sessions include:

- Sampling and Observational Structures. What happens under various conditioning schemes?
- Multivariate and Covariate Methods and Models. Application of the indirect method developed in econometrics to analyze noisy recurrent event data.
- Alternatives to the Cox model. Using additive hazards regression models in multistate modeling.
- Issues on Interval-censored data. Current status data with competing risks.
- Event History Analysis. Cost or Quality of Life processes related to event histories. Structural models, PH version.
- New challenges in survival studies with missing data

Confirmed invited speakers are (in alphabetical order):

P. K. Andersen, University of Copenhagen, Denmark
D. Commenges, Université de Bordeaux II, France
E. Goetghebeur, Ghent University, Belgium
N. Jewell, University of California, Berkeley, USA
J. Kalbfleisch, University of Michigan, USA
N. Keiding, University of Copenhagen, Denmark
J. Klein, University of Wisconsin, USA
S. Lagakos, Harvard University, USA
J. Lawless, University de Waterloo, Canada
S. F. Nielsen, University of Copenhagen, Denmark
R. Prentice, University of Washington, USA
B. Turnbull, Cornell University, USA
L. J. Wei, Harvard University, USA

Local organizer commitee:

Guadalupe Gómez, Universitat Politècnica de Catalunya, Barcelona
M. Luz Calle, Universitat de Vic, Vic
Olga Julià, Universitat de Barcelona, Barcelona
Carles Serrat, Universitat Politècnica de Catalunya, Barcelona

Sponsors:

Institut Estadística de Catalunya (Idescat)
Servei d'Estadística de la Universitat Autònoma de Barcelona

The total number of participants will be limited to approximately 60 people.

If you are interested in the event and want to be included in our mailing list for future information, please send an e-mail to carles.serrat@upc.es. The web site of the meeting is still under construction but, meanwhile you can start visiting, and enjoying Barcelona from the following web pages:

http://www.bcn.es/english/ihome.htm
http://www.bcn.es/english/turisme/saber/frames/ilsaber.htm
http://www.bcn.es/english/barcelon/cultura/frames/ilcultu.htm
http://www.bcn.es/english/barcelon/transport/itransp.htm

Institut d'Educació Contínua
Universitat Pompeu Fabra

Applied Statistics Week

(8th edition, Barcelona 25-29 June 2002)

Short Courses in Statistical Genetics

Aims

The eighth APPLIED STATISTICS WEEK is being organized by the Pompeu Fabra University (UPF) from 25 to 29 June 2002, in Barcelona.

The APPLIED STATISTICS WEEK aims to provide a set of intensive short courses on a particular statistical theme of an applied nature. The courses are presented by acknowledged leading researchers of international stature, who are also known to have excellent teaching skills.

Past themes of the APPLIED STATISTICS WEEK have been: «Statistics in the Health Sciences» (1995), «Statistics in Classification and Pattern Recognition» (1996), «Design and Analysis of Survey Data» (1997), «Statistics in Environmental Science» (1998), «Statistics in Marketing Research» (1999), «Statistics in Society» (2000) and «Statistics in Finance» (2001). This year's theme is «Statistical Genetics». As the Human Genome Project establishes the complete genetic sequence of mankind, science is fast approaching an exciting new stage in its ability to understand genetic forces in man and their relation to disease. In parallel to the human project there are major sequencing efforts in other organisms, enabling us to understand broader questions in evolution and classification. This research has brought with it an explosion of data, leading to an increased need for sophisticated statistical, mathematical and computational tools to enable complex data collection, analysis and interpretation of the results. The three courses in this year's Applied Statistics Week provide a comprehensive view of the statistical tools which are being used extensively in present-day genetic research.

The first two courses are jointly presented by David Balding and John Whittaker of Imperial College, London, and Andrew Morris of the Wellcome Trust Centre for Human Genetics at the University of Oxford. The first course deals with genetic epidemiology, introducing statistical methods for understanding genetic contributions to the causation of human diseases. The second course is a more advanced course on association methods of gene mapping, and is preferably taken in conjunction with the first course, although it can be taken alone by researchers who already have a firm grasp of basic statistics and familiarity with the basic ideas of genetics. The third course, presented by Sandrine Dudoit of the University of California at Berkeley, gives a complete review of the statistical design and analysis of DNA microarray experiments, with special attention to modern methods of pattern recognition and classification. All three courses involve an introduction to and discussion of computational software available in these areas of research. After attending these courses, participants will have an insight into the important role played by statistical methodology in genetic research. These courses are primarily aimed at researchers in genetics and allied fields, but are equally valuable to statisticians interested in initiating themselves into this important area of modern medical science.

We feel that UPF is uniquely placed to foster such intensive courses of high quality in Spain and, indeed, in Europe as a whole. The university and the city of Barcelona enjoy an optimal location for gathering scholars from different parts of Spain and Europe, a city known for its strong work and innovation ethic as well as its richness of leisure and cultural activities. The courses are concentrated into one week to facilitate the enrolment of working professionals and the academic community. We also promote a lively interaction between participants and instructors that assist in achieving the goal of improving the application of statistical concepts and methods to problems in our society.

Programme directors

Michael J. Greenacre
Albert Satorra
Pompeu Fabra University, Barcelona

Target Audience

These courses are aimed at all researchers in the field of genetics, genomic and proteomic research, including biologists, epidemiologists, veterinary scientists, biostatisticians, health scientists in general as well as decision-makers in health economics and insurance. Statisticians interested to move into this field of application will find these courses particularly useful.

Place: classes and enrolment

Classes will be held at IDEC (Institute for Continuing Education), Pompeu Fabra University, at the address: Balmes 132, Barcelona 08008.

Language

All the courses will be taught in English.

Supporting Institution

As in all seven previous editions, the eight APPLIED STATISTICS WEEK is being organized in cooperation with the **Institut d'Estadística de Catalunya (Idescat)**.

Programme

Course 1.

INTRODUCTION TO GENETIC EPIDEMIOLOGY

David Balding, Andrew Morris & John Whittaker
Imperial College, London & Wellcome Trust Centre for Human Genetics, University of Oxford

Genetic epidemiology is concerned with understanding genetic contributions to the causation of human diseases. Rapid advances in molecular biology have led both to new types of genetic data and to a vast increase in the quantity of available data. There is consequently an increasing demand for statisticians with the skills to analyse these data. This course provides an introduction to the methods used to analyse genetic epidemiology data in order to assess the role of genes in causing disease, and to locate and characterize the genes involved.

One of the first tasks in genetic epidemiology is to establish whether or not a disease is familial, and if so whether the familial aggregation is due to shared genes or shared environment, or both. If genes are involved, is there a single major gene, or many each with small effects? After reviewing methods for tackling these problems, the course will focus on methods for locating disease genes, grouped into four broad classes. Linkage analysis includes both likelihood-based methods for analysing transmissions of marker alleles and disease in family trees (parametric linkage) and methods for investigating alleles shared by related affecteds (nonparametric linkage). Allelic association methods also fall into two classes: population-based methods which look for differences in gene frequencies between samples of unrelated affected and unaffected individuals, and family-based methods which study patterns of gene transmission from parents to affected offspring.

In introducing these methods, our emphasis is on the principles of data analysis and study design rather than in the details of particular computational implementations. We will contrast the strengths and weaknesses of the different approaches and try to indicate the circumstances in which each is most appropriate. The course assumes a good grasp of basic statistics including likelihood-based estimation, and hypothesis testing in contingency tables. There will be a brief introduction to the relevant genetics background, but familiarity with the notions of mutation and recombination will be helpful.

Date: Tuesday, 25 June 2002, 9.30-13.30, and 15.00-18.00; Wednesday, 26 June 2002, 9.30-13.30, and 15:00-18:00.

David Balding is Professor of Statistical Genetics at Imperial College of Science, Technology and Medicine, University of London. Before that he was for five years Professor of Applied Statistics at the University of Reading. He was raised and educated in Australia before coming to England to take his PhD in Mathematics at Oxford in 1989. His research is in statistical aspects of genetic epidemiology, population genetics, and bioinformatics. He is an Associate Editor of International Statistical Review and Biostatistics, and an editor of the *Handbook of Statistical Genetics* (Wiley, 2001).

Andrew Morris is Research Fellow in Statistical Genetics at the Wellcome Trust Centre for Human Genetics, University of Oxford. He obtained his PhD in Applied Statistics at the University

of Reading, where he also worked for two-years as a postdoctoral researcher, funded by Pfizer Central Research, UK. His research covers both family-based and population-based association methods for gene mapping.

John Whittaker is Reader in Statistical Genetics at Imperial College of Science, Technology and Medicine, University of London. He obtained his PhD from the Department of Statistics, University of Sheffield, and subsequently worked at the University of Reading where he was Reader in Applied Statistics. His research is in statistical genetics, particularly family-based association methods of gene mapping and applications of generalised estimating equations. He has also developed statistical methods for plant and animal breeding.

Course 2.

ASSOCIATION METHODS OF GENE MAPPING

David Balding, Andrew Morris & John Whittaker
Imperial College, London & Wellcome Trust Centre for Human Genetics, University of Oxford

The course will be at a more advanced level than the introductory course on genetic epidemiology, and will build on the ideas introduced at the end of that course. The two courses may be taken together, or this course may be taken alone by researchers who already have a firm grasp of basic statistics and familiarity with the basic ideas of genetics.

The course will cover candidate-gene analyses, linkage disequilibrium mapping, and family-based tests of linkage and association, such as the transmission disequilibrium test (TDT) and its extensions and generalisations. Each of these approaches relies on statistical associations between observed disease state and genetic markers that are affected by population genetics processes. The course therefore includes an introduction to the key ideas of population genetics, including a brief introduction to coalescent models.

Developments in molecular genetics have led recently to intense interest in association mapping, and statistical methodology has been developing rapidly, so that there are currently no textbooks adequately covering these methods. We will discuss recent developments, and illustrate them with relevant datasets. Topics included are: family-based association methods as well as population-based fine-scale mapping using high density marker maps, and accounting for population substructure using case-control data. The availability of computer software implementing the methods will be discussed, but the course will emphasize underlying statistical principles rather than the details of particular computer packages.

Date: Thursday, 27 June 2002, from 9.30 to 13.30, and from 15.00 to 18.00.

Course 3.

DESIGN AND ANALYSIS OF DNA MICROARRAY EXPERIMENTS

Sandrine Dudoit
Department of Biostatistics, University of California at Berkeley

cDNA microarrays and high-density oligonucleotide chips are novel biotechnologies which allow the monitoring of expression levels in cells for thousands of genes simultaneously. Microarrays

are being applied increasingly in biological and medical research to address a wide range of problems, for example to study the molecular variations among tumors in cancer research, or to study the gene expression response of model organisms such as yeast to different types of stress conditions. Gene expression experiments generate large and complex multivariate datasets. The application of sound statistical design and analysis principles can greatly improve the efficiency and reliability of microarray experiments throughout the data acquisition and analysis process.

This short course, which she has recently presented at the Harvard medical School, will survey statistical and computational issues arising in the design and analysis of DNA microarray experiments. After a brief introduction to genome biology, the following topics will be addressed: experimental design, image analysis, normalization, the identification of differentially expressed genes, pattern discovery and recognition for genes and mRNA samples. Access to an efficient, portable, and distributed statistical computing environment is an essential aspect of the analysis of gene expression data. Statistical computing resources developed as part of the Bioconductor project will also be discussed. This collaborative effort aims more generally to produce an open source and open development computing environment for biologists and statisticians working in bioinformatics (`http://www.bioconductor.org`).

Sandrine Dudoit is Assistant Professor of Biostatistics in the School of Public Health at the University of California at Berkeley. Her doctorate, supervised by Prof. Terence Speed, was on the linkage analysis of complex human traits. Before joining Berkeley she had postdoctoral training in genomics in the laboratory of Professor Patrick O. Brown at Stanford University, working on the development of statistical methods for the design and analysis of DNA microarray experiments. Her research interests are in the application of statistics to problems in genetics and molecular biology.

Date: Friday, 28 June 2002, from 9.30 to 13.30, and from 15.00 to 18.00, and Saturday, 29 June 2002, from 9.30 to 13.30.

Course fees

The course fees include course notes, a Diploma of the Institut d'Educació Contínua (IDEC), meals and refreshments during the courses.

- **Course 1:** 300 € if paid before the 20th of May 2002, otherwise 445 €.
- **Course 2:** 190 € if paid before the 20th of May 2002, otherwise 270 €.
- **Course 3:** 235 € if paid before the 20th of May 2002, otherwise 350 €.

For those who register for 2 courses the fee is reduced by 10%.
For those who register for 3 courses the fee is reduced by 20%.

SPECIAL FEES for doctoral students and participants in a previous APPLIED STATISTICS WEEK:
Course 1: 215 €
Course 2: 135 €
Course 3: 170 €

Registration

The registration form must be sent by the 20th of May to take advantage of to the reduced fees. Payment can be made by bank transfer (including a copy of the bank transfer), by credit card (VISA) or by bank cheque to the Institut d'Educació Contínua. In the case of foreign payments, the cheque should be in convertible pesetas or euros.

Accommodation

The Continuing Education Institute has arranged a special price of 72 euros with two nearby three-star hotels for the accommodation of participants:

Astòria***, address: Paris, 203 (300m from IDEC)

Balmes***, address: Mallorca, 216 (200m from IDEC)

In all cases, prices are for a double room for either one or two persons (breakfast and VAT not included). You can ask for reservation of rooms in the Registration Form.

Information and enrolment

IDEC (Continuing Education Institute)
Balmes, 132
08008 Barcelona-Spain
Telephone: (34) 93 542 18 00
Fax: (34) 93 542 18 08
http://www.upf.es/idec e-mail:idec@upf.es

Informació per als autors i lectors

NORMES PER A LA PRESENTACIÓ D'ARTICLES A *QÜESTIIÓ*

La revista accepta originals no sotmesos a 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, i poden incloure aspectes computacionals i/o de caire docent, i es publicaran exclusivament en anglès, acompanyats d'un breu resum en català a càrrec de la pròpia revista.

Tots els originals destinats a les seccions temàtiques de *Qüestiió* seran sotmesos sistemàticament a una avaluació prèvia a càrrec d'especialistes independents i/o membres del Consell Editor, llevat dels articles convidats i de les reimpressions d'articles. El resultat de l'avaluació es comunicarà a l'autor principal en el cas que es proposin correccions formals o de contingut.

Quan es tramet un original, la revista emet un justificant de recepció, la data del qual figura com a «data de rebuda» en la publicació de l'article, mentre que la «data d'acceptació» de l'article és la data de recepció de la versió definitiva.

Per a la presentació d'articles, l'autor haurà de trametre a l'adreça postal de la Secretaria de *Qüestiió* (Institut d'Estadística de Catalunya) dues còpies del treball complet en DINA4, a una sola cara i a doble espai. Simultàniament, s'haurà d'enviar a l'adreça electrònica questiio@idescat.es la primera pàgina de l'article en format PDF o PS, en la qual s'hi farà constar necessàriament el títol, el nom de l'autor o autors, l'afiliació professional i l'adreça completa, un resum de 75-100 paraules, les paraules clau principals i l'adscripció a la classificació Mathematics Subject Classification 2000 de l'American Mathematical Society (MSC2000). Abans de presentar els articles a la revista, s'aconsella als autors que assegurin la correcció lingüística dels textos en anglès. Les referències bibliogràfiques es faran indicant el cognom de l'autor seguit de l'any de la publicació entre parèntesis [i.e.: Mahalanobis (1936), Rao (1982b)] i es llistaran alfabèticament al final de l'article. Les referències múltiples d'un mateix autor s'ordenaran cronològicament. Les notes explicatives es numeraran correlativament i apareixeran al peu de la pàgina corresponent. Les taules i figures també es numeraran correlativament en el text i es reproduiran directament dels originals tramesos en cas que no sigui possible fer-ne l'autoedició.

Una vegada acceptat l'article, caldrà que l'autor trameti la versió impresa i electrònica definitiva a la Secretaria de *Qüestiió* , d'acord amb les instruccions que li seran indicades pel responsable de l'avaluació del seu article. Es recomana que la versió final es trameti mitjançant el processador de textos LaTeX 2ε .

La Secretaria de *Qüestiió* posa a disposició dels autors que ho sol·licitin plantilles en format LaTeX 2ε per fer-ne l'edició, així com les referències adients de la classificació MSC2000 de l'American Mathematical Society. Per a més informació sobre l'ús i els criteris per a l'adopció d'aquesta classificació es pot consultar directament la Mathematics Subject Classification 2000 (MSC2000) en el web de l'Institut d'Estadística de Catalunya.

Secretaria de QÜESTIIÓ
Institut d'Estadística de Catalunya (Idescat)
Via Laietana, 58
08003 Barcelona

Tel: +34-93 412 09 24 o +34-93 412 00 88
Fax: +34-93 412 31 45
E-mail: questiio@idescat.es

GUIDELINES FOR SUBMITTING ARTICLES TO *QÜESTIIÓ*

The journal accepts for publication only original articles that have not been submitted to any other journal in the areas of Statistics, Operations Research, Official Statistics and Biometrics. Articles may be either theoretical or applied, and may include computational or educational elements. Publication will be exclusively in English, with an abstract in Catalan (abstract translation into Catalan can be done by *Qüestiió*'s staff).

All articles submitted to thematic sections of *Qüestiió* will be forwarded for systematic peer review by independent specialists and/or members of the Editorial Board, except for those articles specifically invited by the journal or reprinted with permission. Reviewer comments will be sent to the primary author if changes are requested in form or content.

When an original article is received, the journal sends the author a dated receipt; the actual acceptance date is assigned when the definitive final version is received.

To submit an article, the author must send two complete copies, double-spaced on one side only of A4 paper, to the postal address for *Qüestiió* Secretary (Institut d'Estadística de Catalunya), as well as an e-mail message to questiio@idescat.es with the title page of the article attached in PDF or PS format. This title page must contain the following items: title, name of the author(s), professional affiliation and complete address, and an abstract (75-100 words) followed by the principal keywords and MSC2000 classification of the American Mathematical Society. Before submitting an article, the author(s) would be well advised to ensure that the text uses correct English. Bibliographic references within the text must follow this format: author surname followed by the year of publication in parentheses [i.e., Mahalanobis (1936), Rao (1982b)], with complete reference citations listed alphabetically at the end of the article. Multiple publications by a single author are to be listed chronologically. Explanatory notes should be numbered sequentially and placed at the bottom of the corresponding page. Tables and figures should also be numbered sequentially and will be reproduced directly from the submitted originals if it is impossible to include them in the electronic text.

Once the article has been approved, the author must submit a final printed and electronic copy to *Qüestiió* (Institut d'Estadística de Catalunya), following the instructions of the editor responsible for that article. It is recommended that this final version be submitted using the LaTeX2ε word processor.

In any case, upon request the *Qüestiió* Secretary will provide authors with LaTeX2ε templates and appropriate references to the MSC2000 classification of the American Mathematical Society. For more information about the criteria for using this classification system, authors may consult directly the Mathematics Subject Classification 2000 (MSC2000) at the website of the Institut d'Estadística de Catalunya.

QÜESTIIÓ Secretary
Institut d'Estadística de Catalunya (Idescat)
Via Laietana, 58
08003 Barcelona

Tel: +34-93 412 09 24 o +34-93 412 00 88
Fax: +34-93 412 31 45
E-mail: questiio@idescat.es

NORMES PER A LA PUBLICACIÓ D'ANUNCIS INSTITUCIONALS A *QÜESTIIÓ*

Qüestiió convida les entitats patrocinadores, les institucions col·laboradores, els organismes públics i privats, i tota la comunitat científica vinculada a l'estadística o la investigació operativa, a la publicació d'anuncis institucionals sobre cursos, seminaris, congressos i activitats similars que, preferentment, tinguin lloc en el nostre país. Els textos poden presentar-se en anglès, francès, català o en qualsevol altra llengua oficial a l'Estat espanyol. Les iniciatives per a una possible publicació sempre són a instància de les entitats interessades, de manera que *Qüestiió* no fa una cerca sistemàtica d'esdeveniments d'aquesta naturalesa, ni té cap ànim d'exhaustivitat en les ressenyes d'activitats finalment publicades.

Una vegada aprovada la inclusió dels anuncis sol·licitats es procedirà a la seva publicació, i es reproduirà directament dels originals tramesos amb les mides adequades i la màxima qualitat tipogràfica possible; en aquest cas, *Qüestiió* no procedeix a cap mena de procés d'autoedició de la versió impresa que l'anunciant hagi tramès. Si els originals es trameten en els mateixos termes electrònics exigits per als articles (vegeu «Normes per a la presentació d'articles a *Qüestiió*»), la revista procedirà a la seva autoedició. Si es desitja una qualitat superior a la reproducció simple o l'autoedició, o bé la seva publicació en color, els sol·licitants hauran de posar-se en contacte amb la Secretaria de *Qüestiió* per tal de trametre els fotolits dels textos originals corresponents.

La disposició dels textos i les figures adjuntes dels anuncis han de procurar la màxima intel·ligibilitat i claredat expositiva, sense atapeir la informació ni utilitzar formats o fonts de lletres excessivament petites. D'altra banda, la publicitat ha de ser fidedigna, exempta d'enganys i respectuosa amb les persones i institucions. En qualsevol cas, la direcció de *Qüestiió* es reserva la decisió final pel que fa a la seva publicació.

L'anunciant es compromet a lliurar els textos/materials amb l'antelació que se li indiqui per a la inserció en els números/volums de *Qüestiió* que prèviament s'hagi establert. La revista no es fa responsable dels retards, per part de l'anunciant, que impedeixin la publicació de l'anunci en els termes previstos.

Mònica M. Jaime
Secretaria de *Qüestiió*
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona
Tel: +34-93 412 15 36
Fax: +34-93 412 31 45
E-mail: questiio@idescat.es

GUIDELINES FOR INSTITUTIONAL ADVERTISEMENTS IN *QÜESTIIÓ*

Qüestiió invites all sponsor entities, collaborating institutions, other public and private bodies and the entire scientific community related to Statistics or Operations Research to submit institutional advertisements on courses, seminars, congress and similar activities that will be held, preferably in our country. These will be accepted in English, French, Catalan or any of the other official languages in Spain. The initiative should always come from the entities interested in advertising them so that *Qüestiió*'s aim is not to do a systematic search of these events and therefore does not publish a comprehensive list of such activities.

Once their insertion is approved the advertisements will be reproduced from the most accurate photocopy of the originals sent by the advertiser to *Qüestiió* in paper copy, with the appropriate size and at the best possible typographic quality. Therefore, in this case the journal does not elaborate any further editing process to the printed version that the advertiser has sent. If the original advertisements are sent in the same electronic format requested by the articles (please see «Guidelines for the submission of articles for *Qüestiió*») the journal will print it directly from the file. If a better quality than the simple reproduction or automatic printing or a colour version of the adverts is desired, the authors should contact the Secretary of *Qüestiió* in order to negotiate this.

The typesetting of texts and figures in the advertisement should have maximum intelligibility and clearness, neither compressing the information too much nor using formats or letter fonts that are too small. Furthermore, the information has to be reliable, without errors and respectful of the people and institutions. The management of *Qüestiió* has the right to a final decision concerning the insertion of the advertisement.

Advertisers commit themselves to give the text/materials on request in order to insert them in the issues of *Qüestiió* that have been previously agreed. The journal is not responsible for any delay from the announcer that could prevent the advertisement from been published on the agreed terms.

Mònica M. Jaime
Secretaria de *Qüestiió*
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona
Tel: +34-93 412 15 36
Fax: +34-93 412 31 45
E-mail: questiio@idescat.es

NORMES PER A LA PUBLICACIÓ D'ANUNCIS PRIVATS O AMB FINALITAT COMERCIAL A *QÜESTIIÓ*

Qüestiió accepta la publicació d'anuncis privats o amb finalitat comercial sobre productes, serveis o altres eines promocionals a l'entorn de l'estadística o la investigació operativa. Els textos poden presentar-se en anglès, francès, català o en qualsevol altra llengua oficial a l'Estat espanyol. Les iniciatives per a una possible publicació sempre són a instància de les organitzacions que hi estiguin interessades, de manera que *Qüestiió* no fa una cerca sistemàtica de novetats o productes d'aquesta naturalesa ni té cap ànim d'exhaustivitat en els anuncis finalment publicats.

Els anuncis en **blanc i negre** s'elaboren a partir de la fotocòpia més acurada possible dels originals que trameti l'anunciant en versió impresa, amb les mides adequades i la màxima qualitat tipogràfica. Per tant, en aquest cas la revista no efectua cap procés d'edició ulterior respecte de la versió impresa que l'anunciant hagi tramès. Alternativament, si els anuncis originals es trameten en els mateixos termes formals exigits per als articles (vegeu «Normes per a la presentació d'articles a *Qüestiió*»), la revista procedirà a la seva autoedició. Igualment, si es desitja una qualitat superior a la reproducció simple, els sol·licitants hauran de trametre els fotolits dels originals corresponents o encarregar-los a *Qüestiió,* que els facturarà separadament.

Els anuncis **en color** requereixen els fotolits dels textos originals, que poden ser subministrats directament per l'anunciant o bé encarregats per la revista a compte de l'anunciant; en el segon cas, l'anunciant ha de trametre a la revista els originals impresos en color amb la màxima qualitat, per tal de filmar-los amb les millors garanties i condicions. El cost dels fotolits realitzats per *Qüestiió* serà sempre a càrrec de l'anunciant, a qui se li repercutirà l'import i l'IVA d'aquests, juntament amb les tarifes que corresponen a la modalitat d'anunci per la qual hagi optat.

La disposició dels textos i figures adjuntes dels anuncis ha de procurar la màxima intel·ligibilitat i claredat expositiva, sense atapeir la informació ni utilitzar formats o fonts de lletres excessivament petites. D'altra banda, la publicitat ha de ser fidedigna, exempta d'enganys i respectuosa amb les persones i institucions. En qualsevol cas, la direcció de *Qüestiió* es reserva la decisió final de la seva inclusió.

L'anunciant es compromet a lliurar els textos/materials amb l'antelació que se li indiqui per a la seva inserció en el(s) número(s)/volum(s) de *Qüestiió* que prèviament s'hagi establert. La revista no es fa responsable dels retards per part de l'anunciant que impedeixin la publicació de l'anunci en els termes previstos.

Imports:

1 pàgina en color (un número aïllat):	125.000 PTA + IVA
1 pàgina en color (tres números consecutius):	200.000 PTA + IVA
1 pàgina en blanc i negre (un número aïllat):	30.000 PTA + IVA
1 pàgina en blanc i negre (tres números consecutius):	50.000 PTA + IVA
1/2 pàgina en blanc i negre (un número aïllat):	20.000 PTA + IVA
1/2 pàgina en blanc i negre (tres números consecutius):	35.000 PTA + IVA

Mides opcionals dels anuncis:

1 pàgina sencera (espai intern):	19.0 cm. × 12.3 cm.
1 pàgina sencera (espai extern):	23.8 cm. × 17.0 cm.
1/2 pàgina (espai intern):	9.5 cm. × 12.3 cm.
1/2 pàgina (espai extern):	11.9 cm. × 17.0 cm.

Forma de Pagament:

- Transferència bancària al compte: 2013-0100-53-0200698577
- Xec bancari nominatiu a l'Institut d'Estadística de Catalunya
- Pagament amb targeta de crèdit

El pagament serà per l'import total de la factura corresponent, on hi figurarà el cost dels fotolits en el cas que l'edició de l'anunci hagi estat a càrrec de l'Institut. En el cas que l'anunciant necessiti una factura proforma, només cal que ho faci saber amb l'antelació suficient.

Correspondència:

Mònica M. Jaime
Secretaria de *Qüestiió*
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona

Tel: +34-93 412 15 36
Fax: +34-93 412 31 45
E-mail: questiio@idescat.es

GUIDELINES FOR THE PRIVATE OR COMMERCIAL ADVERTISEMENTS IN *QÜESTIIÓ*

Qüestiió accepts for their publication both private and commercial advertisements on products, services or other promotional tools related to statistics or operational research and will be accepted in English, French, Catalan or any of the official languages in Spain. The initiatives should always come from entities interested in advertising them so that *Qüestiió*'s aim is not to do a systematic search of news and therefore does not publish a comprehensive list of such private or profit activities.

The **black and white** advertisements are made out from the most accurate photocopy of the originals sent by the advertiser to *Qüestiió* in paper copy with the appropriate size and at the best possible typographic quality. Therefore, in this case the journal does not elaborate any further editorial process to the printed version that the advertiser has sent. Alternatively, if the original advertisements are sent in the same formal terms required by the articles (please see «Guidelines for the submission of articles for *Qüestiió*»), the journal will proceed to its autoedition. In the same way, if a better quality than the simple reproduction is wanted, the authors should send the photolits of the corresponding original texts or, on the other hand, order to *Qüestiió* their fulfilment, which will be invoiced separately from the rates charged as advertisements.

The advertisements **in colour** need the photolits of the original texts, which can be provided directly by the advertiser or requested by *Qüestiió* to the advertiser charge; in the second case, the advertiser must sent to the journal the originals printed in colour with the best possible quality, so that they can be filmed at the best conditions and guarantees. The cost of the photolits made by *Qüestiió* will always be charged to the advertiser together with the VAT derived from it, plus the prices corresponding to the type of the advertisement that has been chosen.

The set up of texts and figures of the advertisement should provide the maximum intelligibility and clearness, neither squeezing together the information nor using set ups or letter types that are too small. On the other hand the publicity has to be reliable, without fraud and respectful to the persons and institutions. The direction of *Qüestiió* has the right of the last decision concerning the insertion of the advertisement.

The advertiser commits 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. x 12.3 cm.
1 full page (external space):	23.8 cm. x 17.0 cm.
1/2 page (internal space):	9.5 cm. x 12.3 cm.
1/2 page (external space):	11.9 cm. x 17.0 cm.

Payment:

- A bank transfer to account number: 2013-0100-53-0200698577
- A bank cheque to Institut d'Estadística de Catalunya
- Charge on a credit card

The payment should be for the amount shown at the invoice, where it will be shouwn the total cost of the photolits, in case that *Qüestiió* would be in charge of the filmation of the advertisement. If advertiser need a pro-forma invoice, he should let us know some time in advance so that *Qüestiió* could send it to the proper address.

Mail address:

Mònica M. Jaime
Secretaria de *Qüestiió*
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona

Tel: +34-93 412 15 36
Fax: +34-93 412 31 45
E-mail: questiio@idescat.es

Butlleta de subscripció a la revista *Qüestiió*

Nom i cognoms _____

Empresa/Institució _____

Adreça _____

Codi postal _____ Ciutat _____

Tel. _____ Fax _____ NIF _____

Data _____

Signatura

Desitjo subscriure'm a ***Qüestiió*** per a l'any 2002

Preu de subscripció vigent:

– Estat espanyol: 21,64 €/any (3.600 Pta) (IVA inclòs)
– Estranger: 24,04 €/any (4.000 Pta) (IVA inclòs)

Forma de pagament

☐ Transferència al compte 2013-0100-53-0200698577

☐ Domiciliació bancària al compte número

☐☐☐☐ ☐☐☐☐ ☐☐ ☐☐☐☐☐☐☐☐☐☐

☐ Xec nominatiu a l'Institut d'Estadística de Catalunya

☐ Gir postal

☐ En efectiu

Retorneu aquesta butlleta (o una fotocòpia) a:

Qüestiió
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona

Preu de números solts (actuals i endarrerits):

– Estat espanyol: 9,02 €/exemplar (1.500 Pta) (IVA inclòs)
– Estranger: 10,22 €/exemplar (1.700 Pta) (IVA inclòs)

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

Novetats editorials en matèria estadística de la Generalitat de Catalunya
gener-juny 2002

- **Estadística de biblioteques 2000**
  Característiques bàsiques
  Idescat
  Gener 2002, 8,50 €, 131 pp.
  ISBN 84-393-5614-5

- **Mercat de treball 2001**
  Ampliació de resultats anuals de l'enquesta de població activa
  Idescat
  Maig 2002, 11,00 €, 326 pp.
  ISBN 84-393-5702-8

- **Classificació catalana de productes per activitats (CCPA-96)**
  Adaptació de la CNPA-96
  Idescat
  Març 2002, 12,50 €, 435 pp.
  ISBN 84-393-5637-4

- **Localització de l'activitat econòmica 1999**
  Empreses, professionals, establiments i superfícies
  Idescat
  Maig 2002, 10,30 €, 365 pp.
  ISBN 84-393-5738-9

- **Estadístiques de la societat de la informació. Catalunya 2001**
  Departament d'Universitats, Recerca i Societat de la Informació
  Secretaria de Telecomunicacions i Societat de la Informació
  Observatori de la Societat de la Informació
  2001, 9,02 €, 176 pp.
  ISBN 84-393-5429-0

- **Migracions internacionals i població jove de nacionalitat estrangera a Catalunya**
  Departament de la Presidència
  Secretaria General de Joventut
  2002, 13,00 €, 140 pp.
  ISBN 84-393-5511-4

- **Cens agrari 1999. Vol. 1**
  Aprofitament de la terra i ramaderia
  Idescat
  Juny 2002, 11,50 €, 238 pp.
  ISBN 84-393-5754-0

- **Moviment natural de la població 2000**
  Dades comarcals i municipals
  Idescat
  Juny 2002, 8,70 €, 106 pp.
  ISBN 84-393-5755-9

- **Classificació catalana d'educació 2000 (CCED-2000).**
  Adaptació de la CNED-2000
  Idescat
  Juny 2002, 10,00 €, 140 pp.
  ISBN 84-393-5761-3

- **Localització de l'activitat econòmica 2000**
  Empreses, professionals, establiments i superfícies
  Idescat
  Maig 2002, 10,30 €, 365 pp.
  ISBN 84-393-5739-7

- **Projeccions de llars de Catalunya 2010**
  Comarques i àmbits del Pla territorial
  Idescat
  Juny 2002, 9,50 €, 310 pp.
  ISBN 84-393-5762-1

- **Les dones a Catalunya: dades estadístiques**
  Departament de la Presidència
  Institut Català de la Dona
  2001, 22,80 €, 246 pp.
  ISBN 84-393-5571-8

- **Anuari de dades meteorològiques 2000**
  Departament de Medi Ambient
  Direcció General de Qualitat Ambiental
  2001, 10,22 €, 112 pp.
  ISBN 84-393-5467-3

LLIBRERIES DE LA GENERALITAT

Barcelona
Rambla dels Estudis, 118 (tel. 93 302 64 62)
llibrbcn@correu.cattel.com

Girona
Gran Via de Jaume I, 38 (tel. 972 22 72 67)
llibrgi@ibernet.com

Lleida
Rambla d'Aragó, 43 (tel. 973 28 19 30)
llibrlle@ibernet.com

Madrid
Blanquerna. Llibreria catalana.
Serrano, 1 (tel. 91 431 00 22)
blanquerna@nauta.es

PUNT DE VENDA
Puigcerdà
Plaça del Rec, 5 (tel. 972 88 05 14)

VENDA PER CORREU
Apartat 2800, 08080 Barcelona
eadop@correu.gencat.es
Plaça del Rec, 5 (tel. 972 88 05 14)

Publicacions de la Generalitat. Apartat de correus 2800, 08080 Barcelona

Nom i cognoms _____

Empresa / Institució _____

Professió _____ E-mail _____

Adreça _____

Població _____ CP _____

NIF / DNI _____ Telèfon _____

Desitjo rebre els volums

_____

_____

○ Carregueu l'import a la meva targeta de crèdit Signatura

☐ American Express ☐ 6000

☐ Master Charge ☐ Visa

○ Contra reemborsament

Núm. de targeta |_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|

Data de caducitat |_|_|_|_|

DOGC